Springer Actuarial

Hirbod Assa
Peng Liu
Simon Wang *Editors*

Quantitative Risk Management in Agricultural Business

Springer

Springer Actuarial

This is a series on actuarial topics in a broad and interdisciplinary sense, aimed at students, academics and practitioners in the fields of insurance and finance.

Springer Actuarial informs timely on theoretical and practical aspects of topics like risk management, internal models, solvency, asset-liability management, market-consistent valuation, the actuarial control cycle, insurance and financial mathematics, and other related interdisciplinary areas.

The series aims to serve as a primary scientific reference for education, research, development and model validation.

The type of material considered for publication includes lecture notes, monographs and textbooks. All submissions will be peer-reviewed.

Hirbod Assa • Peng Liu • Simon Wang
Editors

Quantitative Risk Management in Agricultural Business

Springer

Editors
Hirbod Assa
Model Library
London, UK

Simon Wang
Stable Group Ltd.
London, UK

Peng Liu
Department of Mathematical Sciences
University of Essex
Colchester, Essex, UK

ISSN 2523-3262 ISSN 2523-3270 (electronic)
Springer Actuarial
ISBN 978-3-031-80573-8 ISBN 978-3-031-80574-5 (eBook)
https://doi.org/10.1007/978-3-031-80574-5

This work was supported by University of Essex. The open access publication of this book has been published with the support of the University of Essex's open access fund.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction to Quantitative Risk Management and Risk in Agricultural Business: Cutting Edge Quantitative Concepts and Methodologies

Hirbod Assa, Simon Wang, and Peng Liu

The collection of chapters presented here highlights the latest advancements in insurance, focusing on the integration of cutting-edge AI and statistical techniques with innovative concepts such as parametric and price insurances. These studies address a wide range of topics, from theoretical developments to practical applications and empirical results, pushing the boundaries of traditional insurance models. This introduction summarizes the key contributions of each paper, emphasizing their innovative approaches and their relevance to contemporary insurance practices.

"Index-based Insurance Design for Climate and Weather Risk Management: A Review" provides a comprehensive overview of the design principles and practical applications of index-based insurance. The chapter discusses various types of index insurance, including weather index and area yield index insurance, and their potential to offer cost-effective and scalable risk management solutions. By reviewing the successes and challenges of existing index insurance schemes, the study offers valuable recommendations for improving the design and implementation of these products to enhance their accessibility and effectiveness. This review emphasizes the innovative use of indices in insurance, presenting a framework for future developments in the field.

"Weather and yield index-based insurance schemes in the EU agriculture: A focus on the Agri-CAT fund" discusses various strategies and tools available

H. Assa (✉)
Model Library, London, UK
e-mail: h.assa@essex.ac.uk

S. Wang
Stable Group Ltd., London, UK

P. Liu
Department of Mathematical Sciences, University of Essex, Colchester, Essex, UK
e-mail: peng.liu@essex.ac.uk

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_1

to European farmers for managing weather-related risks. The chapter highlights the role of public and private insurance schemes, mutual funds, and innovative financial instruments in enhancing the resilience of the agricultural sector to climate variability. By providing a detailed analysis of policy frameworks and practical applications, the study discusses the importance of comprehensive risk management approaches in ensuring the sustainability of agriculture in the face of climate change. This research emphasizes the need for integrated risk management strategies that combine traditional insurance with innovative financial instruments.

"Avocado production index insurance: an application of credibility theory on heterogeneous data" introduces an innovative approach to designing index insurance for avocado production by applying a modified version of Bühlmann's credibility theory to handle the challenges posed by heterogeneous data from various countries. By modeling the intensity and frequency of production losses, the research offers a robust framework for calculating insurance premiums that account for variability in data quality and quantity. The study also proposes a two-layer insurance policy to mitigate over-hedging, ensuring that farmers are adequately protected against significant production shortfalls. This methodology, while focused on avocados, provides a scalable model that could be adapted to other agricultural commodities facing similar risks.

In "How do economic variables affect the pricing of commodity derivatives and insurance?" the authors investigate the relationship between macroeconomic variables, namely demand elasticity, and the pricing of commodity derivatives. This study utilizes econometric models such as GMM methods to identify the key drivers of commodity prices and their implications for derivative pricing. The insights gained from this research are crucial for developing more accurate and responsive pricing models for commodity derivatives and particularly insurance on prices in agriculture and other sectors. This chapter exemplifies the intersection of economics and advanced statistical techniques in enhancing financial instruments used in agriculture.

"Empirical Results for Cross Hedging in the Incomplete Market" explores the effectiveness of cross-hedging strategies in markets where perfect hedging instruments are not available. The study evaluates different hedging techniques, including the use of futures and options, to mitigate risks associated with agricultural commodities. The findings show the importance of selecting appropriate hedging instruments and strategies to manage price volatility and protect against adverse market movements. This chapter contributes to the understanding of risk management in incomplete markets, presenting innovative solutions for mitigating financial risks in agriculture.

The chapter "Crop Yield Insurance Analysis for Turkey: Spatiotemporal Dependence" examines the dependencies between crop yields and various risk factors across different regions in Turkey. By employing INLA as an approach developed in the last decade, via using spatial and temporal statistical models, the study identifies patterns that can improve the pricing and structuring of crop yield insurance products. This research is relevant for developing tailored insurance solutions that address specific regional risks, thereby enhancing the effectiveness and uptake

of agricultural insurance. The integration of spatiotemporal dependence in risk assessment represents a cutting-edge approach, ensuring that insurance products are both precise and equitable.

The chapter titled "Model and Forecast Combination for Predictive Yield Distributions in Crop Insurance" explores the use of combined forecasting models to improve the accuracy of yield predictions. By integrating model and forecast methods, the authors demonstrate how the new approach can enhance the reliability of yield estimates, which are critical for setting insurance premiums and coverage levels. This approach leverages the strengths of different models to produce more robust and precise yield forecasts. The use of model and forecast combination represents a significant advancement in predictive analytics for agriculture.

"A Recursive Method on Estimating ARFIMA in agricultural time series" introduces a recursive methodology for estimating Autoregressive Fractionally Integrated Moving Average (ARFIMA) models. ARFIMA models are crucial in time series analysis, especially for data exhibiting long memory properties. The recursive approach presented in this chapter enhances computational efficiency and accuracy, making it a significant contribution to the statistical techniques used in financial and insurance risk modelling in agricultural data set. Specially the use of Hurst exponent has been spelled out and properly used to model the memory in the agricultural data. This work improves the precision and speed of modeling long-term dependencies in time series data, essential for accurate forecasting in insurance and finance.

The impact of weather factors compared with other important factors on agricultural risk is meticulously analyzed in "Examining the impact of weather factors on agricultural market price risk: an XAI approach". This chapter employs machine learning models to quantify the effects of various weather conditions on risk of agricultural prices and see how the many drives of realized volatility can be compared with ENSO as the major weather risk drivers. The chapter has used Shapley values for ranking the most important features impacting price volatility. The results provide valuable insights for designing weather index-based insurance products that can offer financial protection to farmers against weather-related risks. The study enhances the predictive accuracy and reliability of agricultural insurance schemes. This research demonstrates the potential of AI and big data analytics in improving the resilience of agriculture to climatic variability.

In "Textual analysis in agriculture commodities market," the authors apply natural language processing (NLP) techniques to analyze textual data related to agricultural markets price risk. This innovative approach allows for the extraction of sentiment and thematic trends from large volumes of unstructured data, such as news articles and social media posts in modelling price volatility. The insights derived from textual analysis can inform better decision-making in commodity trading and risk management, highlighting the potential of AI-driven tools in the insurance industry. This chapter emphasizes the role of big data and AI in transforming traditional approaches to market analysis and insurance.

"Applications of Singular Spectrum Analysis in Agricultural Financial Time Series" showcases the use of Singular Spectrum Analysis (SSA) for decomposing

and reconstructing time series data. SSA is particularly effective in identifying and isolating the underlying patterns in complex time series, making it a powerful tool for forecasting agricultural prices or yields and other economic indicators. This chapter demonstrates the application of SSA in improving the accuracy of time series forecasts, which is crucial for developing reliable insurance models. The innovative use of SSA in time series analysis represents a cutting-edge approach to enhancing predictive capabilities in insurance.

The collective insights from these chapters highlight the advancement and complexity of contemporary agricultural risk management and insurance practices. Representing the cutting edge of insurance research, these studies integrate advanced statistical methods, including deep neural networks and unstructured data analysis such as NLP and text data, along with modern insurance concepts like parametric and index insurance along with price volatility risk management tools.

These chapters contribute to both theoretical and practical realms, presenting solutions that tackle real-world challenges in the insurance industry. The innovative methodologies and applications they discuss highlight the transformative potential of modern technologies and statistical techniques in agricultural risk management. With climate change increasingly affecting global agriculture and other systematic risks due to globalization, the development of sophisticated insurance products and risk management strategies is becoming ever more crucial.

Chapter 2
Index-based Insurance Design for Climate and Weather Risk Management: A Review

Wenjun Zhu, Jinggong Zhang, Lysa Porth, and Ken Seng Tan

Abstract Index insurance has become a notable risk management tool in response to increasing climate variability and extreme weather events. This chapter offers a thorough review of innovative index-based financial solutions, focusing on index insurance. It explores the essential principles of index insurance, including its actuarial framework, empirical research findings, and practical considerations. Additionally, the chapter explores future advancements in the field, emphasizing the integration of cutting-edge technologies such as artificial intelligence and blockchain. These innovations have the potential to risk modeling, underwriting and claims processing of index insurance. Aimed at researchers, practitioners, and policymakers, this chapter serves as a comprehensive guide for designing effective index insurance programs that enhance resilience in the face of climate uncertainties.

2.1 Introduction

Climate change and its associated weather risks have significantly impacted global production systems and livelihoods [83, 119]. Research by Nordhaus [119] and Hong et al. [83], highlights the severity and widespread nature of weather-related risks, which at times lead to extensive damage across various sectors, particularly in agriculture [2, 10, 13, 111, 127, 169]. Agriculture relies heavily on weather conditions, making it highly susceptible to weather-related risks [56, 133]. Adverse

We are deeply saddened that Professor Ken Seng Tan left us on January 1, 2023.

W. Zhu · J. Zhang · K. S. Tan
Division of Banking & Finance, Nanyang Business School, Nanyang Technological University, Singapore, Singapore

L. Porth (✉)
Lang School of Business and Economics, University of Guelph, Guelph, ON, Canada
e-mail: lporth@uoguelph.ca

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*, Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_2

weather is estimated to cause 70–90% of agricultural production loss [149], creating a strong impetus to hedge against weather risk. However, weather risk is systemic and relevant derivatives are very limited.

Insurance is widely used to hedge weather risk. While traditional insurance models have their complexities, including administrative costs, time-consuming settlement processes, and challenges related to adverse selection and moral hazard, they remain a vital component of risk management. To complement these models and offer a tailored solution for weather-related risks, index insurance has been developed [8, 25, 89, 115]. This alternative approach provides farmers with a means to protect against extreme and systemic weather conditions, enhancing the suite of tools available for effective risk mitigation in agriculture and beyond. Index insurance is a special type of financial contract whose payout is based exclusively on some pre-specified indices. For example, weather index insurance determines the claim payments based on future realizations of weather events determined from certain weather indices. The primary distinction between index insurance and traditional indemnity-based insurance lies in their payout mechanisms. In traditional insurance, payouts are directly tied to the actual losses experienced by policyholders. In contrast, index insurance bases its payouts on predefined indices, which should be built upon data that is both transparent and representative. This structure theoretically reduces information asymmetry issues, such as adverse selection and moral hazard, because the payout is not influenced by the specific losses of an individual policyholder but by the performance of the index. This attribute makes index insurance particularly appealing, as it addresses many of the limitations associated with conventional insurance models. Consequently, index insurance is experiencing rapid growth and increasing interest, especially in sectors like agriculture that are highly susceptible to weather-related risks [141].

Furthermore, the adoption of advanced technologies, including satellite measurements and digitalization, has significantly reduced the costs associated with index insurance. The incorporation of artificial intelligence (AI) and blockchain technology has further improved its efficiency. AI enhances risk modeling through the analysis of vast datasets, including historical weather patterns, satellite imagery, and agricultural data, leading to more accurate risk assessments [130, 131]. This not only refines risk modeling but also simplifies the claims process, enabling quicker and more transparent settlements.

Blockchain technology, known for its decentralized and transparent characteristics, offers innovative possibilities for index insurance by utilizing smart contracts to automate processes, minimizing administrative burdens and fostering trust and transparency [33, 43, 132]. Additionally, index insurance supports farmers in managing risks associated with unpredictable weather, contributing to more stable agricultural practices and economic resilience [113].

Although index insurance holds considerable promise, its uptake has been modest, presenting obstacles to achieving profitability and long-term sustainability. For example, in Canada, enrollment in index insurance programs, like forage index insurance, is notably low, estimated to be as low as 10% in some provinces.

The primary factor contributing to this low demand is basis risk, the risk that the underlying indices and actual losses are mismatched [34, 45, 89]. While challenging

to completely eliminate, basis risk can be mitigated through careful contract design. As emphasized by [22], contract design critically determines the demand for index insurance. As a result, minimizing basis risk is a pivotal strategy to enhance customer demand for index insurance. Besides basis risk, ambiguity aversion [15], farmers' past insurance payout experiences [41], and insurance literacy [18, 69], among others, are also influencing factors for index insurance demand.

This chapter provides a comprehensive exploration of the latest advancements in index-based financial mechanisms, with a focus on index insurance. It situates this innovative insurance model within the wider landscape of managing agricultural and climate-related risks, offering insights into the key considerations for developing and implementing index insurance initiatives. The chapter starts by establishing the actuarial foundation of index-based insurance, discussing the utility-based theoretical framework and the evolution of data-driven methodologies as outlined in Sect. 2.2. Section 2.3 examines the index insurance market, highlighting factors contributing to its limited uptake. Further exploration of various index-based insurance schemes is provided in Sect. 2.4, offering a comprehensive overview of this evolving field.

Additionally, this chapter not only examines current trends but also looks ahead to future developments in index insurance, particularly in Sect. 2.5, which explores the lastest technological advancements. Central to this discussion is how state-of-the-art technologies like AI and blockchain are poised to transform the way index insurance is designed and implemented. These technological breakthroughs are evaluated for their immediate benefits and their long-term potential to reshape the methodologies used in risk assessment and claims management within the context of index insurance. The overarching goal of this chapter is to offer a comprehensive understanding of the evolving field of index insurance, equipping researchers, practitioners, and policymakers with a guide to navigating the complexities of index insurance innovations and their role in enhancing resilience against climate-related challenges.

2.2 The Actuarial Foundation of Index-based Insurance

2.2.1 The Theoretical Framework

The conceptual underpinnings of index insurance design can be traced back to the seminal contributions of [3, 129]. Let us consider an insurance buyer who would like to hedge a potential loss, which is modeled by a nonnegative random variable Y, defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The objective is to design an index insurance payoff based on a p-dimensional index $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. More specifically, the goal is to determine $I(\boldsymbol{X})$, where $I : \mathbb{R}^p \mapsto \mathbb{R}^+$ is the nonnegative payoff function, and this indicates that the insurance payout is completely determined by the realization of the underlying index $\boldsymbol{X}$. In the context

of managing weather-related risks, $\boldsymbol{X}$ is a vector of weather or climate indices which are typically obtainable, transparent, and trustworthy. Denote the premium of the index insurance contract by $\pi(I)$, a functional of the indemnity function I. Then the insurer aims to design an optimal payout function I so that the policyholder's expected utility, $\mathbb{E}(U)$, is optimized under various constraints such as the budget constraint. Mathematically, the optimization problem is expressed as follows [29, 171]:

$$\begin{cases} \sup\limits_{I \in \mathcal{I}} J(I) := \mathbb{E}\big(U[w_0 - Y + I(\boldsymbol{X}) - (1-\theta)\Pi(I)]\big) \\ \text{s.t. } \Pi(I) = \pi_0, \end{cases} \tag{2.1}$$

where U is a utility function such that $U'(\cdot) \geq 0$ and $U''(\cdot) < 0$ representing the policyholder's risk aversion; w_0 is the initial wealth of the policyholder; $0 \leq \theta \leq 1$ is the proportion of insurance premium subsidized by the government; $\mathcal{I} := \{I : \mathbb{R}^p \mapsto \mathbb{R}^+ | I \text{ is measurable}\}$ defines the feasible set of indemnity function I; $\pi(\cdot)$ denotes the premium principle adopted for insurance ratemaking; and π_0 is an exogenously given premium, indicating the price level that the insurance buyer is willing to accept. In the context of agricultural insurance, the exogenous premium level also prevents abusive and speculative use of insurances, due to the popular existence of government subsidies (i.e., $\theta > 0$).

In both current practice and academic literature, the most commonly adopted premium principle for agricultural insurance is the expectation premium principle,

$$\Pi(I) = \lambda \mathbb{E}[I(\boldsymbol{X})], \tag{2.2}$$

where λ is the risk loading parameter and $\lambda \geq 1$.

When $\lambda = 1$, $\Pi(I)$ is called the actuarially fair premium. Typically, λ is strictly larger than 1, which affects the insurance premium, the insurer's profitability, and the demand for insurance among agricultural producers. In practice, it's challenging to measure and select λ. Chen et al. [29] endogenously estimate the equilibrium λ from the supply and demand curves of insurance contracts. Such endogenous insurance premium considers the strategic interaction between insureds and insurers, offering valuable market insights. More general approaches to insurance pricing are discussed in Sect. 2.2.4.[1]

The loss random variable Y in the framework represents either production loss or revenue loss. When Y is production loss, index insurance offers yield protection, ensuring compensation for diminished agricultural outputs. Conversely, if Y represents revenue loss, the insurance extends to cover risks associated with price fluctuations, providing financial compensation for variations in market prices. Currently, most available index-based insurance products in the market focus on

[1] A very useful measure for quantifying and evaluating the selection of risk loading is the relative risk loading. This relates to the notion of Marginal Indemnity Functions (MIF), which measures the increase in ceded loss per unit of increase in the global loss [4, 178].

production risk, and products such as price insurance have only more recently been introduced in the agricultural insurance market. For example, [5] investigates the viability of price index insurances in the market, demonstrating through both theoretical and practical framework that such products can achieve higher Sharpe ratios than traditional financial market indices.

Recent advances in technology and data analytics have greatly enhanced the functionality and appeal of price index insurance. For example, companies use AI techniques to empower stakeholders to make informed decisions and manage financial exposure effectively, providing precise and responsive coverage tailored to the specific needs and risks of policyholders.

It is worthwhile to note that the feasible set $\mathcal{I}$ for index insurance design generally differs from that for convention indemnity-based insurance. In indemnity-based insurance, the payout function is subject to certain conditions, such as the Lipschitz condition, to prevent moral hazard issues associated with the structure of the insurance contract [30, 31]. However, these restrictions are unnecessary for index insurance, as the underlying index $\boldsymbol{X}$ is transparent and cannot be manipulated by either the insurer or the insured. Typically the maximum payout of index insurance can sometimes exceed the largest possible loss (i.e., $\sup Y$ if it is finite), as policyholders in incomplete markets may seek to over-insure their underlying assets to hedge against large losses [50].

Assuming the existence of joint probability density $f(x, y)$, a bounded feasible set $\mathcal{I} = \{I \,|\, I : \mathbb{R}^p \to [0, M] \text{ is measurable}\}$ under some certain technical conditions (see H1 and H2 in [171] for details). [171] establish the existence and uniqueness of an optimal solution to (2.1) and categorizes solution into three different regions. An ordinary differential equation-based method and a corresponding Runge-Kutta-based numerical scheme are proposed in [171] to efficiently solve the problem in the 2-dimensional case, as stated by Theorem 2.1 below where $f(y|x)$ denotes the conditional density function given $X = x$.

Theorem 2.1 *Suppose that the derivative* $\frac{\partial}{\partial x} f(y|x)$ *exists and is continuous on* $\mathbb{R}^2$*, and a function* $\widehat{I} : \mathbb{R} \mapsto \mathbb{R}$ *solves the following ODE problem:*

$$\begin{cases} \frac{\mathrm{d}I}{\mathrm{d}x} = F(x, I), \\ \pi_0 = \gamma \mathbb{E}\left[(I(X) \vee 0) \wedge M\right], \end{cases}$$

where the function $F : \mathbb{R}^2 \mapsto \mathbb{R}$ *is defined by*

$$F(x, I) := -\frac{\int_{\mathbb{R}} U'\left(w + I - y - (1-\theta)\pi_0\right) \frac{\partial}{\partial x} f(y|x)\, \mathrm{d}y}{\int_{\mathbb{R}} U''\left(w + I - y - (1-\theta)\pi_0\right) f(y|x)\, \mathrm{d}y}.$$

Then, $I^* := \left(\widehat{I}(x) \vee 0\right) \wedge M$ *is the optimal solution to problem* (2.1).

Using an empirical case study in where temperature is selected as the underlying index for an insurance contract protecting rice yield in Jiangsu, China, [171] shows that the optimal payout function generally highly non-linear and may even be non-

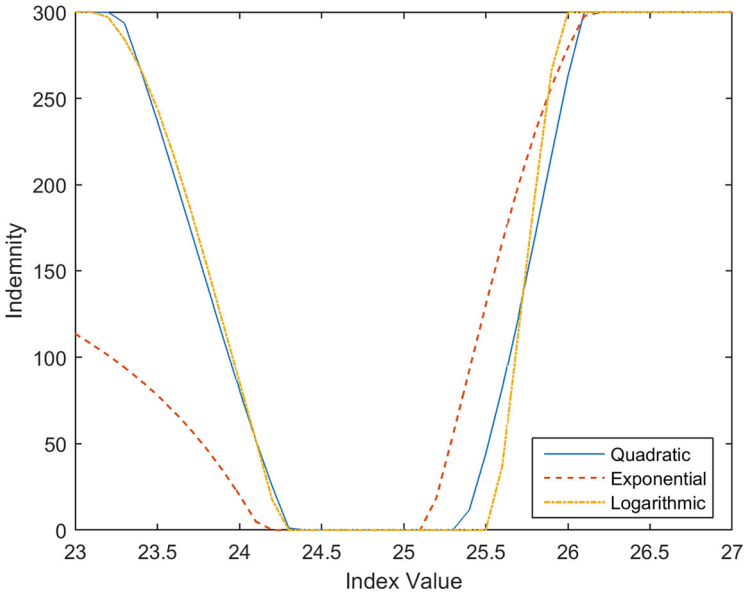


Fig. 2.1 Examples of optimal index payout based on different utility functions (i.e., quadratic, exponential, and logarithmic). (Source: Figure 7 of [171])

monotonic with respect to the index variable. This ensures that the payout aligns closely with actual loss variables, ultimately minimizing basis risk (Fig. 2.1).

2.2.2 Data-based Approaches

Due to the challenges associated with jointly modeling Y and $\boldsymbol{X}$, especially when X is highly dimensional, problem (2.1) may also be formulated in a non-parametric framework, where moment quantities are replaced by their empirical counterparts [29, 54, 142]:

$$\begin{cases} \sup\limits_{I \in \mathcal{I}} J(I) := \frac{1}{N} \sum_{j=1}^{N} \left(U[w_0 - y_j + I(\boldsymbol{x}_j) - (1-\theta)\pi(I)] \right) \\ \text{s.t.} \ \ \frac{\lambda}{N} \sum_{j=1}^{N} I(\boldsymbol{x}_j) = \pi_0, \end{cases} \tag{2.3}$$

where $\{y_j, \boldsymbol{x}_j\}_{j=1,2,\ldots,N}$ is an observed random sample and N is the sample size. In this case, further restrictions on the feasible set $\mathcal{I}_0 \subset \mathcal{I}$ should be imposed to guarantee that the solution to (2.3) is practically meaningful. $\mathcal{I}$ should be properly specified so that it strikes a good balance between model flexibility and stability (see Fig. 2.2). For example, [29] formulates and solves problem (2.3) through a neural network-based framework, and the optimal solution sacrifices some stability but

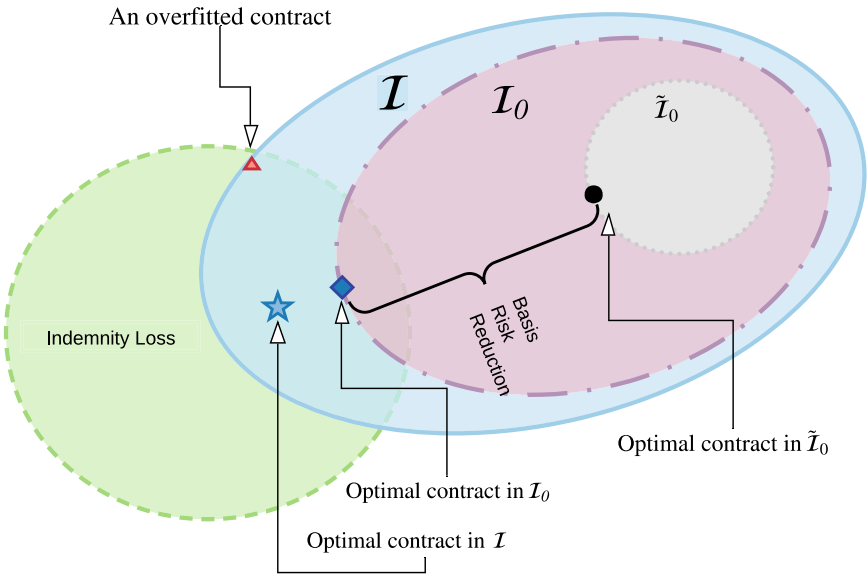


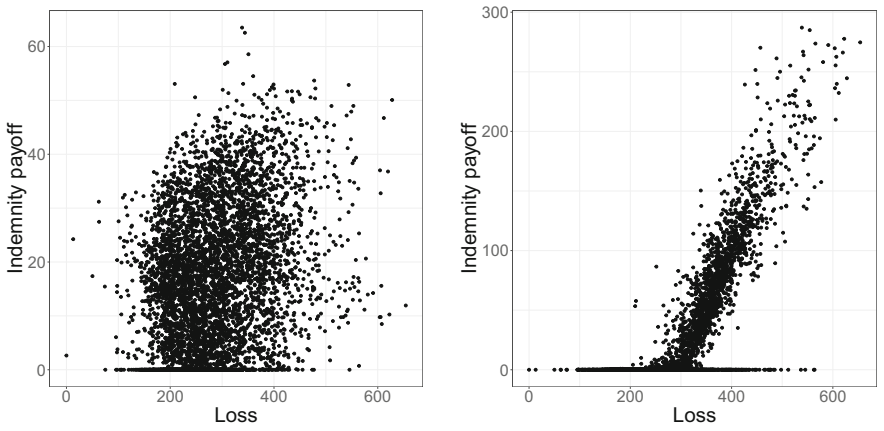
Fig. 2.2 Feasible sets and optimal contracts under a non-parametric framework. This figure compares three different feasible sets and their corresponding optimal contracts. The dashed-green circle area represents the indemnity loss, which is the actual loss experienced by the policyholder. The general feasible set, $\mathcal{I}$, is represented by the solid-blue circle area and the blue star denotes the global optimal contract. The dotted-grey circle area, $\tilde{\mathcal{I}}_0$, is a feasible set of all piecewise linear contracts. The black dot at the edge of $\tilde{\mathcal{I}}_0$ is the optimal piecewise linear contract, i.e., the contract with the smallest basis risk within $\tilde{\mathcal{I}}_0$. The dotted-blue area, $\mathcal{I}_0$, represents the specific feasible set explored in this study, and its optimal contract is denoted by the blue diamond. The red triangle illustrates an overfitted contract. (Source: Figure B.1 of [29])

achieves significantly enhanced flexibility and hence significant basis risk reduction as evidenced by Fig. 2.3.

2.2.3 *Other Optimization Objectives*

Motivated by the asymmetric and heavy-tailed nature of financial and insurance risks, [54] extends this design framework (2.1) by replacing the objective function with tail risk measures such as Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR), and Entropic Value-at-Risk (EVaR). The optimization problem is then reformulated as follows:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \rho\left(\left\{ y_i - \left[\left(\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j}\right) \vee 0\right] \wedge M + \Pi(\boldsymbol{\theta})\right\}_{i=1,2,\ldots,N}\right), \quad (2.4)$$



(a) Piecewise-linear (b) NN-based

Fig. 2.3 Insurance payoffs against actual losses to illustrate the basis risk for: Panel (**a**) a piecewise-linear insurance contract based on a single temperature index; Panel (**b**) an neural network-based index insurance contract proposed by [29]. (Source: Figure 1 of [29])

where $\boldsymbol{\theta}$ is a parametric representation of the insurance payout function. For any $\alpha \in (0, 1)$, the empirical measures of VaR, CVaR, and EVaR are given, respectively, by

$$\text{VaR}_\alpha\left(\{Y_i\}_{i=1,2,\ldots,N}\right) = Y_{k_N:N},$$

$$\text{CVaR}_\alpha\left(\{Y_i\}_{i=1,2,\ldots,N}\right) = \frac{(k_N - \alpha N)\, Y_{k_N:N}}{N(1-\alpha)} + \frac{1}{N(1-\alpha)} \sum_{i=k_N+1}^{N} Y_{i:N},$$

and,

$$\text{EVaR}_\alpha\left(\{Y_i\}_{i=1,2,\ldots,N}\right) = \inf_{t>0} \left\{ t^{-1} \ln\left[\frac{\sum_{i=1}^{N} e^{tY_i}}{N(1-\alpha)} \right] \right\},$$

where $k_N = \lceil \alpha N \rceil$ denotes the smallest integer greater than or equal to αN, and $Y_{i:N}$ represents the i-th order statistic from a random sample of size N from population Y. The main challenge to numerically solve (2.4) is that the objective function may not be convex or differentiable. Fan et al. [54] show that the objective function, when formulated with VaR, CVaR, and EVaR, is at least continuous with respect to θ. To address the numerical difficulties, they propose using a model-based annealing random search method. This extended framework is of great practical relevance for corporate farming operations and the internal risk management strategies of insurance companies [29].

2.2.4 Pricing Framework with Premium Principles

A crucial element in any successful agricultural insurance program, including index insurance, is the establishment of actuarially fair and sustainable premium rates [6, 157]. Setting appropriate premiums is essential yet challenging. Failure to set sound premiums can lead to various issues, such as stunted growth in agricultural insurance programs and insufficient reinsurance capacity. Even minor improvements in the precision of premium rates can lead to significant benefits, particularly in reducing government spending and ensuring the long-term viability of these programs [36, 121]. A pivotal step in establishing a robust framework for insurance pricing involves selecting an appropriate premium principle. Premium principles serve as a quantitative approach to pricing risk. This subsection introduces key premium principles used in agricultural insurance.

Consider an insurable risk, Y, which is often defined as a non-negative random variable, with its cumulative distribution function $F_Y(y)$, survival function $S_Y(y)$, and probability density function $f_Y(y)$. In agricultural insurance, the insurable risk can be crop production losses, revenue losses, or other financial risks. We denote the collection of all nonnegative random variables as a set $\mathscr{Y}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Mathematically, a premium principle is a functional Π assigned to an insurable risk Y.

In order for a premium principle to reflect the underlying riskiness of an insurance exposure, it should possess some desirable properties for it to be actuarially sound [168]. The key properties of a premium principle can be summarized as follows:

1. Positive risk loading: $\Pi(Y) \geq \mathbb{E}(Y)$ for all $Y \in \mathscr{Y}$. This property requires that the premium charged for insuring the risk is no less than the expected payout.
2. No unjustified risk-loading: For a degenerate risk Y, i.e. there exists a constant c such that $\mathbb{P}(Y = c) \equiv 1$, then $\Pi(Y) = c$. This property implies that if a risk results in a constant loss of c for certain, then the corresponding insurance premium should assign no risk loading.
3. No ripoff: $\Pi(Y) \leq \operatorname{ess\,sup}(Y)$ for all $Y \in \mathscr{Y}$. This property ensures that the insurer should not charge higher than the maximum loss of the risk.
4. Translation invariance: $\Pi(Y + a) = \Pi(Y) + a$ for all $Y \in \mathscr{Y}$ and $a \geq 0$. If a risk Y is increased by a constant amount a, then the premium for the combined risk $Y + a$ should just be the premium of the original risk plus a.
5. Scale invariance: $\Pi(aY) = a\Pi(Y)$ for all $Y \in \mathscr{Y}$ and $a \geq 0$. This property is also known as homogeneity of degree one in economic literature, and its significance is to avoid arbitrage opportunities. For example, the premium for $2Y$ should correspond to the premium for two insurance policies for the risk Y, otherwise, there is a possibility of arbitrage.

   Combining Property 4 and Property 5 implies linearity.
6. Subadditivity: $\Pi(Y_1 + Y_2) \leq \Pi(Y_1) + \Pi(Y_2)$ for all $Y_1 \in \mathscr{Y}$ and $Y_2 \in \mathscr{Y}$. Subadditivity requires that insuring a portfolio of homogeneous risk should be less expensive than insuring individual risks separately because of the risk

pooling and diversification benefit. This property implies that the insurer cannot benefit from dividing risk into pieces.
7. First stochastic dominance (FSD) preserving: If $S_{Y_1}(y) \leq S_{Y_2}(y)$ for all $y \geq 0$, then $\Pi(Y_1) \leq \Pi(Y_2)$.
8. Stop-loss (SL) order preserving: If $\mathbb{E}\big[(Y_1 - d)_+\big] \leq \mathbb{E}\big[(Y_2 - d)_+\big]$ for all $d > 0$, then $\Pi(Y_1) \leq \Pi(Y_2)$. Here $(x)_+ = \max(x, 0)$.

While there are a large number of premium principles, some commonly used ones include the following:

(a) *Expectation Premium Principle:* $\Pi^e(Y) = (1 + \theta)\mathbb{E}(Y)$, where $\theta > 0$. This is the most widely used premium principle in agricultural insurance ratemaking, including index insurance pricing, due to its simplicity.

(b) *Standard Deviation Premium Principle:* $\Pi^{sd}(Y) = \mathbb{E}(Y) + \theta\sqrt{\mathbb{V}(Y)}$, where $\theta > 0$ and $\mathbb{V}(Y)$ is the variance of the random variable Y. This premium principle incorporates a risk loading that is proportional to the standard deviation of the insured risk. While widely used in general P&C insurance, this premium principle has received little attention in agricultural insurance, with the exception of the work by [124], which empirically analyzed the optimal reinsurance contract structure for Manitoba Agricultural Services Corporation (MASC) in Manitoba, Canada.

(c) *Esscher Premium Principle:* $\Pi^{ess} = \dfrac{\mathbb{E}(Ye^{\theta Y})}{\mathbb{E}(e^{\theta Y})}$, where $\theta > 0$. This premium principle, which is based on *Esscher Transform*, is widely used in option pricing in the context of incomplete markets [17, 70, 86]. Therefore, this premium principle is applicable for weather derivative pricing. In addition, it is a special case of the *Equilibrium Premium Principle* proposed by [16]. Additionally, a more general form of the Esscher premium principle is referred to as the *Exponential Tilting Premium Principle* [80, 91].

(d) *Distortion Premium Principle:* For any increasing concave function $g : [0, 1] \mapsto [0, 1]$ with $g(0) = 0, g(1) = 1$, the premium is calculated as $\Pi^d(Y) = \int_0^\infty g\big(S_Y(u)\big)du$ [153, 156]. The function g is called the *distortion* function and $g\big(S_Y(u)\big)$ is called the distorted probability. A special case of this premium class is called *Proportional Hazards Premium Principle* [152].

(e) *Multivariate Weighted Premium Principle (MWPP):* The MWPP [177] is defined based on the weighted density transform: $\mathfrak{T}^w : (Y, \boldsymbol{X}) \to Y^w$, associated with the random vector $\boldsymbol{X}$ and the weighting function $w : [0, \infty] \mapsto [0, \infty]$, with $\mathbb{C}\big(Y, w(\boldsymbol{X})\big) \geq 0$. More formally, $\Pi^w(Y, \boldsymbol{X}) = \mathbb{E}(\mathfrak{T}^w) = \dfrac{\mathbb{E}\big(w(\boldsymbol{X})Y\big)}{\mathbb{E}\big(w(\boldsymbol{X})\big)}$. The MWPP is a general premium principle that depends not only on the underlying risk Y, but also the auxiliary factors $\boldsymbol{X}$ and the weighting function w. MWPP adds flexibility to the premium principle, including other existing premium principles as special cases, depending on the choices of $\boldsymbol{X}$ and w, including the univariate weighted premium in [62], the Esscher premium and the Equilibrium Premium Principle. Empirically, one can choose random vector

$\boldsymbol{X}$ to reflect systemic and often geographically correlated climate risks and adverse weather conditions. In addition, from an asset allocation point of view, the resulting economic weighted pricing functional Π incorporates a number of risk capital allocation rules, such as those based on modified tail covariance and exponential tilting [60, 61, 63, 64, 154].

2.2.5 Data Scarcity

Data availability is a significant challenge in agricultural insurance, primarily because loss experience data is often scarce and concerns frequently arise regarding its credibility [37, 38, 125, 126]. There are many contributing factors to data scarcity. First, in most regions, there is only one crop growing season per year, therefore, there is only one yield (or loss) observation per year. This means that approximately 30–40 years of annual historical loss observations can be used for rating products. Further, due to crop rotation and other market forces farmers do not grow the same crop each year, and this leaves fewer observations or an inconsistent time series at the farm-level.

In addition to limited data, there can also be concerns over the credibility of data. For example, older loss experience may not be as relevant today due to program modifications, technological advancements, deviations in farming practices, changes in climate, etc. [120, 155, 159]. The result is a need to balance using as much of the time series as possible so that those infrequent, but, extreme weather events are considered, versus the concern that older data is not representative and, therefore, should be discarded. There are two possible ways to blance this incongruity. First, we can "restate" historical yield or loss experience to bring it on level with the current environment so that as many of the older observations can be used as possible [12, 37, 126, 157, 158, 176, 177]. Alternatively, we can borrow information from regions with better data quality. For example, [128] develops a relational model to predict farm-level crop yield distributions in the absence of farm-level yield data.

From a data availability and data quality point of view, index insurance offers several advantages over traditional indemnity-based agricultural insurance. Weather index insurance policies rely on readily available weather variables, such as rainfall, temperature, and satellite images. This information is often more accessible and less susceptible to manipulation or fraud compared to individual farm-level data required for traditional agricultural insurance [160]. Moreover, from a statistical inference viewpoint, the modelling and pricing of index-based policies may face less challenges, since large volumes of reliable and extensive weather data records are typically available in daily frequency, facilitating more sound statistical modeling.

2.3 The Demand and Supply of Index Insurance Market

Despite its potential benefits, the adoption of index insurance programs in practice has been lower than anticipated. Therefore, most index insurance programs around the world remain at the pilot stage and struggle to scale up. This persistent lack of demand is a recognized concern in existing literature, and in particular, an important empirical puzzle is that the most risk-averse farmers display particularly low interest in index insurance [40, 72–74]. This paradox has driven researchers to investigate various factors contributing to the low demand for index insurance, exploring both behavioral and structural barriers that limit its widespread adoption.

2.3.1 Basis Risk

Cited as the most important factor that contributes to the low demand of index insurance programs in the literature [25, 34, 45, 47, 74, 89], basis risk refers to the risk of contractual nonperformance, i.e., the underlying indices and actual losses are mismatched, and hence could lead to situations in which farmers are not indemnified for actual losses, or are paid indemnities despite having no actual losses. There are three primary sources of basis risk in agricultural index insurance [173]. The first is *Variable Basis Risk*. When weather variables used for hedging and the loss exposures originate from the same geographic region, yet their correlation remains imperfect. The mismatch occurs due to the nonlinear relationship between weather conditions and actual losses. A high-dimensional weather index system and improved modeling of nonlinear relationships can help mitigate this type of basis risk [25, 102, 176]. The second type is *Spatial Basis Risk* occurs when the geographic location of the underwriting risk exposure differs from where the weather indices are recorded. Mobarak and Rosenzweig [113] highlight the significance of spatial basis risk, showing that placing rainfall gauges in specific villages can significantly increase rainfall insurance demand. Enhancing spatial and temporal modeling of weather variables is critical for reducing spatial basis risk and improving the effectiveness of weather insurance programs [175]. The last type is *Temporal Basis Risk*. It arises when there is a mismatch between the timing of the occurrence of agricultural losses and the timing of the index-insurance coverage, leading to potential gaps or inadequacies in coverage.

Minimizing basis risk is crucial because even if premium rates are actuarially sound, farmers may still lack full coverage if basis risk is present [112]. In a theoretical framework, [34] analyzes the impact of basis risk to index insurance demand, where basis risk is defined as the joint probability that the producer incurs a loss but receives no payment from the index insurance contract. Insurance demand is found to decrease with increasing basis risk but exhibits a non-monotonic relationship with risk aversion. In particular, for an infinitely risk averse producer, index insurance demand is zero, because of the high disutility associated with

contract nonperformance—that is, the scenario where basis risk is high and the insurance fails to provide indemnity despite actual losses. These results imply that an increase in risk aversion does not necessarily lead to an increase in demand for index insurance [51].

A few studies have empirically assessed the impact of basis risk on index insurance demand. Collaborating with a private insurance company in India—HDFC ERGO General Insurance, [82] finds that demand falls with basis risk. Jensen et al. [89] examine the distribution of basis risk associated with the Index Based Livestock Insurance(IBLI) product in Kenya [25].

2.3.2 Prospect Theory

Prospect theory [90, 148] presents an alternative approach to conventional expected utility theory. It sheds light on the negative relationship observed in the literature between risk aversion and index insurance demand. Prospect theory assumes that the utility function is convex for losses and concave for gains, and that low probability events are overweighted. These assumptions imply that index insurance policies become very valuable for covering extreme losses under prospect theory. However, when index insurance fails to provide coverage for certain extreme events, the insurance is perceived as significantly less valuable, since these worst-case scenarios are overweighted the most. This results in a disproportionate amount of disutility when coverage fails, contributing to low demand for index insurance.

Research by [151] conducts willingness-to-pay surveys in the U.S., revealing that insurance demand decreases by more than 20% when a policy has just a 1% chance of contractual nonperformance. This suggests that insurance is highly valued when it gives the impression of completely eliminating a risk, rather than just reducing it. Moreover, [35] conducted experiments with index insurance in rural Ethiopia and found that demand surpassed predictions made by expected utility theory. Their results align more with an S-shaped probability weighting function, where extreme events are underweighted, contrary to prospect theory's assumption of overweighting such events.

In a recent study by[134], the authors analyzed index insurance demand within a prospect theory framework considering basis risk. They confirmed a negative relationship between loss aversion and insurance demand, particularly emphasizing the pronounced impact of loss aversion on farmers exposed to high basis risk. Their findings from a lab-in-the-field experiment in Kenya showed that a one standard deviation increase in loss aversion reduced desired insurance coverage by 20% relative to the mean condition for high basis risk index insurance, while the effect was marginal for low basis risk index insurance.

2.3.3 Ambiguity Aversion

Ambiguity aversion, as introduced by [53], presents an alternative perspective on index insurance demand. This theory assumes that an ambiguity-averse individual dislikes being uncertain about the distribution of outcomes rather than simply disliking that outcomes are uncertain. This aversion to ambiguity can significantly impact the demand of index insurance, especially in scenarios involving uncertainty about outcome distributions, technological adoption, basis risk, and government intervention in premium subsidies.

The study by [15] highlights how ambiguity aversion intersects with the adoption of new technologies, creating uncertainty that may hinder the demand of index insurance. For an ambiguity-averse individual, assessing insurance involves considering the worst-case scenarios, leading to a perception of lower value in the insurance contract. Analyzing data from randomized controlled trials conducted in Malawi and Kenya [72], [15] show that the negative impact of risk aversion on insurance demand primarily stems from ambiguity-averse individuals. Interestingly, individuals not averse to ambiguity show an increasing demand for insurance as predicted by standard theories. Chi et al. [32] provides a theoretical framework to explain the low index insurance demand from an ambiguity aversion perspective. This model predicts that an ambiguous individual, who knows the marginal distributions of the crop yield but lacks information about dependence structures, would choose not to purchase index insurance even if the premium is lower than the actuarially fair level, implying that government subsidies are essential to encourage participation. With government subsidies, the participation rate becomes positive but decreases as the insurance premium, risk aversion coefficient, and volatility of index payouts increase. In the case that farmer has more information about the correlation between the index and crop yield, they may purchase some index insurance even without subsidies, and additional information further increases demand.

2.3.4 Complexity Aversion

In practice, contract complexity often acts as a deterrent, reducing participation in insurance markets, i.e., complexity aversion [9, 26, 87, 88, 95, 138, 139]. This issue is especially pronounced in health insurance markets [21, 143]. For index insurance, there exists a delicate balance between basis risk and the complexity of the insurance contract. While a more flexible contract offers better protection, it tends to be significantly more intricate for farmers compared to a simpler contract. The complexity of a contract that exceeds farmers' understanding heightens their perceived uncertainty about the insurance payout [96]. Farmers' aversion to complexity is often reflected in their struggle to comprehend insurance contracts, where greater complexity translates to a larger variance in index insurance payouts, creating difficulties in contract understanding [96].

To alleviate farmers' concerns over contract complexity, enhancing the interpretability of index insurance becomes crucial, fostering better comprehension and trust among farmers [29]. In addition, literature suggests that education significantly bolsters insurance demand [18, 69, 75, 82]. For example, [82] find empirical evidence that incentivizing learning or learning by doing is more effective in improving both understanding and demand for insurance. Moreover, government subsidies can play a pivotal role in enhancing communication and trust between insurance companies and farmers through public-private partnerships (PPPs). By subsidizing aspects like education, outreach, and contract simplification efforts, these partnerships can alleviate the complexity burden and foster greater understanding and acceptance of insurance products among farmers.

2.3.5 Other Factors

The low uptake of index insurance contracts is influenced by various other factors. For example, liquidity constraints and differing intertemporal preferences impact farmers' ability and willingness to invest in insurance [11, 74]. The lasting effects of premium subsidies shape the dynamics of insurance demand over time [82]. Moreover, farmers' past experiences with insurance payouts significantly influence their future decisions regarding insurance adoption [18, 41, 82]. Asymmetric information between insurers and farmers impacts participation in index insurance programs [79]. The level of insurance literacy, education, and awareness among farmers also plays a pivotal role in their understanding and acceptance of index insurance products [18, 69, 75, 82]. Additionally, farmer heterogeneity contributes to varying preferences and behaviors regarding index insurance demand [24]. Finally, the effectiveness and accessibility of marketing and distribution channels used to promote and deliver index insurance products are crucial factors influencing their adoption [39, 110, 172]. See [123] for a comprehensive literature review of the demand for microinsurance, including index insurance.

2.4 Index Insurance Programs

2.4.1 Weather Index Insurance

Thus far, most index insurance products available in the market have taken the form of weather-based, meaning that payouts are based on weather station measurements.[2] For example, rainfall and temperature-based index insur-

[2] Weather derivatives are a concept closely related to weather index insurance in weather risk management. While they differ in regulatory, accounting, tax, and legal aspects, both serve

ance have been the primary focus of most practical implementations. However, in many cases there are limitations in terms of the density of weather stations, and possibly the weather data availability and credibility from each station, especially in developing countries. Therefore, alternative weather variables, such as wind and sunshine, can be considered to develop weather index insurance programs. [3]

Consider rainfall index insurance as an example, where the index insurance payout, I, can be represented as follows,

$$I(X) = \begin{cases} I_m & 0 \le X \le \tau_1 \\ \dfrac{-X - \tau_2}{\tau_2 - \tau_1} & \tau_1 < X \le \tau_2 \\ 0 & X > \tau_2. \end{cases}$$

Here X is the rainfall variable, τ_1 and τ_2 are two triggers, and I_m is the maximum insurance payout. The resulting indemnity payment from the index insurance contract has the step-wise form. The step-wise loss cost shape of such index-based insurance can be justified by the typical positive relationship between agricultural yield and developed indices such as rainfall.

A critical consideration in designing a viable contract based on a single index is a careful selection of the weather variable and a precise determination of triggers, τ_1 and τ_2, to mitigate basis risk. To assess the effectiveness of the proposed index, a straightforward approach is to compare the index with actual loss data. The degree of their alignment can provide evidence of efficacy of the proposed index. Figures 2.4 and 2.5 illustrate examples of correlation analyses between rainfall index and temperature index, respectively, and their relationship with agricultural losses.

While piece-wise linear contracts, e.g., Eq. (2.5), widely used in both the literature and practical applications [73, 105, 150, 160], their functional form is inherently limited in accommodating nonlinearity and high-dimensionality. Chen et al. [29] compare a single-index piece-wise linear contract with a highly nonlinear counterpart constructed using a high-dimensional weather index system, and find that the former exhibits significantly larger basis risk and inferior efficiency in downside risk reduction and welfare improvement.

Indeed, most of the pilot weather index insurance programs have faced limited adoption and concerns about commercial sustainability. Barnett and Mahul [8]

as instruments for transferring weather risks and share a similar mathematical foundation. Transactions involving producers in developing countries often take the form of insurance, whereas derivatives are typically designed for large-scale buyers and can play a role in reinsurance markets to offset weather risks. This includes their application in the international reinsurance market for traditional agricultural insurance, where weather risk remains a significant exposure [5, 144, 145, 147, 160].

[3] Many studies have examined various interpolation techniques, such as the Inverse Distance Weighting (IDW) approach and the kriging method, to address the limitations of weather station data [20, 28, 49, 101, 140].

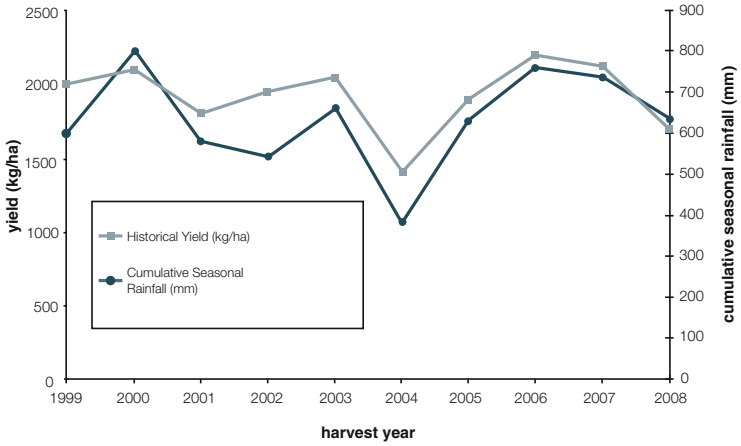


Fig. 2.4 Correlation analyses of rainfall index versus maize yields in Alaba Wereda. (Source: Figure 5.3 of [160])

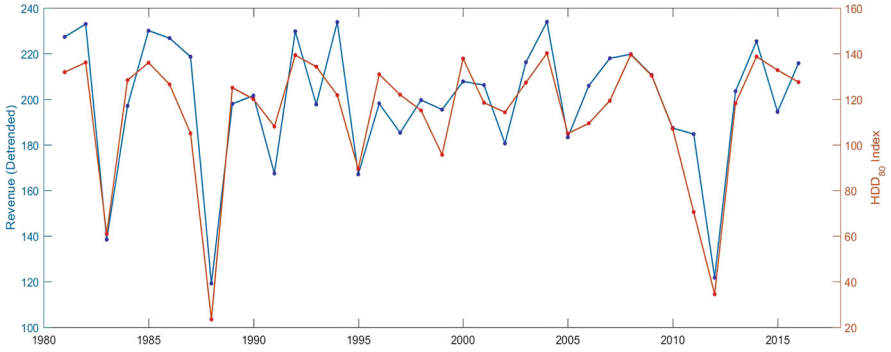


Fig. 2.5 Correlation analyses of temperature index versus corn yields in the US. (Source: Authors)

present an early overview of weather index insurance in developing countries. World Bank [160] provides a comprehensive introduction to weather index insurance and its research development efforts, shedding light on the technological and practical barriers faced by developing countries in creating effective weather index insurance products. Smith and Watts [137] broaden the scope by studying index insurance feasibility, scalability, and sustainability. Carter et al. [22] conduct an empirical review of weather index insurance in developing countries, highlighting the low take-up issue and corresponding actions to increase it.

Rainfall Insurance in India In 2003–2004, one of India's largest private general insurance companies, in collaboration with the World Bank, pioneered the country's first weather index insurance product. This policy's payout structure was based on rainfall deficits and was linked to crop loans provided to farmers through the

Microfinance Institution (MFI) BASIX. Inspired by this initiative, the state-owned Agricultural Insurance Company of India (AICI) introduced rainfall insurance and the Weather-Based Crop Insurance Scheme (WBCIS), making India the world's largest weather index insurance market. Despite challenges in the actuarial performance, subsequent modifications and the growth of WBCIS demonstrate the evolving landscape of index insurance in India. Recommendations from the World Bank [106] continue to guide improvements aimed at enhancing the program's efficiency and resilience.

Caribbean Catastrophe Risk Insurance Facility The Caribbean Catastrophe Risk Insurance Facility (CCRIF) is a risk-pooling mechanism operated by the Caribbean governments. Linking to natural disasters, particularly hurricanes and earthquakes, CCRIF is established to mitigate the negative impact of catastrophes by providing quick and reliable financial assistance to Caribbean countries. The facility operates on a parametric insurance (i.e., index insurance) model, where payouts are triggered based on pre-defined parameters, such as the magnitude of an earthquake or the intensity of a hurricane. This parametric approach allows for swift and transparent payouts, enabling member countries to access funds within a short time frame after a covered event.

El Niño—Southern Oscillation (ENSO)-induced Rainfall Index Insurance In Northern Peru, the occurrence of El Niño events contributes to excessive rainfall and catastrophic flooding, resulting in significant damages to crop and infrastructure. These adverse conditions, in turn, impair borrowers' ability to meet loan obligations. To address these challenges, an innovative index insurance product has been developed, using Pacific Ocean surface temperatures as a key indicator. Elevated temperatures signal the onset of El Niño conditions triggering insurance payouts. By offering a financial safeguard against losses incurred during these weather events, microfinance institutions (MFIs) are motivated to expand agricultural lending and provide improved rural financial services. Programs like this not only mitigate the financial impact of El Niño-induced losses but also enhance community resilience by fostering climate-adaptive financial solutions [42].

Malaysian Solar Energy Shortfall Insurance The Solar Energy Shortfall Insurance (SESI) program offers parametric insurance coverage for solar energy operators in Malaysia, offering index-triggered protection against financial losses resulting from lack of sunlight. Launched by a Malaysian insurer and backed by a global reinsurer, SESI aims to enhance investor confidence and support financial institutions, thereby encouraging greater investment in solar energy infrastructure. The involvement of reinsurers in SESI's development highlights the collaboration between insurers and reinsurers to develop innovative risk transfer solutions for the renewable energy sector. The index utilized in SESI can be seamlessly integrated as a time-saving automated tool for various solar parametric insurance needs, including pricing, reinsurance acceptance, and provisional claims compensation assessments. SESI aligns with Malaysia's national goal of increasing its renewable energy mix and contributes to the broader effort in mitigating climate risk through sustainable energy solution.

2.4.2 Satellite-based Index Insurance

As previously noted, the limitation of weather station data arises from the sparse distribution of weather stations. Despite the various interpolation techniques that can be used to help address the situation of limited or missing data, a remaining problem is that the low density of stations has been proven to systemically underestimate extreme values, which are precisely those extreme events that the insurance program is intended to cover. Consequently, the effectiveness of weather station-based index insurance programs may be compromised.

Recent advancements in satellite-based remote sensing offer a transformative solution for index-based insurance, leveraging publicly available and transparent "big data." This technology presents a potential avenue for reducing basis risk, thereby enhancing the relevance and effectiveness of index insurance policies. Satellite-based crop yield estimation has been improving over time, and will continue to improve rapidly with advances in satellite technology. Satellites continue to improve with more bands, better sensors, and better resolution. Also, software and image processing capability continues to improve, along with more computing power (e.g. cloud computing), and more data storage is available to deal with big data at lower cost. As well, advances in deep learning models can improve the computing and processing capabilities, enabling more precise yield estimations.

Studies show that indemnities from insurance contracts based on satellite data exhibit higher correlations with actual yield losses resulting from droughts, compared to conventional weather indexis such as rainfall [107]. As a result, remote sensing-based index insurance has gained popularity for monitoring pasture productivity and providing coverage for livestock losses [25, 89, 107, 161, 162].

In general, there are two main types of remote sensing index approaches to estimate crop yield, including vegetation indices and biophysical variable indices. The most commonly used vegetation index is Normalized Difference Vegetation Index (NDVI) [114, 122, 146], which is widely used for index-based insurance design. Several operational commercial grassland insurance programs have adopted this approach, including in Spain, Mexico, Canada and the U.S. NDVI is computed from visible (VIS) light and near-infrared (NIR) light, reflected by vegetation. It is calculated as the normalized difference between NIR and VIS, expressed as:

$$NDVI = \frac{(NIR - VIS)}{(NIR + VIS)}. \tag{2.5}$$

NDVI ranges from -1 to 1, with higher NDVI indicating greener vegetation (higher yield), and lower NDVI indicating less green vegetation (lower yield). It is commonly well above zero for moderate vegetation, and nearer to 1 for very dense vegetation. NDVI will be typically slightly above zero for no vegetation (bare soil), and may be negative for clouds, snow, or water, which have high reflectance.

Despite its widespread use, NDVI poses challenges, such as those associated with soil reflectance. Consequently, researchers have developed numerous enhanced

methods building upon NDVI for more accurate yield estimation [114]. These improved versions include the Green Normalized Difference Vegetation Index (GNDVI), Enhanced Vegetation Index (EVI), Modified Soil Adjusted Vegetation Index (MSAVI2), and Optimized Soil Adjusted Vegetation Index (OSAVI). These refined indices aim to address limitations and enhance the accuracy of vegetation-related assessments in diverse environmental conditions.

An alternative to a vegetation index is an approach that uses biophysical variables (parameters). Two widely used biophysical variables are Leaf Area Index (LAI) and the Fraction of Photosynthetically Active Radiation (FPAR), which can be leveraged for yield estimation. LAI measures the ability of the plant to absorb sunlight that reflects photosynthesis and can be used to indicate crop yields. Related to LAI is FPAR, which refers to the fraction of absorbed PAR (APAR), to incoming photosynthetically active radiation (PAR), i.e., FPAR = APAR/PAR, and is between 0 and 1. The biophysical parameter values are usually obtained by fitting a semi-empirical model to spectral data. Studies suggest that approaches based on biophysical variables outperform vegetation indices, such as NDVI, particularly in the context of quantifying biomass [7, 19]. Brock Porth et al. [14] compared various production indices, including those derived from satellite-derived vegetation and biophysical parameters, evaluating their performance in constructing index-based insurance and their efficacy in reducing basis risk.

2.4.3 Area-yield Index Insurance

Area-yield insurance represents another type of index insurance program. Initially designed for soybean farmers in specific U.S. counties, the program has evolved over time, extending its coverage to encompass major commodities such as corn, wheat, and cotton. A distinctive feature of area-yield insurance policies is that indemnity payments are determined by the average yield within a county rather than the losses incurred by individual farms. This approach effectively mitigates issues related to information asymmetry, as the actions of any single producer are unlikely to have a substantial impact on the overall yields at the county level.

Area-yield insurance organizes a group of producers into K distinct risk pools. Consider producer i operating within the kth risk pool. Under the protection of area-yield insurance, this producer is eligible to receive an indemnity payment if the area yield of the kth risk pool, denoted as y_k, falls below a pre-established threshold, $\bar{y}_c$. The indemnity function, applicable universally within this pool, is formulated as follows:

$$I_k = \max\left(\bar{y}_c - y_k, 0\right) \times scale, \tag{2.6}$$

where $\bar{y}_c$ is the critical yield, calculated as

$$\bar{y}_c = \mu_k \times coverage, \tag{2.7}$$

and μ_k represents the expected area yield level, that is, $\mu_k = \mathbb{E}[y_k]$. $coverage$ and $scale$ offer producers additional flexibility. Typically farm-level volatility is higher than the county-level. Increasing their $scale$ and $coverage$ ensure that producers can have sufficient coverage in years when production significantly drops.

[109] establishes a connection between individual yield y_i, and area yield y_k. By projecting the producer's individual yield y_i onto the area yield y_k, y_i can be modeled as follows:

$$y_i = \mu_i + \beta_i \cdot (y_k - \mu_k) + \epsilon_i. \tag{2.8}$$

Here, Eq. (2.8) decomposes individual yield variation into a systemic component $\beta_i \cdot (y_k - \mu_k)$ that is perfectly correlated with the area yield and a non-systemic component ϵ_i that is uncorrelated with area yield. The coefficient $\beta_i = \mathbb{C}(y_i, y_k)/\sigma^2_{y_k}$ quantifies the sensitivity of producer's individual yield to the systemic factors that affect the area yield.[4]

Typically area-yield index-based insurance programs, such as the Area Risk Protection Insurance (ARPI) in the US, defines risk pools based on county boundaries. While this method provides a structured approach to risk assessment, it has limitations. As [136] point out, relying on county yields for an area-index may not be ideal. They argue that county boundaries often fail to accurately group together producers with similar year-to-year percentage deviations from forecasted yields. This mismatch can lead to a misalignment between the actual risk profiles of individual producers and the broader risk pool defined by these administrative boundaries.

To minimize basis risk, payouts should be based on average yields of smaller areas. This might however introduce moral hazard. On the other hand, when payouts are based on average yields of a bigger area, moral hazard is limited but basis risk is higher. Therefore, in area-yield index insurance design, it is essential to determining the risk pools by selecting the optimal number of risk pools to achieve a trade-off between reducing basis risk and mitigating moral hazard, and grouping together the producers with similar risks. Xu et al. [166] introduced a framework to optimize sustainable risk pooling in area-yield insurance using behavior-based machine learning. This method not only significantly reduces contract basis risk and mitigates producers' tail risk, but also offers geographical and economic insights. Figure 2.6 illustrates the risk pooling outcomes for the state of Illinois. Elabed et al. [52] propose a multi-scale index insurance approach to enhance risk diversification and improve insurance efficiency. Under such a contract, payouts are based on average yields at multiple levels, for instance at both the village and the regional level. Farmers who collude to reduce yields in the absence of any shock, will not receive a payout. In this way, the village trigger ensures low levels of basis risk while the regional one addresses moral hazard. While not fully eliminating either of

[4] Given that the area-yield insurance contract aims to hedge the systemic risk faced by producers, β_i needs to be positive to ensure the contract's effectiveness.

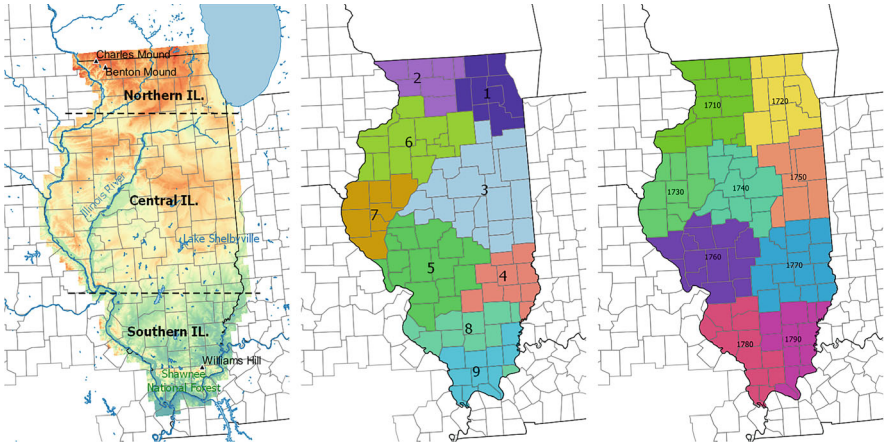


Fig. 2.6 Illinois risk pooling visualization. Illinois can be divided into three distinct geographical regions from north to south: Northern Illinois, including the Chicago Municipal Area and surrounding Charles Mountain; Central Illinois, known for its prairie landscapes and the Illinois River; Southern Illinois, distinguished by its warmer climate and location between the Mississippi and Ohio Rivers. The risk pooling in the proposed approach (Fig. 2.6b captures this geographical insights). (**a**) Geography map. (**b**) Risk pooling result. (**c**) Agricultural district. (Source: Figure 4 of [166])

the two problems, this demonstrates how careful contract design can be utilized to enhance the quality of index insurance.

2.4.4 Blended Index Insurance

Motivated by the distinct advantages of both conventional indemnity-based insurance and index insurance under different real-world scenarios, [170] proposes the concept of "blended insurance", which seeks to combine the benefits of both approaches to achieve greater cost efficiency and enhanced risk mitigation. Theoretical results demonstrate is that blended insurance can provide more effective risk mitigation than either conventional insurance or index insurance alone. The payout function for blended insurance, $\pi(I_b(Y, \mathbf{X}))$, is defined as a combination of the payout structures from both conventional indemnity-based insurance and index insurance, offering a more balanced and comprehensive risk management solution:

$$I_b(Y, \mathbf{X}; t) := I_c(Y) \cdot 1_{[T(\mathbf{X})>t]} + I_i(\mathbf{X}) \cdot 1_{[T(\mathbf{X})\leq t]}, \tag{2.9}$$

where $I_c(Y)$ and $I_i(\mathbf{X})$ represent the conventional and index components of the blended payout function, respectively; $1_{[\cdot]}$ is the indicator function that determines

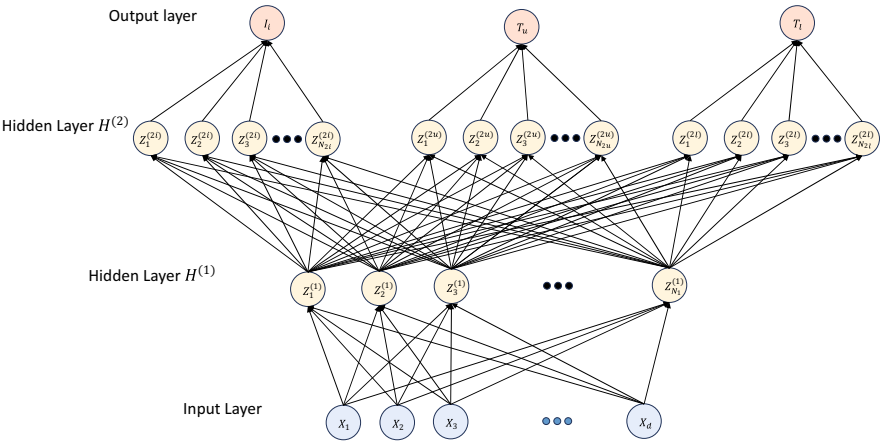


Fig. 2.7 Schematic of a multi-output NN with two hidden layers. (Source: Figure 1 of [170])

which component to be triggered under certain circumstances; and $T(\cdot)$ is a "triggering score" such that $T(\mathbf{X}) \leq t$ defines the contingent events when the insurance payouts are calculated from index $\mathbf{X}$. Since index insurance generally have a lower risk loading compared to conventional indemnity-based insurance, blended insurance (2.9) attains improved cost efficiency as well as reduced moral hazard and adverse selection. In [170], a machine learning-based algorithm is established to solve for the optimal blended payout function. Figure 2.7 illustrates the multi-output neural network for this blended index insurance, with output I_i describing the index payout level, and outputs T_u and T_l determine whether the conventional or index component should be triggered. Based on soybean production data in Iowa, USA, [170] show that the blended payout function can be viewed as a combination of conventional and index insurance and thus attains enhanced basis risk reduction, as illustrated in Fig. 2.8.

2.5 Advancements in Index-based Financial Facilities

While current index insurance programs have demonstrated effectiveness in mitigating risks associated with weather-related events, several challenges threaten their long-term profitability and sustainability. In particular, data availability and basis risk remain key concerns in index insurance design. Recent advances in technology and index-based financial markets in provide promising solutions to enhance the development and scalability of index-based insurance programs for weather and climate-related risks.

Blockchain Technology for Smart Contract The integration of blockchain technology offers a promising avenue for the development of transparent, automated,

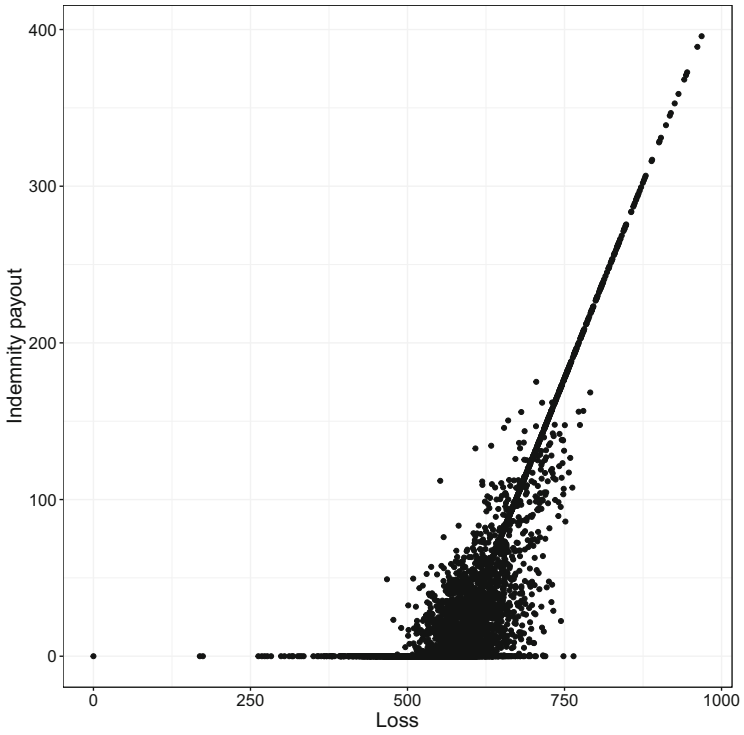


Fig. 2.8 Basis risk for blended index insurance. (Source: Figure 6 of [170])

and tamper-proof smart contracts in index insurance. By leveraging decentralized consensus mechanisms, blockchain has the potential to expand the contracting space through the utilization of smart contracts, which can address issues related to informational asymmetry, enhance market entry and competition, and ultimately improve welfare and consumer surplus [43]. A key advantage of blockchain technology is its ability to expedite settlement, eliminating fragmented post-trade infrastructures and facilitating a more flexible and efficient settlement cycle. Chiu and Koeppl [33] construct theoretical model for a hypothetical blockchain-based securities settlement system and estimate that the U.S. corporate debt market could yield net gains ranging from 1 to 4 basis points (bps) with blockchain implementation. In the context of weather-based index insurance, Salem et al. [132] propose a blockchain-based smart contract framework for weather-based index insurance, selecting Docker Container and the Neo blockchain platform for experimentation.

The index insurance market has been evolving, with blockchain technology playing an increasing role in expanding coverage and enhancing efficiency. For example, in Africa, a collaborative effort has utilized blockchain technology to implement weather index insurance to support smallholder farmers and address

the low penetration of agricultural insurance. So far, the project has reached over 12,567 farmers in Kenya with at least 511 farmers receiving mid-season payouts during the Long Rains 2021 season. [5] Similarly, in Vietnam, a regional insurtech company, in collaboration with an insurance provider, the Vietnam Meteorological and Hydrological Administration, and a reinsurer, has launched the country's first blockchain-based Weather Index Insurance product. This parametric insurance solution automates insurance claims processing for rice farmers, mitigating the financial risks posed by severe climate conditions. The insurance product offers an affordable premium starting from $8/hectare, providing coverage against irregular rainfall, while ensuring faster, simpler, and more objective claims settlement. The initiative aims to cover 50,000 hectares in collaboration with public-private enterprises, reinforcing the role of blockchain technology in scaling sustainable agricultural insurance solutions.[6]

AI and ML Methods to Improve Loss Modeling Basis risk is the major factor that hinders the success of index-based insurance programs. Therefore, improving loss prediction is critical. These technologies offer sophisticated data analysis capabilities, enabling insurers to enhance risk assessment, improve underwriting processes, and predict potential losses more accurately. In recent years, there has been a substantial and growing body of literature exploring the utilization of various machine/deep learning algorithms across various fields of actuarial science, including the life sector and nonlife sector. Examples from the nonlife side include, for example, risk modeling and prediction [48, 57, 65, 67, 68, 81, 84, 93, 94, 98–100, 108, 118, 163, 165, 174], insurance reserving [1, 66, 97, 104, 164], climate and weather risk management [14, 29, 71, 167], lapse risk management [103], fraud detection [46, 77], cyber insurance [55], and regulation [23]. For an in-depth review of the recent advancements of ML in actuarial science, refer to [130] and [131], as well as the references therein. Recent advances in AI and ML methods provide the technical foundation to achieve the objective of improving loss prediction and reducing basis risk in designing index insurance programs.

Insurance-linked Security (ILS) Market An alternative efficient risk sharing and risk management strategy is to seek capital market solutions via the insurance-linked security (ILS) market. With value linked to insurance-related risks, such as natural disasters, health and life insurance risks, etc., ILS provides a distinctive facet of the financial landscape, offering innovative solutions for climate risk and weather risk transfer. Within this market, innovative instruments include contingent capital, catastrophe (CAT) futures and options, CAT swaps, as well as CAT bonds [27, 44, 45, 58, 76, 78, 85, 92, 116]. In particular, CAT bonds are by far the most popular and successful hedging instrument in the market. For example, as

[5] More information is available here: https://acreafrica.com/reimagining-agriculture-insurance-using-blockchain-technology/.

[6] More information is vailabile here: https://iglooinsure.com/press/igloo-launches-weather-index-insurance-for-rice-farmers/.

of 2020, catastrophe bonds and ILS issued and outstanding were $14.17 billion and $46.36 billion, respectively.[7] In a CAT bond agreement, investors forgive some principal and/or coupon payments if a predefined catastrophic event, such as a hurricane, earthquake, or flood, occurs triggering significant losses for the issuer. In this case, the coupon/principal payout retained may help to stabilize the bond issuer's cash flows during natural disasters. If no catastrophe occurs, investors receive their principal plus coupons that are much higher than LIBOR [59, 117]. Thus, a CAT bond arrangement offers potential benefits for both the bond issuer and investors. This financial mechanism has gained significant traction as it allows issuers to enhance their risk resilience, while providing investors with a unique opportunity to diversify their portfolios. CAT bonds are particularly attractive due to their high yields and low correlation with traditional financial market returns, making them an appealing option for institutional investors seeking alternative risk exposure.

The Role of Reinsurance in Index Insurance Market Development Reinsurance plays an important role in the development of the index insurance market, especially in its sustainability and expansion. According to a survey conducted by Access to Insurance Initiative (A2ii) in 2020 that included 27 jurisdictions, most current index insurance products are reinsured [135]. The market is also experiencing growing demand for aggregate loss coverage, revenue protection, and catastrophe protection—key areas where reinsurers play a crucial role in providing substantial support. Reinsurers enhance the capacity of insurers to offer broader coverage by absorbing a portion of the risk. This risk distribution is essential, especially for systemic weather risk protection, where events like droughts or floods can trigger simultaneous payouts across many policies. With global exposure, reinsurers have the capacity to redistribute climate and weather-related risks across diversified portfolios, providing crucial capital relief. Moreover, reinsurers play a leading role innovative risk modeling developing advanced climate risk models and designing new index-based insurance products that help minimize basis risk. These advancements enhance the accuracy and reliability of underlying insurance policies, making them more effective in managing climate-related risks.

2.6 Conclusion

This chapter has provided a comprehensive examination of index insurance as a pivotal instrument in managing weather risks, particularly in the context of increasing climate variability and extreme weather events. Through a detailed discussion of its actuarial framework, empirical foundations, and practical considerations, we have highlighted the multifaceted dimensions that influence the effectiveness

[7] Source: Artemis Catastrophe Bond & Insurance-Linked Securities Deal Directory (https://www.artemis.bm).

and viability of index insurance in mitigating climate risks. The incorporation of innovative index-based financial facilities and the infusion of cutting-edge technologies, such as AI and blockchain, underscore the dynamic evolution of this risk management tool. These advancements not only enhance precision in risk modeling but also streamline processes, marking a transformative shift in the landscape of index insurance.

While existing index insurance programs exhibit efficacy in mitigating weather-related risks, challenges to their long-term profitability and sustainability persist. Beyond concerns related to basis risk and data availability, factors such as affordability, regulatory support, and limited awareness among policyholders contribute to the intricacies of the landscape. Active engagement and support from governments, NGOs, international agencies, and the private sector are crucial in navigating these challenges.

Looking ahead, this chapter has emphasized the importance of adopting a forward-thinking approach that considers not only the unique needs of smallholder farmers and vulnerable communities but also the challenges associated with insuring difficult and systemic risks. As we explore the future directions for index insurance design, the focus must be on fostering inclusivity and developing innovative financial instruments that effectively address the diverse and complex risks posed by climate change and extreme weather events. By advancing scalable, data-driven, and adaptive solutions, we can work toward building a more resilient and sustainable financial ecosystem—one that not only mitigates the impact of weather-related risks across different scales but also supports long-term economic stability and aligns with global development goals in an increasingly uncertain climate.

References

1. Al-Mudafer, M. T., Avanzi, B., Taylor, G., & Wong, B. (2022). Stochastic loss reserving with mixture density neural networks. *Insurance: Mathematics and Economics*, *105*, 144–174.
2. Alley, R. B., Marotzke, J., Nordhaus, W. D., Overpeck, J. T., Peteet, D. M., Pielke, R. A., Pierrehumbert, R. T., Rhines, P. B., Stocker, T. F., Talley, L. D., et al. (2003). Abrupt climate change. *Science*, *299*(5615), 2005–2010.
3. Arrow, K. J. (1974). Optimal insurance and generalized deductibles. *Scandinavian Actuarial Journal*, *1974*(1), 1–42.
4. Assa, H. (2015). On optimal reinsurance policy with distortion risk measures and premiums. *Insurance: Mathematics and Economics*, *61*, 70–75.
5. Assa, H., & Wang, M. (2021). Price index insurances in the agriculture markets. *North American Actuarial Journal*, *25*(2), 286–311.
6. Babcock, B. A., Hart, C. E., & Hayes, D. J. (2004). Actuarial fairness of crop insurance rates with constant rate relativities. *American Journal of Agriculture Economics*, *86*(3), 563–575.
7. Baret, F., & Weiss, M. (2010). Towards an operational gmes land monitoring core service - biopar methods compendium - lai, fapar, f cover ndvi. Technical Report 1.5, European Research Project Geoland 2 (EC Proposal Reference No. FP-7-218795).
8. Barnett, B. J., & Mahul, O. (2007). Weather index insurance for agriculture and rural areas in lower-income countries. *American Journal of Agricultural Economics*, *89*(5), 1241–1247.

9. Bernheim, B. D., & Sprenger, C. (2020). On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica*, *88*(4), 1363–1409.
10. Bernstein, A., Gustafson, M. T., & Lewis, R. (2019). Disaster on the horizon: The price effect of sea level rise. *Journal of Financial Economics*, *134*(2), 253–272.
11. Binswanger-Mkhize, H. P. (2012). Is there too much hype about index-based agricultural insurance? *Journal of Development Studies*, *48*(2), 187–200.
12. Borman, J. I., Goodwin, B. K., Coble, K. H., Knight, T. O., & Rejesus, R. (2013). Accounting for short samples and heterogeneous experience in rating crop insurance. *Agriculture Finance Review*, *73*(1), 88–101.
13. Born, P., & Viscusi, W. K. (2006). The catastrophic effects of natural disasters on insurance markets. *Journal of Risk and Uncertainty*, *33*(1–2), 55–72.
14. Brock Porth, C., Porth, L., Zhu, W., Boyd, M., Tan, K. S., & Liu, K. (2020). Remote sensing applications for insurance: A predictive model for pasture yield in the presence of systemic weather. *North American Actuarial Journal*, *24*(2), 333–354.
15. Bryan, G. (2019). Ambiguity aversion decreases the impact of partial insurance: Evidence from african farmers. *Journal of the European Economic Association*, *17*(5), 1428–1469.
16. Bühlmann, H. (1980). An economic premium principle. *ASTIN Bulletin*, *11*, 52–60.
17. Bühlmann, H., Delbaen, F., Embrechts, P., & Shiryaev, A. N. (1996). No-arbitrage, change of measure and conditional Esscher transforms. *CWI Quarterly*, *9*(4), 291–317.
18. Cai, J., De Janvry, A., & Sadoulet, E. (2020). Subsidy policies and insurance demand. *American Economic Review*, *110*(8), 2422–2453.
19. Camacho, F., & Torralba, I. (2011). Towards an operational gmes land monitoring core service – validation report high resolution vegetation parameters. Technical Report 1.1, European Research Project Geoland 2 (EC Proposal Reference No. FP-7-218795).
20. Cao, X., Okhrin, O., Odening, M., & Ritter, M. (2015). Modelling spatio-temporal variability of temperature. *Computational Statistics*, *30*(3), 745–766.
21. Capuno, J. J., Kraft, A. D., Quimbo, S., Tan, J. C. R., & Wagstaff, A. (2014). Effects of interventions to raise voluntary enrollment in a social health insurance scheme: A cluster randomized trial. World Bank Policy Research Working Paper (6893).
22. Carter, M. R., De Janvry, A., Sadoulet, E., & Sarris, A. (2014). *Index-based weather insurance for developing countries: A review of evidence and a set of propositions for up-scaling*.
23. Castellani, G., Fiore, U., Marino, Z., Passalacqua, L., Perla, F., Scognamiglio, S., & Zanetti, P. (2018). An investigation of machine learning approaches in the Solvency II valuation framework. Available at SSRN 3303296.
24. Ceballos, F., & Robles, M. (2020). Demand heterogeneity for index-based insurance: The case for flexible products. *Journal of Development Economics*, *146*, 102515.
25. Chantarat, S., Mude, A. G., Barrett, C. B., & Carter, M. R. (2013). Designing index-based livestock insurance for managing asset risk in northern Kenya. *Journal of Risk and Insurance*, *80*(1), 205–237.
26. Charness, G., & Levin, D. (2005). When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity, and affect. *American Economic Review*, *95*(4), 1300–1309.
27. Charpentier, A. (2008). Insurability of climate risks. *Geneva Papers on Risk and Insurance: Issues and Practice 33*, 91–109. https://doi.org/10.1057/palgrave.gpp.2510155
28. Chen, F. W., & Liu, C. W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of taiwan. *Paddy and Water Environment*, *10*(3), 209–222.
29. Chen, Z., Lu, Y., Zhang, J., & Zhu, W. (2024). Managing weather risk with a neural network-based index insurance. *Management Science 70*(7), 4306–4327.
30. Chi, Y., & Tan, K. S. (2011). Optimal reinsurance under var and cvar risk measures: A simplified approach. *ASTIN Bulletin: The Journal of the IAA*, *41*(2), 487–509.
31. Chi, Y., & Weng, C. (2013). Optimal reinsurance subject to vajda condition. *Insurance: Mathematics and Economics*, *53*(1), 179–189.

32. Chi, Y., Zhang, J., & Zhu, W. (2021). Low demand for index insurance. Nanyang Business School Working Paper.
33. Chiu, J., & Koeppl, T. V. (2019). Blockchain-based settlement for asset trading. *The Review of Financial Studies*, *32*(5), 1716–1753.
34. Clarke, D. J. (2016). A theory of rational demand for index insurance. *American Economic Journal: Microeconomics*, *8*(1), 283–306.
35. Clarke, D., & Kalani, G. (2011). Microinsurance decisions: Evidence from ethiopia. In *CSAE 25th anniversary conference*. Citeseer.
36. Coble, K. H., Harri, A., Anderson, J. D., Ker, A. P., & Goodwin, B. (2008). Review of County Yield Trending Procedures and Related Topics. Technical Report, USDA Risk Management Agency.
37. Coble, K. H., Miller, M. F., Rejesus, R. M., Boyles, R., Knight, T. O., & Goodwin, B. K. (2011). Methodology analysis for weighting of historical experience. Technical Report, USDA Risk Management Agency.
38. Coble, K., Knight, T., Miller, M., Goodwin, B., Rejesus, R., & Boyles, R. (2013). Estimating structural change in U.S. crop insurance experience. *Agricultural Finance Review*, *73*(1), 74–87.
39. Cole, S. A., & Xiong, W. (2017). Agricultural insurance and economic development. *Annual review of Economics*, *9*, 235–262.
40. Cole, S., Giné, X., Tobacman, J., Topalova, P., Townsend, R., & Vickery, J. (2013). Barriers to household risk management: Evidence from India. *American Economic Journal: Applied Economics*, *5*(1), 104–135.
41. Cole, S., Stein, D., & Tobacman, J. (2014). Dynamics of demand for index insurance: Evidence from a long-run field experiment. *American Economic Review*, *104*(5), 284–290.
42. Collier, B., Katchova, A. L., & Skees, J. R. (2011). Loan portfolio performance and el niño, an intervention analysis. *Agricultural Finance Review*, *71*(1), 98–119.
43. Cong, L. W., & He, Z. (2019). Blockchain disruption and smart contracts. *The Review of Financial Studies*, *32*(5), 1754–1797.
44. Cummins, J. D., & Trainar, P. (2009). Securitization, insurance, and reinsurance. *Journal of Risk and Insurance*, *76*, 463–492. https://doi.org/10.1111/j.1539-6975.2009.01319.x.
45. Cummins, J. D., Lalonde, D., & Phillips, R. D. (2004). The basis risk of catastrophic-loss index securities. *Journal of Financial Economics*, *71*(1), 77–111.
46. Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*, *90*(3), 743–768.
47. Deng, X., Barnett, B. J., & Vedenov, D. V. (2007). Is there a viable market for area-based crop insurance? *American Journal of Agricultural Economics*, *89*(2), 508–519.
48. Diao, L., & Weng, C. (2019). Regression tree credibility model. *North American Actuarial Journal*, *23*(2), 169–196.
49. Dirks, K., Hay, J., Stow, C., & Harris, D. (1998). High-resolution studies of rainfall on norfolk island: Part II: interpolation of rainfall data. *Journal of Hydrology*, *208*(3), 187–193.
50. Doherty, N. A., & Schlesinger, H. (1983). Optimal insurance in incomplete markets. *Journal of Political Economy*, *91*(6), 1045–1054.
51. Doherty, N. A., & Schlesinger, H. (1990). Rational insurance purchasing: consideration of contract nonperformance. *The Quarterly Journal of Economics*, *105*(1), 243–253.
52. Elabed, G., Bellemare, M. F., Carter, M. R., & Guirkinger, C. (2013). Managing basis risk with multiscale index insurance. *Agricultural Economics*, *44*(4–5), 419–431.
53. Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669.
54. Fan, Q., Tan, K. S., & Zhang, J. (2023). Empirical tail risk management with model-based annealing random search. *Insurance: Mathematics and Economics*, *110*, 106–124.
55. Farkas, S., Lopez, O., & Thomas, M. (2021). Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, *98*, 92–105.

56. Fisher, A. C., Hanemann, W. M., Roberts, M. J., & Schlenker, W. (2012). The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather: Comment. *American Economic Review*, *102*(7), 3749–3760.
57. Fissler, T., Merz, M., & Wüthrich, M. V. (2023). Deep quantile and deep composite triplet regression. *Insurance: Mathematics and Economics*, *109*, 94–112.
58. Froot, K. A. (1999). The evolving market for catastrophic event risk the distribution of catastrophic event risk. *Risk Management and lnsurance Review*, *2*(3), 1–28.
59. Froot, K. A. (2001). The market for catastrophe risk: A clinical examination. *Journal of Financial Economics*, *60*(2–3), 529–571.
60. Furman, E., & Landsman, Z. (2006). Tail variance premium with applications for elliptical portfolio of risks. *ASTIN Bulletin*, *36*(2), 433–462.
61. Furman, E., & Landsman, Z. (2008). Economic capital allocations for non-negative portfolio of dependent risks. *ASTIN Bulletin*, *38*(2), 601–619.
62. Furman, E., & Zitikis, R. (2008). Weighted premium calculation principles. *Insurance: Mathematics and Economics*, *42*(1), 459–465.
63. Furman, E., & Zitikis, R. (2008). Weighted risk capital allocations. *Insurance: Mathematics and Economics*, *43*(2), 263–269.
64. Furman, E., & Zitikis, R. (2010). General Stein-type covariance decompositions with applications to insurance and finance. *ASTIN Bulletin*, *40*(1), 369–375.
65. Gabrielli, A., & Wüthrich, M. V. (2018). An individual claims history simulation machine. *Risks*, *6*(2), 29.
66. Gabrielli, A., Richman, R., & Wüthrich, M. V. (2020). Neural network embedding of the over-dispersed poisson reserving model. *Scandinavian Actuarial Journal*, *2020*(1), 1–29.
67. Gao, G., Meng, S., & Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, *2019*(2), 143–162.
68. Gao, G., Wang, H., & Wüthrich, M. V. (2022). Boosting Poisson regression models with telematics car driving data. *Machine Learning*, *111*(1), 243–272.
69. Gaurav, S., Cole, S., & Tobacman, J. (2011). Marketing complex financial products in emerging markets: Evidence from rainfall insurance in India. *Journal of Marketing Research*, *48*(SPL), S150–S162.
70. Gerber, H., & Shiu, S. (1994). Option pricing by Esscher transforms. *Transaction of Society of Actuaries*, *46*, 99–140.
71. Ghahari, A., Newlands, N. K., Lyubchich, V., & Gel, Y. R. (2019). Deep learning at the interface of agricultural insurance risk and spatio-temporal uncertainty in weather extremes. *North American Actuarial Journal*, *23*(4), 535–550.
72. Giné, X., & Yang, D. (2009). Insurance, credit, and technology adoption: Field experimental evidencefrom malawi. *Journal of development Economics*, *89*(1), 1–11.
73. Giné, X., Townsend, R., & Vickery, J. (2007). Statistical analysis of rainfall insurance payouts in southern India. *American Journal of Agricultural Economics*, *89*(5), 1248–1254.
74. Giné, X., Townsend, R., & Vickery, J. (2008). Patterns of rainfall insurance participation in rural India. *World Bank Economic Review*, *22*(3), 539–566.
75. Gine, X., Karlan, D., & Ngatia, M. (2013). *Social networks, financial literacy and index insurance*. World Bank Washington.
76. Golden, L. L., Wang, M., & Yang, C. (2007). Handling weather related risks through the financial markets: Considerations of credit risk, basis risk, and hedging. *Journal of Risk and Insurance*, *74*(2), 319–346. https://doi.org/10.1111/j.1539-6975.2007.00215.x.
77. Gomes, C., Jin, Z., & Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, *88*(3), 591–624.
78. Götze, T., & Gürtler, M. (2022). Risk transfer beyond reinsurance: The added value of CAT bonds. *Geneva Papers on Risk and Insurance: Issues and Practice*, *47*(1), 125–171.
79. Hartman-Glaser, B., & Hébert, B. (2020). The insurance is the lemon: Failing to index contracts. *Journal of Finance*, *75*(1), 463–506.
80. Heilmann, W. R. (1989). Decision theoretic foundations of credibility theory. *Insurance: Mathematics and Economics*, *8*(1), 77–95.

81. Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, *25*(2), 255–285.
82. Hill, R. V., Robles, M., & Ceballos, F. (2016). Demand for a simple weather insurance product in India: Theory and evidence. *American Journal of Agricultural Economics*, *98*(4), 1250–1270.
83. Hong, H. G., Karolyi, G. A., & Scheinkman, J. (2020). Climate finance. *Review of Financial Studies*, *33*(3), 1011–1023.
84. Hu, C., Quan, Z., & Chong, W. F. (2022). Imbalanced learning for insurance using modified loss functions in tree-based models. *Insurance: Mathematics and Economics*, *106*, 13–32.
85. Huang, S., Zhang, J., & Zhu, W. (2023). Storm cat bond: Modeling and valuation. *North American Actuarial Journal*, 1–26.
86. Hubalek, F., & Sgarra, C. (2006). On the Esscher transform and the minimal entropy martingale measure for exponential lévy models. *Quantitative Finance*, *6*(2), 125–145.
87. Huck, S., & Weizsäcker, G. (1999). Risk, complexity, and deviations from expected-value maximization: Results of a lottery choice experiment. *Journal of Economic Psychology*, *20*(6), 699–715.
88. Iyengar, S. S., & Kamenica, E. (2010). Choice proliferation, simplicity seeking, and asset allocation. *Journal of Public Economics*, *94*(7–8), 530–539.
89. Jensen, N. D., Barrett, C. B., & Mude, A. G. (2016). Index insurance quality and basis risk: Evidence from northern Kenya. *American Journal of Agricultural Economics*, *98*(5), 1450–1469.
90. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292.
91. Kamps, U. (1998). On a class of premium principles including the Esscher principle. *Scandinavian Actuarial Journal*, *1998*(1), 75–80.
92. Karagiannis, N., Assa, H., Pantelous, A., & Turvey, C. (2016). Modelling and pricing of catastrophe risk bonds with a temperature-based agricultural application. *Quantitative Finance*, *16*(12), 1949–1959.
93. Karimi, N., Assa, H., Salavati, E., & Adibi, H. (2023). Calibration of storage model by multi-stage statistical and machine learning methods. *Computational Economics*, *62*(4), 1437–1455.
94. Karimi, N., Salavati, E., Assa, H., & Adibi, H. (2023). Sensitivity analysis of optimal commodity decision making with neural networks: A case for covid-19. *Mathematics*, *11*(5), 1202.
95. Kreps, D. M. (1979). A representation theorem for "preference for flexibility". *Econometrica*, 565–577.
96. Kubitza, C., Hofmann, A., & Steinorth, P. (2020). Financial literacy and precautionary insurance. Working Paper, University of Bonn.
97. Kuo, K. (2019). Deeptriangle: A deep learning approach to loss reserving. *Risks*, *7*(3), 97.
98. Lally, N., & Hartman, B. (2018). Estimating loss reserves using hierarchical bayesian gaussian process regression with input warping. *Insurance: Mathematics and Economics*, *82*, 124–140.
99. Lee, S. C. (2021). Addressing imbalanced insurance data through zero-inflated poisson regression with boosting. *ASTIN Bulletin: The Journal of the IAA*, *51*(1), 27–55.
100. Lee, S. C., & Lin, S. (2018). Delta boosting machine with application to general insurance. *North American Actuarial Journal*, *22*(3), 405–425.
101. Li, J., Zhang, J., Zhang, C., & Chen, Q. (2006). Analyze and compare the spatial interpolation methods for climate factor. *Pratacultural Science*, *8*, 001.
102. Li, H., Porth, L., Tan, K. S., & Zhu, W. (2021). Improved index insurance design and yield estimation using a dynamic factor forecasting approach. *Insurance: Mathematics and Economics*, *96*, 208–221.
103. Loisel, S., Piette, P., & Tsai, C. H. J. (2021). Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin: The Journal of the IAA*, *51*(3), 839–871.

104. Lopez, O., Milhaud, X., & Thérond, P. E. (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin: The Journal of the IAA* , *49*(3), 741–762.
105. Mahul, O., & Skees, J. R. (2007). Managing agricultural risk at the country level: The case of index-based livestock insurance in mongolia. World Bank Policy Research Working Paper (4325).
106. Mahul, O., & Verma, N. (2011). *Enhancing crop insurance in India*. World Bank Washington.
107. Makaudze, E. M., & Miranda, M. J. (2010). Catastrophic drought insurance based on the remotely sensed normalised difference vegetation index for smallholder farmers in zimbabwe. *Agrekon*, *49*(4), 418–432.
108. Meng, S., Gao, Y., & Huang, Y. (2022). Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. *Insurance: Mathematics and Economics*, *106*, 115–127.
109. Miranda, M. J. (1991). Area-yield crop insurance reconsidered. *American Journal of Agricultural Economics*, *73*(2), 233–242. https://doi.org/10.2307/1242708.
110. Miranda, M. J., & Farrin, K. (2012). Index insurance for developing countries. *Applied Economic Perspectives and Policy*, *34*(3), 391–427.
111. Miranda, M. J., & Glauber, J. W. (1997). Systemic risk, reinsurance, and the failure of crop insurance markets. *American journal of Agricultural Economics*, *79*(1), 206–215.
112. Mobarak, A. M., & Rosenzweig, M. R. (2012). Selling formal insurance to the informally insured. Yale Economics Department Working Paper No. 97; Yale University Economic Growth Center Discussion Paper No. 1007.
113. Mobarak, A. M., & Rosenzweig, M. R. (2013). Informal risk sharing, index insurance, and risk taking in developing countries. *American Economic Review*, *103*(3), 375–380.
114. Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering*, *114*(4), 359–371.
115. Musshoff, O., Odening, M., & Xu, W. (2011). Management of climate risks in agriculture–will weather derivatives permeate? *Applied Economics*, *43*(9), 1067–1077.
116. Nell, M., & Richter, A. (2004). Improving risk allocation through indexed cat bonds. *Geneva Papers on Risk and Insurance*, *29*(2), 183–201. https://about.jstor.org/terms.
117. Niehaus, G. (2002). The allocation of catastrophe risk. *Journal of Banking & Finance*, *26*, 585–596.
118. Noll, A., Salzmann, R., & Wuthrich, M. V. (2020). Case study: French motor third-party liability claims. Available at SSRN 3164764.
119. Nordhaus, W. (2019). Climate change: The ultimate challenge for economics. *American Economic Review*, *109*(6), 1991–2014.
120. Okhrin, O., Odening, M., & Xu, W. (2013). Systemic weather risk and crop insurance: The case of china. *Journal of Risk and Insurance*, *80*(2), 351–372.
121. Ozaki, V. A., Goodwin, B. K., & Shirota, R. (2008). Parametric and nonparametric statistical modeling of crop yield: implications for pricing crop insruance contracts. *Applied Economics*, *40*(9), 1151–1164.
122. Pineiro, G., Oesterheld, M., & Paruelo, J. (2006). Seasonal variation in aboveground production and radiation-use efficiency of temperate rangelands estimated through remote sensing. *Ecosystems*, *9*, 357–373.
123. Platteau, J. P., De Bock, O., & Gelade, W. (2017). The demand for microinsurance: A literature review. *World Development*, *94*, 139–156.
124. Porth, L., Tan, K. S., & Weng, C. (2013). Optimal reinsurance analysis from a crop insurer's perspective. *Agriculture Finance Review*, *73*(2), 310–328.
125. Porth, L., Pai, J., & Boyd, M. (2015). A portfolio optimization approach using combinatorics with a genetic algorithm for developing a reinsurance model. *Journal of Risk and Insurance*, *82*(3), 687–713.
126. Porth, L., Zhu, W., & Tan, K. S. (2014). A credibility-based Erlang mixture model for pricing crop reinsurance. *Agricultural Finance Review*, *74*(2), 162–187.

127. Porth, L., Pai, J., & Boyd, M. (2015). A portfolio optimization approach using combinatorics with a genetic algorithm for developing a reinsurance model. *Journal of Risk and Insurance*, *82*(3), 687–713.
128. Porth, L., Tan, K. S., & Zhu, W. (2019). A relational data matching model for enhancing individual loss experience: An example from crop insurance. *North American Actuarial Journal*, *23*(4), 551–572.
129. Raviv, A. (1979). The design of an optimal insurance policy. *American Economic Review*, *69*(1), 84–96.
130. Richman, R. (2021). Ai in actuarial science–a review of recent advances–part 1. *Annals of Actuarial Science*, *15*(2), 207–229.
131. Richman, R. (2021). Ai in actuarial science–a review of recent advances–part 2. *Annals of Actuarial Science*, *15*(2), 230–258.
132. Salem, M. J., Ndolu, F. H. E., Hidayatullah, D. E. R., & Sari, R. F. (2021). Developing neo smart contract for weather-based insurance. In *2021 4th international seminar on research of information technology and intelligent systems (ISRITI)* (pp. 603–608). IEEE.
133. Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, *106*(37), 15594–15598.
134. Shin, S., Magnan, N., Mullally, C., & Janzen, S. (2022). Demand for weather index insurance among smallholder farmers under prospect theory. i, *202*, 82–104.
135. Simões, R. (2020). Index Insurance: 2020 Status and Regulatory Challenges. Technical Report, Access to Insurance Initiative.
136. Skees, J. R., Black, J. R., & Barnett, B. J. (1997). Designing and rating an area yield crop insurance contract. *American Journal of Agricultural Economics*, *79*(2), 430–438. https://doi.org/10.2307/1244141.
137. Smith, V. H., & Watts, M. (2019). Index based agricultural insurance in developing countries: Feasibility, scalability and sustainability. *Gates Open Res*, *3*(65), 65.
138. Sonsino, D., & Mandelbaum, M. (2001). On preference for flexibility and complexity aversion: Experimental evidence. *Theory and Decision*, *51*(2), 197–216.
139. Sonsino, D., Benzion, U., & Mador, G. (2002). The complexity effects on choice with uncertainty–experimental evidence. *The Economic Journal*, *112*(482), 936–965.
140. Sun, X., Manton, M., & Ebert, E. E. (2003). Regional rainfall estimation using double-kriging of raingauge and satellite observations. *Bureau of Meteorology*.
141. Swiss Re (2013). Partnering for food security in emerging markets. *Sigma*, *1*, 1–44.
142. Tan, K. S., & Zhang, J. (2023). Flexible weather index insurance design with penalized splines. *North American Actuarial Journal*, 1–26.
143. Thornton, R. L., Hatt, L. E., Field, E. M., Islam, M., Solís Diaz, F., & González, M. A. (2010). Social security health insurance for the informal sector in nicaragua: A randomized evaluation. *Health Economics 19*(S1), 181–206.
144. Turvey, C. G. (2001). Weather derivatives for specific event risks in agriculture. *Applied Economic Perspectives and Policy*, *23*(2), 333–351.
145. Turvey, C. G. (2005). The pricing of degree-day weather options. *Agricultural Finance Review*, *65*(1), 59–85.
146. Turvey, C. G., & Mclaurin, M. K. (2012). Applicability of the normalized difference vegetation index (ndvi) in index-based crop insurance design. *Weather, Climate, and Society*, *4*(4), 271–284.
147. Turvey, C. G., Weersink, A., & Chiang, S. H. C. (2006). Pricing weather insurance with a random strike price: The ontario ice-wine harvest. *American Journal of Agricultural Economics*, *88*(3), 696–709.
148. Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297–323.
149. USDA (2014). World agricultural supply and demand estimates report (WASDE). Office of the chief economist (OCE), United States Department of Agriculture.

150. Vedenov, D. V., & Barnett, B. J. (2004). Efficiency of weather derivatives as primary crop insurance instruments. *Journal of Agricultural and Resource Economics*, 387–403.
151. Wakker, P., Thaler, R., & Tversky, A. (1997). Probabilistic insurance. *Journal of Risk and Uncertainty*, *15*, 7–28.
152. Wang, S. (1995). Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathetatics and Economics*, *17*, 43–54.
153. Wang, S. (1996). Premium calculation by transforming the layer premium density. *ASTIN Bulletin*, *26*(1), 71–92.
154. Wang, S. (2007). Normalized exponential tilting: pricing and measuring multivariate risks. *North American Actuarial Journal*, *11*(3), 89–99.
155. Wang, H. H., & Zhang, H. (2003). On the possibility of a private crop insurance market: A spatial statistics approach. *Journal of Risk and Insurance*, *70*(1), 111–124.
156. Wang, S., Young, V. R., & Panjer, H. H. (1997). Axiomatic characterization of insurance prices. *Insurance: Mathematics and Economics*, *21*(2), 173–183.
157. Woodard, J. D. (2014). A conditional distribution approach to assessing the impact of weather sample heterogeneity on yield risk estimation and crop insurance ratemaking. *North American Actuarial Journal*, *18*(2), 279–293.
158. Woodard, J. D., Pavlista, A. D., Schnitkey, G. D., Burgener, P. A., & Ward, K. A. (2012). Governement insurance program design, incentive effects, and technology adoption: The case of skip-row crop insurance. *American Journal of Agriculture Economics*, *94*(4), 823–837.
159. Woodard, J. D., Schnitkey, G. D., Sherrick, B. J., Lozano-Gracia, N., & Anselin, L. (2012). A spatial econometric analysis of loss experience in the U.S. crop insurance program. *Journal of Risk and Insurance*, *79*(1), 261–286.
160. World Bank. (2011). *Weather index insurance for agriculture: guidance for development practitioners*. World Bank.
161. World Bank. (2012). *Ndvi pasture index-based insurance for livestock producers in south west buenos aires province*. World Bank.
162. World Bank. (2013). *Ndvi pasture index-based insurance for livestock producers in uruguay*. World Bank.
163. Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, *7*(1), 89–108.
164. Wüthrich, M. V. (2018). Neural networks applied to chain–ladder reserving. *European Actuarial Journal*, *8*(2), 407–436.
165. Wuthrich, M. V., & Buser, C. (2021). Data analytics for non-life insurance pricing. Swiss Finance Institute Research Paper (16–68).
166. Xu, Y., Pan, G., Tan, K. S., & Zhu, W. (2020). A sustainable area-yield insurance program with optimal risk pooling: A behavior-based machine learning approach. Nanyang Business School Working Paper.
167. Xu, Y., Tan, K. S., & Zhu, W. (2022). From meteorology to market: A geo-hierarchical deep learning approach for flood risk pricing. Nanyang Business School Working Paper.
168. Young, V. R. (2004). Premium principles. In *Encyclopedia of actuarial science*. John Wiley & Sons, Ltd.
169. Zanjani, G. (2002). Pricing and capital allocation in catastrophe insurance. *Journal of Financial Economics*, *65*(2), 283–305.
170. Zhang, J. (2024). Blended insurance scheme: A synergistic conventional-index insurance mixture. *Insurance: Mathematics and Economics*, *119*, 93–105.
Zhang, J. (2023). Blended insurance scheme: A synergistic conventional-index insurance mixture. Nanyang Business School Working Paper.
171. Zhang, J., Tan, K. S., & Weng, C. (2019). Index insurance design. *ASTIN Bulletin: The Journal of the IAA*, *49*(2), 491–523.
172. Zhang, Y., Tang, C. S., & Yu, J. J. (2023). Improving farmer welfare via weather index insurance: Sell-direct and sell-through mechanisms. UCLA Working Paper.
173. Zhu, W. (2015). Actuarial Ratemaking in Agricultural Insurance. Ph.D. Thesis, University of Waterloo.

174. Zhu, W. (2023). A deep factor model for crop yield forecasting and insurance ratemaking. *North American Actuarial Journal*, 1–16.
175. Zhu, W., Tan, K. S., Porth, L., & Wang, C. W. (2018). Spatial dependence and aggregation in weather risk hedging: A lévy subordinated hierarchical archimedean copulas (lshac) approach. *ASTIN Bulletin: The Journal of the IAA*, *48*(2), 779–815.
176. Zhu, W., Porth, L., & Tan, K. S. (2019). A credibility-based yield forecasting model for crop reinsurance pricing and weather risk management. *Agricultural Finance Review*, *79*(1), 2–26.
177. Zhu, W., Tan, K. S., & Porth, L. (2019). Agricultural insurance ratemaking: Development of a new premium principle. *North American Actuarial Journal*, *23*(4), 512–534.
178. Zhuang, S. C., Weng, C., Tan, K. S., & Assa, H. (2016). Marginal indemnification function formulation for optimal reinsurance. *Insurance: Mathematics and Economics*, *67*, 65–76.

Chapter 3
Weather and Yield Index-Based Insurance Schemes in the EU Agriculture: A Focus on the Agri-CAT Fund

F. G. Santeramo, T. Balezentis, and M. Tappi

3.1 Introduction

Agriculture is the most vulnerable sector to climate change, e.g., temperatures or rainfall may significantly affect the crop yields, also leading the proliferation of pathogens and hence pests and diseases [1]. The total economic losses from weather- and climate-related have caused damages reaching nearly 487 billion of euros in EEA member countries since 1980, and just 3% of all events are responsible for 60% of economic losses [2]. Extreme weather events such as heavy precipitation, flood, drought, frost, heat, and strong wind are more and more frequent, intense, long-lasting, and they are the major drivers of agricultural losses [3, 4]. Heavy precipitation may reduce photosynthetically active radiation up to irreversible tissue damages, setting the conditions for diseases due to the proliferation of pathogens, nutrient leaching, soil erosion, and oxygen deficit [5, 6], also inducing flash flood events, in combination with other factors as the antecedent soil moisture [7, 8]. Drought and water shortage may affect the metabolism of plants with changes in root growth and architecture, and other tissue-specific responses that modify the flux of cellular signals [9]. The stress due to drought events is the main factor limiting the development of crop and its productivity [10]. Cold may damage the leaf and seedling survival, also leading to the sterility and the abortion of formed

F. G. Santeramo (✉)
University of Foggia, Foggia, Italy

Institute of Economics and Rural Development Vilnius, Lithuania
e-mail: fabio.santeramo@unifg.it

T. Balezentis
Institute of Economics and Rural Development Vilnius, Lithuania

M. Tappi
Regione Puglia, Puglia, Italy

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_3

grains, especially for the cereal crops [11]. Heat directly affects the crop physiology, reducing photosynthesis rates, leading the acceleration of leaf senescence processes, oxidative damages, and pollen sterility [12]. Strong wind may also be very impactful (i.e., abrasions on the leaves and fruits, defoliations, water loss, desiccation, loss of flowers and poor fruit set), although the plants can change the structure and properties of cells and tissues, re-configuring their canopies as a defensive response [13]. On-farm and risk-sharing strategies are available to improve the resilience of farming systems to weather risks. The former includes risk control (i.e., risk prevention such as irrigation, shading, pest control, improved planning and monitoring activities), reserves (i.e., stocking, financial savings, additional labour input), and diversification (i.e., agricultural and structural diversification as nature conservation or agrotourism, off-farm allocation of resources); the latter includes risk pooling (i.e., mutual funds, agricultural insurance, membership in cooperatives, credit unions, producer organizations), and risk transfer (i.e., forwards, futures contracts) [14]. Member States may grant support for risk management tools (e.g., financial contribution to insurance premiums and to mutual funds) which can help farmers to manage production and income risks related to their agricultural activity and over which they have no control [15]. The new Common Agricultural Policy (CAP) reform is putting increasing emphasis on instrument supporting proactive management of the effects of extreme weather events due to climate change [15]. We provide an overview of the spread of risk management tools subsidised by new CAP 2023–2027, focusing on two promising tools: the weather index-based insurance and the Agri-CAT fund. We also discuss on their feasibility at farm-level, highlighting pros and cons, also animating the debate on how policymakers may improve the attractiveness of risk management tools.

3.2 Agricultural Risk Management Tools Provided by CAP 2023–2027: An Overview

Farmers are exposed to different types of risks that may affect the agricultural activity, and have diverse attitudes and preferences [16]: (1) price risks, due to price volatility and uncertainty about future prices as result of competition, geopolitical, climate change, and phytosanitary risks, etc.; (2) production risks, when the outputs are lower than expected mainly due to extreme weather events, pest, and disease; (3) income risks, characterised by an imbalance between revenue and costs, e.g., the prices of inputs as fertiliser, or seed, increase while the prices of outputs remain stable [17]. The new CAP 2023–2027 of the European Commission devotes around 4.6 billion of euros of total public expenditure supported by European Agricultural Fund for Rural Development (EAFRD) for the risk management in agriculture, of which 2.7 billion of euros as EU contribution and 1.9 billion of euros as national contribution. The priority given to support for risk management in agriculture is presented in the Fig. 3.1.

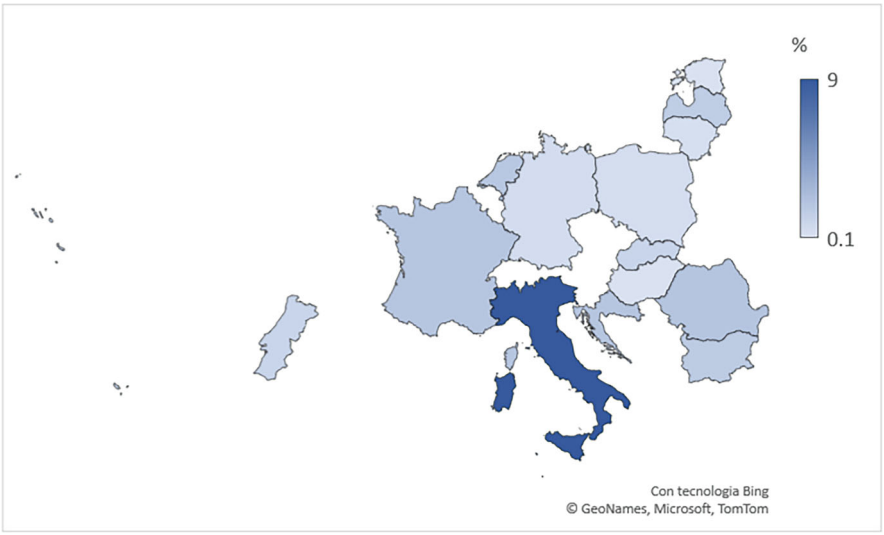


Fig. 3.1 Share of total public CAP expenditure for risk management tools. Source: adaptation from Approved 28 CAP Strategic Plans (2023–2027)

The aim of new CAP is to cover around 15% of EU farms. Currently, 14 Member States include support for risk management tools in their CAP Strategic Plans (CSPs) and propose 25 interventions under the possibility given by Article 76 of the CSP Regulation for support through insurance schemes and mutual funds, covering insurance premium, and other (Fig. 3.2). In addition, sectorial interventions are provided for fruit and vegetables, wine, olive oil, and other sectors. Moreover, producer organisations recognised by Member States (exception made for Bulgaria and Latvia) have several tools (e.g., withdrawals, harvest and production insurance, mutual funds, green and non-harvesting, collective storage) to cope with risks that may affect the agricultural production.

In addition, Bulgaria, Italy, and Romania apply a withholding from the direct payments (i.e., 1.5%, 3.0%, and 3.0%, respectively) as contribution to manage the agricultural risks. This basic coverage is complementary to the other risk management tools under EAFRD. Italy stands out at the European level for the provision of subsidised risk management tools (see Fig. 3.1), and it establishes four risk management interventions worth almost 3 billion of euros. These interventions aim to help farmers to better face growing climatic adversities through subsidised insurances, income stabilisation tools, and a new national mutual fund for catastrophic events (covering frost, floods, and drought damage). The latter has seen for the first time ever the participation of all Italian beneficiaries of direct payments (around 800,000 farmers) with support from the EAFRD and a contribution for 3% from the European Agricultural Guarantee Fund (EAGF) to the mutual fund scheme. The largest multiple-peril crop insurance schemes are in France, Spain, and Italy, single-peril hail insurances are widespread in Germany, while crop insurances

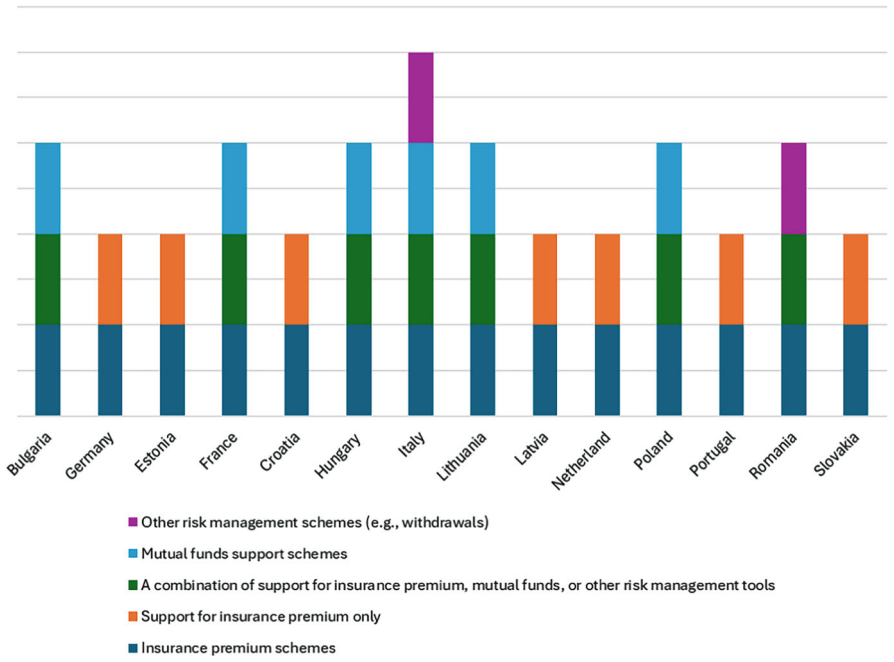


Fig. 3.2 Type of support for risk management under EAFRD (Article 76). Source: approved 28 CAP Strategic Plans 2023–2027

complementary to the weather risk covering phytosanitary risks are not widely available in EU, except in Denmark, Germany, Hungary, Italy, Netherlands, and Spain [18]. The increasing amount of available data on weather conditions, and the higher frequency of extreme weather events and natural disaster allow have led to the development of new tools (e.g., weather index-based insurance and mutual fund for catastrophic weather events) meant both to complement traditional insurance schemes and to avoid the default of many farms [19].

3.3 The Insurances Based on Weather-Index Among European Countries

Coping risks though crop insurance schemes is not straightforward, due to information asymmetries and partial knowledge from both sides [20, 21]. Weather index-based insurance is a type of insurance scheme designed to protect farmers against weather-related risks. Unlike traditional indemnity crop insurances which rely on yield losses and physical damage observations, weather index-based insurances indemnify the farmers based on predefined weather parameters (or indexes) as triggers, e.g., excess rainfall or extremely hot temperatures, easily measurable and

directly related with production yields [14, 22, 23].[1] Therefore, the indemnification is triggered whenever the value of the index exceeds or falls short the threshold [24–26]. The index, recorded by weather stations or provided by other data sources, is independent, objective, transparent, and free from manipulations [23, 27, 28]. These characteristics represent the strengths of weather index-based insurances when compared with the traditional indemnity insurances. Sure enough, the latter show some issues as the asymmetric information, high transaction costs, moral hazard, and adverse selection [29, 30]. However, the weather index-based insurances present the limit of basis risk, i.e., the discrepancy between insurance payouts and agricultural output ([31]). More specifically, the insured farmers may experience yield losses, while the weather index does not trigger a payment, or may obtain a compensation without having any yield losses because the index trigger the reimbursement, in other words, the index is not perfectly correlated with the actual losses which affected insurance policyholders [32, 33]. Furthermore, following the previous reasoning, the insured farmers may not be adequately compensated as result of production losses [31]. Basis risk can be separated into three categories depending on the causes: (1) spatial (or geographical) basis risk due to the distance of weather station from the location covered by insurance contract; (2) design basis risk due to the imperfect correlation between weather index and crop yields; (3) temporal basis risk due to imperfect time period selection for deriving the index, e.g., the weather index does not capture the phenological stage more susceptible to a specific weather event [29, 31, 34]. Morsink et al. [35], offered two different definition of basis risk: (1) insured peril basis risk that compares claim payments with losses from perils explicitly named in the insurance contract; (2) production smoothing basis risk that compares claim payments with losses from agricultural production. In other words, the challenge is to decide whether the claim payments should be compared to the value of losses (e.g., in dollar) or just the amount of losses (e.g., tons of crop or number or animals). For both types of basis risk two indicator are proposed: the probability of catastrophic basis risk and the catastrophic performance ratio. The former establishes the probability that a farmer experiences more that 70% loss of agricultural production without reaching the payment threshold; the latter reflects what, on average, a farmer gets back per $1 of commercial premium paid in the case that she experiences catastrophic crop loss [35]. Despite this main limitation, weather index insurance allows farmers to diversify their risk portfolio, also playing a crucial role in enhancing farmers' access to credit and investment [36]. Among the EU largest agricultural producers in terms of production values, Germany offers the largest number of insurance schemes, while Spain has the largest number of insurance providers, as shown in the Fig. 3.3:

The indemnity insurances remain the most offered products; the weather index-based insurances are widespread in Germany with 10 insurance products, while

[1] In Appendix we dive into methods to compute yield risks and present an empirical application to EU member States.

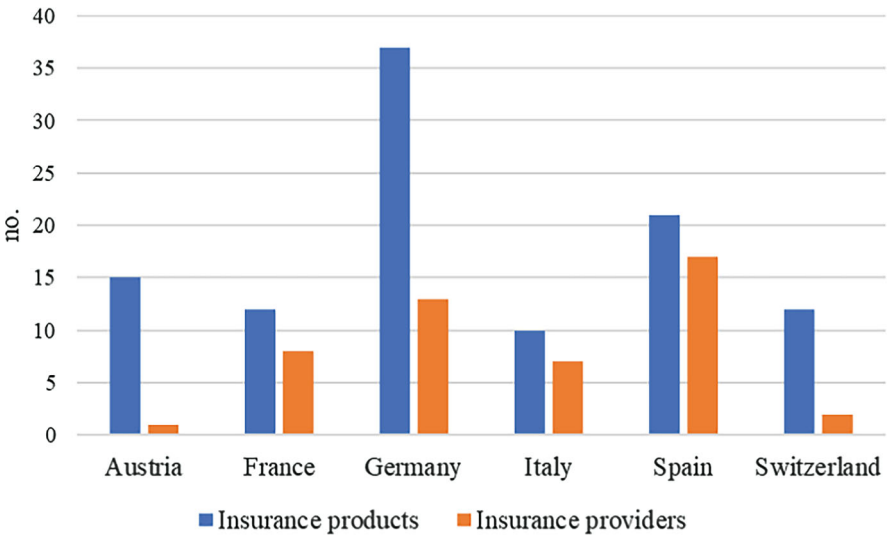


Fig. 3.3 Number of insurance schemes and insurance providers among main EU producers. Source: Bucheli et al. [3]

other countries (i.e., Austria, France, Italy, Spain, and Switzerland) offer less than 5 [3]. The weather index-based insurance covers only drought, heavy precipitation, heat, and frost events. The cumulative precipitation is the most used index to assess the damages due to drought and heavy precipitation in Germany, Italy, and Switzerland [3]. The heat days index (as daily temperature above 30 °C) is used in Austria for drought and heat insurance. The soil moisture index is used in Germany to cover drought risk, while the temperature index is used in Italy and Germany for heat and frost (Fig. 3.4).

Several challenges need to be addressed to maximize the potential of these tools: (1) investments in weather monitoring systems, remote sensing technologies, and data collection networks to ensure accurate index calculations and timely payouts, e.g., using satellite-derived datasets based on Normalized Difference Vegetation Index, also supported by low-cost in situ sensors [37]; (2) design appropriate indices that accurately reflect the risk exposure of farmers and calibrated them to capture variations in crop performance due to changes in weather, also considering the phenological stages more susceptible [25, 26, 38]; (3) affordability and accessibility of weather index insurance remain significant barriers for small-scale and marginalized farmers, e.g., premium costs should be affordable, and insurance products need to be tailored to the specific needs and limitation of different farming systems and regions; (4) knowledge and awareness, e.g., dissemination through seminars and workshops, partnerships between stakeholders, public institutions, and insurance companies, in order to improve the participation in crop insurance schemes which il still low [39].

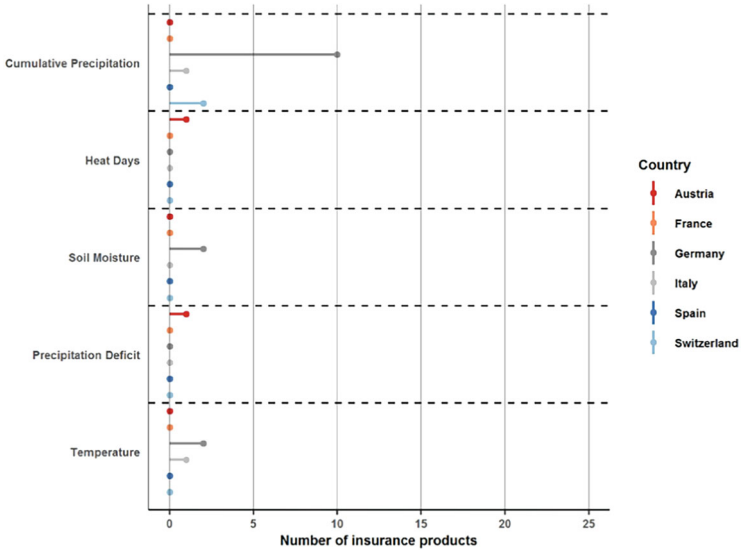


Fig. 3.4 Weather indexes used for weather index insurances. Source: Bucheli et al. [3]

3.4 The Agri-CAT Fund: A Basic Coverage to Cope with Catastrophic Weather Risks

The National Strategic Plan (NSP) 2023–2027 of the Italian Ministry of Agricultural, Sovereignty Food and Forestry provides some agricultural risk management tools subsidized by public contribution and available for the farmers: (1) insurance schemes and (2) mutual funds to cover damages on plants and animal production, farm structures and livestock farms due to adverse weather conditions, epizootic diseases, plant diseases, parasitic infestations, environmental emergencies; (3) mutual funds to cover income losses due to price volatility and market fluctuations (i.e., Income Stabilization Tool), limited to the poultry, sugar beet, durum wheat, cow and sheep milk, olive, fruit and vegetable, rice and pig farming sectors; (4) national mutual funds for catastrophic events to cope with damages due to extreme weather such as flood, frost, and drought. The latter represents an absolutely novelty, namely Agri-CAT National Mutual Fund for catastrophic weather events, established pursuant to Article 1, paragraph 515, of Law no. 234, December 30, 2021, which provides for all farms receiving direct payments of a basic mutual coverage to cope with weather catastrophic events (i.e., flood, frost, and drought), formally recognized by national public bodies. More specifically, flood is defined as a natural calamity that occurs as a result of heavy rainfall or overflow, caused by exceptional atmospheric events, affecting natural and artificial waterways and inundating surrounding areas, accompanied by the transport and deposition of usual and incoherent materials; frost occurs when the temperature drops below

0 °C; drought is an extraordinary lack of precipitation compared to the period's normal conditions, resulting in a decrease in soil moisture content below the critical moisture limit and depletion of water sources to the extent that even emergency irrigation measures are impossible (Agricultural Risk Management Plan, 2023). The aim is to improve the resilience of farming systems to climate change, also increasing the number of farms participating in risk management schemes and promoting a territorial and sectorial rebalancing of public support. In fact, the Agri-CAT fund will act in complementarity with the other risk management tools, especially with the insurance policies which will operate on the weather risks not covered by the Fund. Therefore, the exposure level of insurance companies will be lower and more sustainable in financial terms. Farmers who experience yield losses due to catastrophic weather events must present a claim report on the National Agricultural Information System (SIAN) platform to be compensated. The fund identifies the affected areas based both on maps developed using agrometeorological indicators provided by ISMEA and insurance adjuster activity, and it covers exclusively the production losses resulting from catastrophic weather events (i.e., flood, drought, and frost) specified in the annual plan that exceed the minimum threshold of 20% of the farmer's average annual production. The average annual production, identified in monetary values to provide a summary of the yields of different types of crops cultivated in the farms, is determined using "value indexes" provided by the Agricultural Risk Management Plan (PGRA) as the baseline for calculating compensations in the case of damages (Fig. 3.5).

The compensation amounts to the product of the index value and the areal damage percentage determined by insurance adjuster activity. The Agency for Agri-

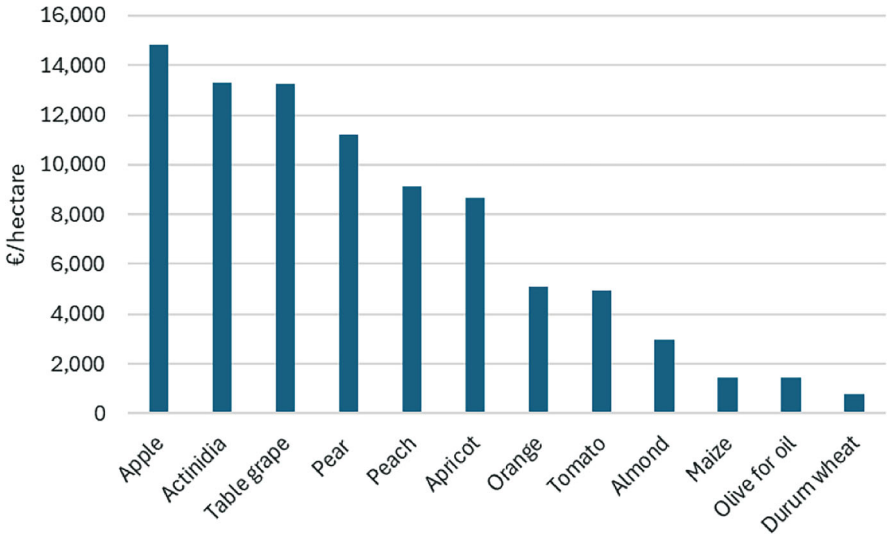


Fig. 3.5 Value indexes of the main crops for the Agri-CAT compensation. Source: PGRA, 2023

cultural Payments (AGEA), as the entity responsible to the financial compensation, verifies any overcompensation resulting from the accumulation of interventions from the Agri-CAT fund with other public or private risk management schemes, also ensuring that the compensation value does not exceed the maximum value of the production losses. More specifically, the Agri-CAT compensation is explained by the following formula:

$$C = f\left(w, t_p\right) + (l_c * v_c) - d$$

where the compensation C is function of the catastrophic weather events w when the threshold t exceeds the 20% of the annual farmer's production p, added by the production losses l per value index v provided by PGRA for the crop c, subtracted from the deductible d which may amount from 20% for permanent crops (except citrus and olive), and horticulture, to 30% for arable crops and other crops (including citrus and olive). The Agri-CAT fund is financed through a withholding tax of 3% of direct payments (in accordance with the Article 19 of Regulation (EU) 2021/2115, a State Member can decide to allocate up to 3% of the direct payments to be paid to a farmer valid as a private share for the activation of a risk management tool). Currently, out of 150,000 agricultural insurance contracts in Italy, approximately only 18,000 cover catastrophic risks. The Agri-CAT fund allows a coverage of around 700,000 farms receiving CAP direct payments from catastrophic weather events which may severely affect the crop yields. The financial allocation of the fund for 2023–2027 years is approximately 350 million euros per year of which around 105 million euros from the EAGF (derived from 3% withholding), and around 245 million euros from the EAFRD. In 2022, an experimental activity was conducted in Italy to validate the entire functioning of fund (e.g., risk coverage, monitoring of catastrophic events, claims management, area-based damage estimation). For this purpose, 12 crops representing the *Made in Italy* brand were selected, approximately one-third of Italy's agricultural production, covering around 4 million hectares with a value of over 10 billion euros: actinidia, almond, apple, apricot, durum wheat, industrial tomato, maize, olive for oil, orange, peach, pear, wine grape. Moreover, 13 test areas[2] were identified across the following provinces: Bari, Bolzano, Caserta, Catania, Chieti, Ferrara, Foggia, Latina, Mantova, Ravenna, Sondrio, Trento, and Verona (see the Table 3.1).

Clearly, some provinces are leaders for certain crop production, e.g., Bolzano and Trento produce together around the 60% of the national production of apples, while Latina and Foggia produced around 42% and 25% of the national production of Actinidia and industrial tomato for 2023 year, respectively. The experimental activity of Agri-CAT fund involved 85,000 farmers for a total of 435,000 cultivated hectares. All test areas were affected by catastrophic weather events as drought, flood, and frost. More specifically, the provinces more affected by drought were

[2] At the end of 2022, Ancona and Pesaro-Urbino provinces were added as test area due to flood events which affected Marche region in September 2022.

Table 3.1 Overview of Agri-CAT fund experimental activity and crop production data

Crop	Test area (province)	N. of CAT event	N. of exceeding the threshold	Share of crop production over the national crop production (%)	
				2018	2023
Actinidia	Latina	7	0	32.0	41.9
Almond	Bari	9	6	18.6	8.4
Apple	Bolzano	12	1	40.6	42.0
	Sondrio	6	0	1.4	1.5
	Trento	11	0	20.7	22.5
Apricot	Ravenna	4	3	8.6	10.3
Durum wheat	Foggia	11	3	17.2	16.3
Industrial tomato	Foggia	11	0	30.0	24.9
	Mantova	7	0	5.5	4.5
Maize	Mantova	7	3	5.1	5.6
Olive for oil	Bari	9	0	3.1	9.8
Orange	Catania	8	0	26.7	22.1
Peach	Caserta	8	3	15.2	35.2
Pear	Ferrara	3	2	28.5	7.9
Wine grape	Chieti	8	1	4.9	3.1
	Verona	9	1	7.2	6.8

Source: ISMEA [40]; elaboration from ISTAT, 2023

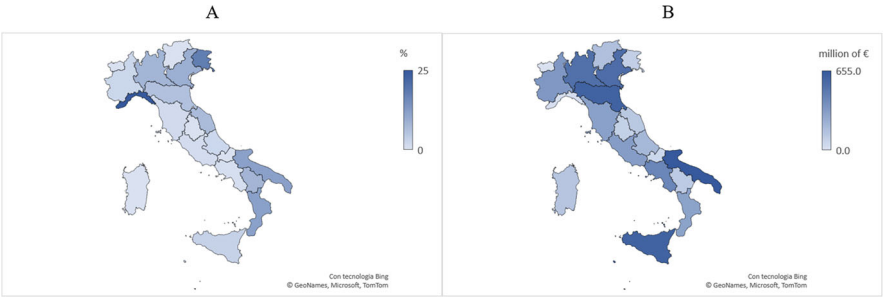


Fig. 3.6 Share of Agri-CAT compensation (**a**) and production losses due to CAT event (**b**). Source: ISMEA [40]

those in which crops lacked in emergency irrigation systems (e.g., Foggia for durum wheat with losses of 35%) or characterized by water scarcity (e.g., Chieti and Mantova for wine grapes and maize with losses of approximately of 50%). Regarding flooding, the provinces of Ancona and Pesaro-Urbino were the most affected, with hourly precipitation reaching peaks of 90 mm and losses of up to 100%. Moreover, the provinces of Ravenna, Caserta, and Bari experienced damages to apricots (around of 50%), peaches (around of 30%), and almonds (over 70%), respectively. Focusing on the technical balance sheet of the Agri-CAT fund a regional level, Liguria and Friuli-Venezia Giulia regions were the most indemnified regions with 25% and 17% of their gross saleable production, respectively (Fig. 3.6). Based on the detected damages across Italian regions, Puglia and Sicilia regions were the worst in terms of production losses due to catastrophic weather events, amounting to 655.3 and 601.4 million of euros, respectively, while the Italian agricultural losses amounted to 5.6 billion of euros (Fig. 3.6). Cereals and vegetables are the agricultural sector with greater production losses, i.e., around 1.2 and 1.6 billion of euros, respectively. The share of Agri-CAT compensation is higher for cereals and oil with 126.8 and 67.1 million of euros, respectively (Fig. 3.7).

A substantial equalization of indemnifications between geographic macrodivisions and production sectors emerged, in relation to the damages detected, unlike what has been observed in the insurance market, historically unbalanced (even more clearly on risks from catastrophic events) on territories in the north and on specific compartments, particularly on wine and fruit production. This would allow the activation of potential processes of local mutuality between crops in the same provinces or at least at the regional level. The experimentation highlighted the importance of defining the timeliness of the insurance appraisals with respect to the time of occurrence of weather event and, subsequently, with respect to the time of time of harvest. A monitoring system based on phenological and meteorological data (independent of damage claims), on "sentinel" farms randomly selected in homogeneous areas, would be efficient. As shown in Tappi et al. [25, 26], earliness and phenological phases are crucial information to reduce yield losses due to temperature shocks.

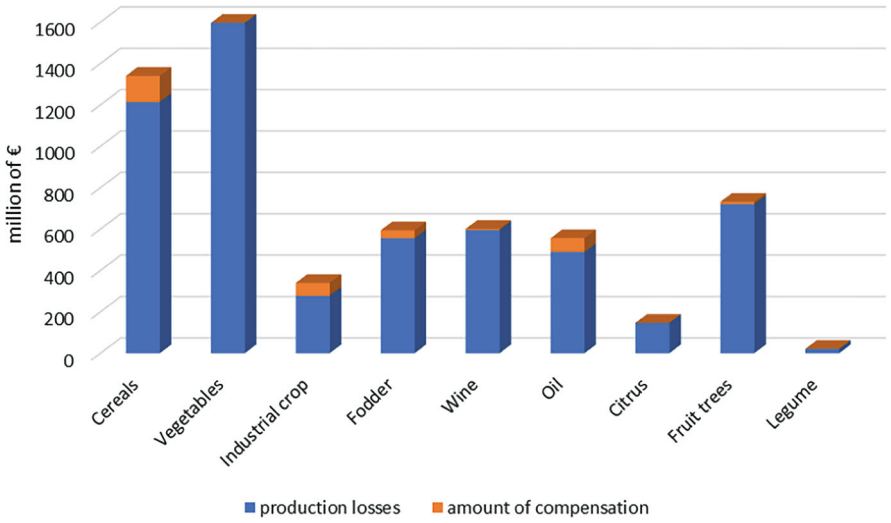


Fig. 3.7 Production losses and amount of Agri-CAT compensation across main agricultural sectors. Source: ISMEA [40]

3.5 Assessment of Yield Risks: The EU Case

We assess the yield risk for the cereal crops in the EU-27, using yield data form Eurostat (cereals to produce grain, including seed), from 2013 to 2022 (Table 3.2). The yield risk is measured as the expected (downside) deviation from the expected value. The expected value can be measured statistically. Following Zhang and Wang [41], the use the relative deviation from the trend.

$$g_t = \frac{\left(y_t - \hat{y}_t\right)}{\hat{y}_t}, \tag{3.1}$$

where g_t is the relative measure of the deviation from trend, y_t is the observed yield (t/ha), and $\hat{y}_t$ is the trend of the yield (t/ha). The trend can be calculated in several ways. We follow (and modify) two rules envisaged in EU Regulation 2021/2115: we use a three-year moving average is used as a trend, and a five-year moving average without the lowest and highest values (i.e., a trimmed mean). As we are looking at the historical data, the average values are calculated by including the current period (which would have been impossible for risk management under the regulation). The two candidate options for the 3-year average and 5-year trimmed mean are formally defined as:

$$\hat{y}_t = \frac{\sum_{\tau=0}^{2} y_{t-\tau}}{3} \quad \text{and} \tag{3.2}$$

Table 3.2 Average cereal yields for the EU countries, 2013–2022

Country	Average, t/ha	SD, t/ha	RSD
Belgium	8.44	0.78	0.09
Netherlands	8.24	0.44	0.05
Ireland	8.01	0.66	0.08
Germany	7.17	0.47	0.07
France	6.94	0.60	0.09
Austria	6.71	0.54	0.08
Denmark	6.51	0.66	0.10
Croatia	6.44	0.65	0.10
Slovenia	6.21	0.79	0.13
Czechia	5.82	0.38	0.07
Luxembourg	5.77	0.36	0.06
Hungary	5.75	0.83	0.14
Sweden	5.53	0.82	0.15
Slovakia	5.46	0.67	0.12
Italy	5.43	0.29	0.05
EU-27	5.41	0.17	0.03
Bulgaria	5.17	0.53	0.10
Portugal	4.81	0.35	0.07
Romania	4.43	0.94	0.21
Greece	4.08	0.31	0.08
Poland	4.08	0.50	0.12
Lithuania	4.03	0.44	0.11
Latvia	3.95	0.53	0.13
Spain	3.72	0.54	0.15
Estonia	3.70	0.71	0.19
Finland	3.62	0.43	0.12
Cyprus	1.72	0.86	0.50

$$\hat{y}_t = \frac{\sum_{\tau=0}^{4} y_{t-\tau} - \min\{y_{t-\tau}\}_{\tau=0}^{\tau=4} - \max\{y_{t-\tau}\}_{\tau=0}^{\tau=4}}{3}. \tag{3.3}$$

The probability of a hazard is defined by calculating the occurrence of time periods (years) with the relative deviation from the trend exceeding a certain value, $\lambda\hat{y}_t$, where coverage level $\lambda \in (0, 1]$ indicates which proportion of the expected yield is covered by the risk analysis. The most pessimistic measure is obtained by setting $\lambda = 1$ when any downward deviation from the trend is considered. The EU regulation mentioned above is related to a 20% deviation below the trend. The intermediate values can also be considered. The yield risk is calculated as a product of the expected hazard and a probability of obtaining a negative hazard:

$$R = E\left(h_t | y_t < \lambda\hat{y}_t\right) \Pr\left(y_t < \lambda\hat{y}_t\right). \tag{3.5}$$

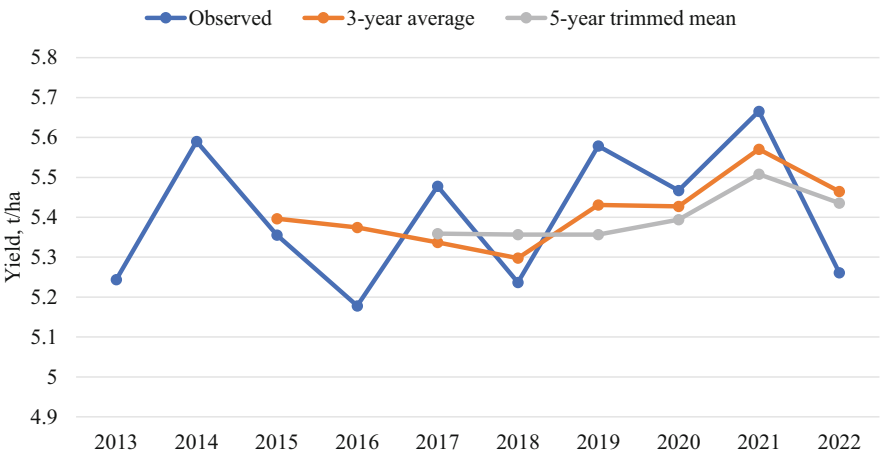


Fig. 3.8 The observed cereal yields and fitted trends for the EU-27 average value, 2013–2022

where the hazard is computed as follows: $h_t = \min\{0, g_t\}$.

The trends were calculated as described above, i.e., the 3-year average and 5-yuear trimmed mean were used. An illustrative example of the trends fitted for the EU-27 average is provided in Fig. 3.8. As the period needed for the calculation extends, more observations are lost at the initial periods. As one can note, the two trend lines intersected in between 2018 and 2019. This marked a generally upward swing in the observed data series. Thus, the 3-year average is more sensitive to such developments.

The expected hazard (given a loss in yield is observed) can be calculated for any level of coverage. The levels of 1, 0.95, 0.9, and 0.8 are chosen for the analysis (note that 0.8 corresponds to a 20% downside deviation from the trend). The resulting (conditional) expected hazards are summarised in Table 3.3.

The probability to observe a downward deviation is measured simply as the ratio of the time periods satisfying the condition compared to the total number of time periods covered. Alternatively, one could estimate an underlying density function and embark on integration. This is particularly relevant if the deviations from trend are assumed to follow a non-normal distribution.

The descriptive statistics are informative. Belgium, the Netherlands, and Ireland perform the best with the average cereal yields exceeding 8 t/ha over 2013–2022. These countries also show relatively low standard deviation. The measure of the relative standard deviation (RSD) is used to compare the countries in the sense of yield variability with respect to the average values. This measure considers both positive and negative deviation from the mean, and the trend is not considered.

The lowest expected yield loss (given a loss occurs) is observed for Portugal. Specifically, it is expected that cereal yield would fall below the trend value by some 1.4% assuming the 3-year moving average is used as a trend. In case the 5-year trimmed mean is applied, there is no hazard obtained as the observed values of the

Table 3.3 The expected hazard (proportion of the expected yield lost) obtained by using either a 3-year moving average or a 5-year trimmed mean as the expected yield for different coverage rates

	3-year average				5-year trimmed mean			
Country	$\lambda = 1$	$\lambda = 0.95$	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 1$	$\lambda = 0.95$	$\lambda = 0.9$	$\lambda = 0.8$
EU-27	−0.023				−0.027			
Belgium	−0.091	−0.134	−0.205	−0.205	−0.051	−0.063		
Bulgaria	−0.058	−0.165	−0.165		−0.124	−0.124	−0.186	
Czechia	−0.056	−0.076			−0.060	−0.088	−0.114	
Denmark	−0.099	−0.099	−0.181		−0.154	−0.154	−0.254	−0.254
Germany	−0.036	−0.077	−0.101		−0.049	−0.155	−0.155	
Estonia	−0.185	−0.185	−0.185	−0.255	−0.175	−0.175	−0.175	−0.235
Ireland	−0.067	−0.106	−0.106		−0.117	−0.117	−0.154	
Greece	−0.038	−0.087	−0.122		−0.065	−0.115	−0.115	
Spain	−0.102	−0.131	−0.168		−0.133	−0.178	−0.178	
France	−0.089	−0.128	−0.182		−0.037	−0.057		
Croatia	−0.039	−0.086	−0.120		−0.066	−0.171	−0.171	
Italy	−0.034	−0.107	−0.107		−0.049	−0.124	−0.124	
Cyprus	−0.264	−0.264	−0.369	−0.615	−0.035	−0.056	#DIV/0!	
Latvia	−0.079	−0.137	−0.182		−0.138	−0.138	−0.215	−0.215
Lithuania	−0.074	−0.093	−0.153		−0.086	−0.209	−0.209	−0.209
Luxembourg	−0.049	−0.116	−0.116		−0.047	−0.065		
Hungary	−0.088	−0.167	−0.272	−0.272	−0.197	−0.349	−0.349	−0.349
Netherlands	−0.032	−0.082			−0.028			
Austria	−0.035	−0.061			−0.035	−0.054		
Poland	−0.065	−0.085	−0.118		−0.088	−0.140	−0.140	
Portugal	−0.014							
Romania	−0.128	−0.165	−0.211	−0.314	−0.187	−0.270	−0.270	−0.270
Slovenia	−0.068	−0.129	−0.129		−0.090	−0.117	−0.142	
Slovakia	−0.079	−0.133	−0.133		−0.084	−0.119	−0.150	
Finland	−0.080	−0.179	−0.179	−0.216	−0.142	−0.193	−0.193	−0.218
Sweden	−0.154	−0.208	−0.208	−0.292	−0.245	−0.245	−0.245	−0.380

cereal yields in Portugal are above the 5-year trimmed means. This finding suggests that the use of the 3-year moving average may be more operational in empirical analysis and decision making. Indeed, the trimmed mean excludes some data from the calculations thus making the expected values less responsive to the underlying trends in the yields. A similar situation is observed for Cyprus, which showed a 26% expected hazard for a 3-year average and just 6.5% for a 5-year trimmed mean.

The highest expected hazard (ignoring Cyprus) is observed for Estonia (19% and 18%) and Sweden (15% and 25%) depending on the trend applied. The expected mean for deviations exceeding 5% compared to the expected yield were not obtained for some of the countries due to a relatively short period covered in this study. Note

Table 3.4 The probabilities of experiencing loss indicated by the level of coverage obtained by using either a 3-year moving average or a 5-year trimmed mean as the expected yield

	3-year average				5-year trimmed mean			
Country	$\lambda = 1$	$\lambda = 0.95$	$\lambda = 0.9$	$\lambda = 0.8$	$\lambda = 1$	$\lambda = 0.95$	$\lambda = 0.9$	$\lambda = 0.8$
EU-27	0.50	0.00	0.00	0.00	0.25	0.00	0.00	0.00
Belgium	0.38	0.25	0.13	0.13	0.38	0.25	0.00	0.00
Bulgaria	0.38	0.13	0.13	0.00	0.25	0.25	0.13	0.00
Czechia	0.38	0.25	0.00	0.00	0.38	0.25	0.13	0.00
Denmark	0.38	0.38	0.13	0.00	0.25	0.25	0.13	0.13
Germany	0.63	0.25	0.13	0.00	0.50	0.13	0.13	0.00
Estonia	0.38	0.38	0.38	0.13	0.25	0.25	0.25	0.13
Ireland	0.38	0.13	0.13	0.00	0.25	0.25	0.13	0.00
Greece	0.63	0.25	0.13	0.00	0.25	0.13	0.13	0.00
Spain	0.50	0.38	0.25	0.00	0.38	0.25	0.25	0.00
France	0.38	0.25	0.13	0.00	0.50	0.25	0.00	0.00
Croatia	0.63	0.25	0.13	0.00	0.38	0.13	0.13	0.00
Italy	0.63	0.13	0.13	0.00	0.50	0.13	0.13	0.00
Cyprus	0.38	0.38	0.25	0.13	0.25	0.13	0.00	0.00
Latvia	0.50	0.25	0.13	0.00	0.25	0.25	0.13	0.13
Lithuania	0.50	0.38	0.13	0.00	0.38	0.13	0.13	0.13
Luxembourg	0.50	0.13	0.13	0.00	0.25	0.13	0.00	0.00
Hungary	0.50	0.25	0.13	0.13	0.25	0.13	0.13	0.13
Netherlands	0.63	0.13	0.00	0.00	0.38	0.00	0.00	0.00
Austria	0.63	0.13	0.00	0.00	0.38	0.13	0.00	0.00
Poland	0.38	0.25	0.13	0.00	0.25	0.13	0.13	0.00
Portugal	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Romania	0.50	0.38	0.25	0.13	0.38	0.25	0.25	0.25
Slovenia	0.50	0.25	0.25	0.00	0.38	0.25	0.13	0.00
Slovakia	0.50	0.25	0.25	0.00	0.38	0.25	0.13	0.00
Finland	0.63	0.25	0.25	0.13	0.38	0.25	0.25	0.13
Sweden	0.38	0.25	0.25	0.13	0.25	0.25	0.25	0.13

that the average value for the EU-27 shows low expected hazard suggesting that pooling is a promising strategy for risk management in the EU crop sector.

The probabilities of experiencing a loss that falls beyond a chosen level of coverage are presented in Table 3.4. Looking at the EU average, one can note a 50% probability of observing a loss in yields which implies a symmetric distribution of the relative deviations from the trend in case a 3-year moving average is used. As for the 5-year trimmed mean, the probability of observing a loss is just 25%. Thus, the latter option for setting a trend seems less reasonable for risk management.

Even though Estonia and Sweden showed high expected hazards at $\lambda = 1$, the probability of such event is just 0.38. The countries with non-zero probabilities of catastrophic yield loss falling below the coverage level of 80% include Belgium, Estonia, Cyprus, Hungary, Romania, Finland, and Sweden in case a 3-year moving

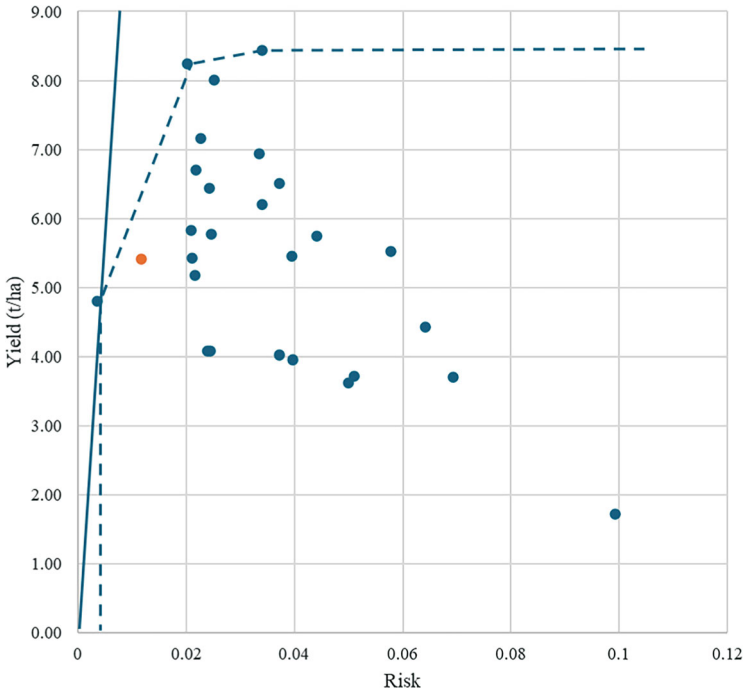


Fig. 3.9 Return to risks

average is applied. As for the 5-year trimmed mean trend, such countries include Denmark, Estonia, Latvia, Lithuania, Hungary, Romania, Finland, and Sweden. Thus, there exists an overlap across the two sets of countries, yet it is not a perfect one.

The countries can be ranked according to the returns to the risk. The yield can be normalized with respect to the risk following a well-known mean-variance approach, and operationalised via the data envelopment analysis [42, 43]. We outline the possibilities for assuming either variable or constant returns to risk when constructing the best practice frontiers. Figure 3.9 presents the case of the EU countries, assuming the cereal yield based on a 3-year moving average and the coverage level of 100%.

The solidline in Fig. 3.9 indicates the best practice frontier based on the highest returns to the risk observed in the sample. It is obtained as a line passing from the point of origin through the corresponding observation. The variable returns to risk assumption are imposed by assuming that benchmarks are set by considering the dominating observations only. Inefficient countries are located below the frontier and can approach it by adjusting risk and/or yield.

The constant returns to risk frontier is straightforward to be applied in the case of the constant returns to risk. Indeed, it is enough to find the maximum yield-to-risk

Table 3.5 Yield-risk efficiency for the EU countries ($\lambda = 1$)

Country	Average Yield	Risk (0)	Yield/Risk	Rank (Yield/Risk)	Rank (Yield)	Efficiency
EU-27	5.41	0.01	465.64	2	16	0.34
Belgium	8.44	0.03	247.44	10	1	0.18
Bulgaria	5.17	0.02	239.86	11	17	0.18
Czechia	5.82	0.02	277.99	7	10	0.20
Denmark	6.51	0.04	175.19	15	7	0.13
Germany	7.17	0.02	317.21	5	4	0.23
Estonia	3.70	0.07	53.39	26	25	0.04
Ireland	8.01	0.03	319.44	4	3	0.24
Greece	4.08	0.02	170.34	16	20	0.13
Spain	3.72	0.05	72.90	23	24	0.05
France	6.94	0.03	207.00	13	5	0.15
Croatia	6.44	0.02	265.03	8	8	0.20
Italy	5.43	0.02	257.82	9	15	0.19
Cyprus	1.72	0.10	17.36	27	27	0.01
Latvia	3.95	0.04	99.61	21	23	0.07
Lithuania	4.03	0.04	108.32	20	22	0.08
Luxembourg	5.77	0.02	234.52	12	11	0.17
Hungary	5.75	0.04	130.08	19	12	0.10
Netherlands	8.24	0.02	407.33	3	2	0.30
Austria	6.71	0.02	307.93	6	6	0.23
Poland	4.08	0.02	166.79	17	21	0.12
Portugal	4.81	0.00	1357.46	1	18	1.00
Romania	4.43	0.06	69.05	25	19	0.05
Slovenia	6.21	0.03	182.83	14	9	0.13
Slovakia	5.46	0.04	138.22	18	14	0.10
Finland	3.62	0.05	72.52	24	26	0.05
Sweden	5.53	0.06	95.53	22	13	0.07

ratio and use it as a numeraire when comparing each country against this ratio. The results are presented in Table 3.5.

Portugal shows full efficiency which is caused by extremely low risk value (one may expect an increase in case a longer time series were covered). Cyprus, Romania, and Estonia show the lowest efficiency levels. Note that the rankings of the EU countries based on the average yield and yield-to-risk ratio are different, e.g., in the case of Belgium. Therefore, the yield risk is an important measure when devising risk management policies in the EU crop farming.

3.6 Prospects of the Index-Based Insurance Schemes in the EU

The AGRI-CAT scheme offers an interesting case study to scale the index-based insurance schemes at the EU level, as for other successful risk management schemes

such as the IST [44]. Scaling up this initiative comes with several questions and concerns that need to be addressed in the next Common Agricultural Policy, following a long trend of reforms that have been able to improve the effectiveness of the EU scheme [45]. We list here some of the many issues that will have to be faced by the theorists of the EU CAT scheme, by the analysists of the crop insurance markets and by the policymakers that will design the national rules to operationalize the interventions.

The EU schemed must be scalable and adaptable to the national specificities, in terms of climate conditions, farming, market and trade structures, farmers representation, and culture. The scheme has to be peculiar and specific, but also inclusive as it should promote diversification and inclusivity within the agricultural sector, discouraging disparities or exclusionary practices that would undermine the effectiveness of the scheme. The EU CAT should also be economically, environmentally, and socially sustainable. Failure to do so would result in a short-lived instrument that will not survive the pressure of the most affected categories. Moreover, the EU CAT must be politically feasible and sustainable.

These challenges cannot be taken and solved in a single stage. Differently, it would be advisable to plan a stepwise implementation of the large-scale EU CAT, that may build on long-lasting experience in insurance schemes in major economies, particularly in the US [39].

3.7 Conclusion

Understanding climate change and future projections is useful to support the development of risk management tools tailored to the needs of farmers. Premium subsidies encourage farms to increase both crop acreage and insurance coverage [46]. Enlarging the insured area leads to higher insurance premium, while securitization through CAT bonds turned out to be an effective tool for reducing systems risk at reasonable costs [47].

The new CAP provides several risk management tools that may play a crucial role in enhancing the resilience of farming systems to climate change. These tools, such as crop insurance, may help the farmers to protect the crops and their incomes from damages due to extreme weather events that are more and more frequent, and impactful. Clearly, farmers need to be informed about the availability of these tools and encouraged to adopt them, and only through a combination of complementary risk management strategies (e.g., agricultural diversification together to crop insurances or mutual funds) may build more resilient farms. Likewise, it is important that stakeholders consider that the indicators are based on data that are imperfectly measured and that efforts are required to standardise data quality [35]. Furthermore, it is essential to involve the stakeholders, policymakers, researchers, professional association, and agronomist to develop effective risk management solutions and protect agricultural and environmental heritage for generations to come. It needs to tailor contractual schemes also taking into account the need to limit the Fund's

financial exposure, to encourage maximum integration with the insurance system, to encourage the spread of policies subsidized against weather and climate risks, to overcome territorial and sectoral asymmetries and to reverse the upward trend in insurance rates. Clearly, insurance does not compensate for the entire loss but represented an aid to the farmer to stay in the market.

Acknowledgements This study was carried out within the RESEARCH PROJECTS OF RELEVANT NATIONAL INTEREST (PRIN)—2020 Call (Prot. 20205L79R8) "Towards a holistic approach to SUStainable RISK management in agriculture (Sus-Risk)" and the RESEARCH PROJECTS OF RELEVANT NATIONAL INTEREST (PRIN) Call (Prot. P2022JTSFB) —2022 "REcovering the agrifood system from Shocks Induced by Labour Inputs, ENergy, Climate Extremes" (Resilience).

Bibliography

1. Malhi, G. S., Kaur, M., & Kaushik, P. (2021). Impact of climate change on agriculture and its mitigation strategies: A review. *Sustainability, 13*(3), 1318.
2. EEA. (2022). *Economic Losses and Fatalities from Weather-and Climate-Related Events in Europe*. Available at: https://www.eea.europa.eu/publications/economic-losses-and-fatalities-from. Last accessed: 25 July 2023.
3. Bucheli, J., Conrad, N., Wimmer, S., Dalhaus, T., & Finger, R. (2023). Weather insurance in European crop and horticulture production. *Climate Risk Management, 100525*.
4. Walsh, J. E., Ballinger, T. J., Euskirchen, E. S., Hanna, E., Mård, J., Overland, J. E., Vihma, T., et al. (2020). Extreme weather and climate events in northern areas: A review. *Earth-Science Reviews, 209*, 103324.
5. Mäkinen, H., Kaseva, J., Trnka, M., Balek, J., Kersebaum, K. C., Nendel, C., Kahiluoto, H., et al. (2018). Sensitivity of European wheat to extreme weather. *Field Crops Research, 222*, 209–217.
6. Zampieri, M., Ceglar, A., Manfron, G., Toreti, A., Duveiller, G., Romani, M., Djurdjevic, V., et al. (2019). Adaptation and sustainability of water management for rice agriculture in temperate regions: The Italian case-study. *Land Degradation & Development, 30*(17), 2033–2047.
7. Diakakis, M. (2012). Rainfall thresholds for flood triggering. The case of Marathonas in Greece. *Natural Hazards, 60*, 789–800.
8. Kaiser, M., Günnemann, S., & Disse, M. (2020). Providing guidance on efficient flash flood documentation: an application based approach. *Journal of Hydrology, 581*, 124466.
9. Gupta, A., Rico-Medina, A., & Caño-Delgado, A. I. (2020). The physiology of plant responses to drought. *Science, 368*(6488), 266–269.
10. Iqbal, M. S., Singh, A. K., & Ansari, M. I. (2020). Effect of drought stress on crop production. *New Frontiers in Stress Management for Durable Agriculture*, 35–47.
11. Barlow, K. M., Christy, B. P., O'leary, G. J., Riffkin, P. A., & Nuttall, J. G. (2015). Simulating the impact of extreme heat and frost events on wheat crop production: A review. *Field Crops Research, 171*, 109–119.
12. Rezaei, E. E., Webber, H., Gaiser, T., Naab, J., & Ewert, F. (2015). Heat stress in cereals: Mechanisms and modelling. *European Journal of Agronomy, 64*, 98–113.
13. Gardiner, B., Berry, P., & Moulia, B. (2016). Wind impacts on plant growth, mechanics and damage. *Plant Science, 245*, 94–118.
14. Vroege, W., & Finger, R. (2020). Insuring weather risks in European agriculture. *EuroChoices, 19*(2), 54–62.

15. Devot, A., Royer, L., Arvis B., Deryng, D., Caron Giauffret, E., Giraud, L., Ayral, V., & Rouillard, J. (2023). *Research for AGRI Committee – The impact of extreme climate events on agriculture production in the EU*. European Parliament, Policy Department for Structural and Cohesion Policies, Brussels.
16. Coletta, A., Giampietri, E., Santeramo, F. G., Severini, S., & Trestini, S. (2018). A preliminary test on risk and ambiguity attitudes, and time preferences in decisions under uncertainty: Towards a better explanation of participation in crop insurance schemes. *Bio-based and Applied Economics, 7*(3), 265–277.
17. European Commission. (2017). *Study on Risk Management in EU Agriculture*.
18. Meuwissen, M. P., Mey, Y. D., & van Asseldonk, M. (2018). Prospects for agricultural insurance in Europe. *Agricultural Finance Review, 78*(2), 174–182.
19. Arata, L., Cerroni, S., Santeramo, F. G., Trestini, S., & Severini, S. (2023). Towards a holistic approach to sustainable risk management in agriculture in the EU: A literature review. *Bio-based and Applied Economics, 12*(3), 165–182.
20. Santeramo, F. G. (2018). Imperfect information and participation in insurance markets: evidence from Italy. *Agricultural Finance Review, 78*(2), 183–194.
21. Santeramo, F. G. (2019). I learn, you learn, we gain experience in crop insurance markets. *Applied Economic Perspectives and Policy, 41*(2), 284–304.
22. Barnett, B. J., & Mahul, O. (2007). Weather index insurance for agriculture and rural areas in lower-income countries. *American Journal of Agricultural Economics, 89*(5), 1241–1247.
23. Shirsath, P., Vyas, S., Aggarwal, P., & Rao, K. N. (2019). Designing weather index insurance of crops for the increased satisfaction of farmers, industry and the government. *Climate Risk Management, 25*, 100189.
24. Belissa, T., Bulte, E., Cecchi, F., Gangopadhyay, S., & Lensink, R. (2019). Liquidity constraints, informal institutions, and the adoption of weather insurance: A randomized controlled Trial in Ethiopia. *Journal of Development Economics, 140*, 269–278.
25. Tappi, M., Carucci, F., Gatta, G., Giuliani, M. M., Lamonaca, E., & Santeramo, F. G. (2023a). Temporal and design approaches and yield-weather relationships. *Climate Risk Management, 40*, 100522.
26. Tappi, M., Carucci, F., Gagliardi, A., Gatta, G., Giuliani, M. M., & Santeramo, F. G. (2023b). Crop varieties, phenological phases and the yield-weather relationship: evidence from the Italian durum wheat production. *Bio-based and Applied Economics*.
27. Conradt, S., Finger, R., & Spörri, M. (2015). Flexible weather index-based insurance design. *Climate Risk Management, 10*, 106–117.
28. Bucheli, J., Dalhaus, T., & Finger, R. (2021). The optimal drought index for designing weather index insurance. *European Review of Agricultural Economics, 48*(3), 573–597.
29. Abdi, M. J., Raffar, N., Zulkafli, Z., Nurulhuda, K., Rehan, B. M., Muharam, F. M., Tangang, F., et al. (2022). Index-based insurance and hydroclimatic risk management in agriculture: A systematic review of index selection and yield-index modelling methods. *International Journal of Disaster Risk Reduction, 67*, 102653.
30. Ceballos, F., Kramer, B., & Robles, M. (2019). The feasibility of picture-based insurance (PBI): Smartphone pictures for affordable crop insurance. *Development Engineering, 4*, 100042.
31. Dalhaus, T., Musshoff, O., & Finger, R. (2018). Phenology information contributes to reduce temporal basis risk in agricultural weather index insurance. *Scientific Reports, 8*(1), 1–10.
32. Clement, K. Y., Botzen, W. W., Brouwer, R., & Aerts, J. C. (2018). A global review of the impact of basis risk on the functioning of and demand for index insurance. *International Journal of Disaster Risk Reduction, 28*, 845–853.
33. Erec Heimfarth, L., & Musshoff, O. (2011). Weather index-based insurances for farmers in the North China Plain: An analysis of risk reduction potential and basis risk. *Agricultural Finance Review, 71*(2), 218–239.
34. Schmidt, L., Odening, M., Schlanstein, J., & Ritter, M. (2022). Exploring the weather-yield nexus with artificial neural networks. *Agricultural Systems, 196*, 103345.
35. Morsink, K., Clarke, D., & Mapfumo, S. (2016). *How to measure whether index insurance provides reliable protection*. World Bank Policy Research Working Paper (7744).

36. Tadesse, M. A., Shiferaw, B. A., & Erenstein, O. (2015). Weather index insurance for managing drought risk in smallholder agriculture: Lessons and policy implications for sub-Saharan Africa. *Agricultural and Food Economics, 3*, 1–21.
37. Enenkel, M., Osgood, D., Anderson, M., Powell, B., McCarty, J., Neigh, C., Brown, M., et al. (2019). Exploiting the convergence of evidence in satellite data for advanced weather index insurance design. *Weather, Climate, and Society, 11*(1), 65–93.
38. Santeramo, F. G., Lamonaca, E., Maccarone, I., & Tappi, M. (2024). Extreme weather events and crop insurance demand. *Heliyon, 10*(7), e27839.
39. Santeramo, F. G., & Ford Ramsey, A. (2017). Crop Insurance in the EU: Lessons and Caution from the US. *EuroChoices, 16*(3), 34–39.
40. ISMEA. (2023). *Il Fondo Mutualistico Nazionale per la copertura dei danni catastrofali*. Available at https://www.ismea.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/12034. Last accessed 25 July 2023.
41. Zhang, Q., & Wang, K. (2010). Evaluating production risks for wheat producers in Beijing. *China Agricultural Economic Review, 2*(2), 200–211.
42. Lim, S., Oh, K. W., & Zhu, J. (2014). Use of DEA cross-efficiency evaluation in portfolio selection: An application to Korean stock market. *European Journal of Operational Research, 236*(1), 361–368.
43. Joro, T., & Na, P. (2006). Portfolio performance evaluation in a mean–variance–skewness framework. *European Journal of Operational Research, 175*(1), 446–461.
44. Cordier, J., & Santeramo, F. (2020). Mutual funds and the Income Stabilisation Tool in the EU: Retrospect and Prospects. *EuroChoices, 19*(1), 53–58.
45. Santeramo, F. G., Russo, I., & Lamonaca, E. (2023). Italian subsidised crop insurance: What the role of policy changes. *Q Open, 3*(3), qoac031.
46. Yu, J., Smith, A., & Sumner, D. A. (2018). Effects of crop insurance premium subsidies on crop acreage. *American Journal of Agricultural Economics, 100*(1), 91–114.
47. Shen, Z., & Odening, M. (2013). Coping with systemic risk in index-based crop insurance. *Agricultural Economics, 44*(1), 1–13.
48. Michel-Kerjan, E. O. (2010). Catastrophe economics: the national flood insurance program. *Journal of Economic Perspectives, 24*(4), 165–186.
49. Moore, F. C., & Lobell, D. B. (2014). Adaptation potential of European agriculture in response to climate change. *Nature Climate Change, 4*(7), 610–614.

Chapter 4
Avocado Production Index Insurance: An Application of Credibility Theory on Heterogeneous Data

Hirbod Assa

Abstract This chapter focuses on assessing avocado production index insurances and investigates insurance pricing utilizing credibility theory on a heterogeneous data set. The paper presents a methodology for analyzing and designing insurances, specifically addressing the challenges that arise when dealing with a data sets with varying characteristics. To enhance the reliability of the results, the analysis modifies Bühlmann's credibility theory to refine parameter distributions. Considering data sets from different countries on avocado production, this chapter provides global and local premium rates for each country, revealing rates based on historical trends. The study also proposes a two-layer policy for insurances that covers production return risks, incorporating both a standard deduction and a preventive measure to mitigate moral hazard risks.

4.1 Introduction

Agricultural insurances is one of the oldest and most active areas of research in actuarial science, see [1, 2]. There are two major risks in the agricultural industry: the production risk, and the risk of the prices. To manage these risks, there are different categories of agricultural insurances as listed below

- crop insurances [3, 4];
- revenue insurances; [5, 6], [7], [8];
- index insurances [9];
- price index insurances [10].

Due to lack of futures market, and also low cross correlation of the avocado prices with other commodities, the risk management of avocado is always a concern for the farmers. Likewise, in the literature the risk of avocado businesses is poorly studied, which warrants a separate study to propose a kind of insurance that can be used

H. Assa (✉)
Model Library, London, UK

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_4

worldwide. In this chapter, we propose an index insurance that uses production data on avocado as a trigger, reported by Food and Agriculture Organization (FAO) . Our goal is to develop a pricing strategy that effectively accounts for the risks and uncertainties associated with avocado production across different countries with a high data heterogeneity. Avocado production is influenced by numerous factors, including climate variability, pest outbreaks, and agricultural practices, which significantly impact annual yields. These dynamics necessitate a robust risk management model that accurately accounts for the probability and intensity of potential production shortfalls. We adopt an intensity-frequency framework, under model uncertainty and utilize credibility theory to better understand the risk of low yield.

Index insurance literature is an active research area in the literature. Here we briefly review most recent works in the field of agricultural index insurances in actuarial science. First, [10] explores the development of price index insurances. Their study stresses the need for tailored insurance solutions that address the specific volatility and risk profiles inherent in agricultural markets, thereby enhancing the financial resilience of farmers. Brock Porth et al. [11] demonstrate the potential of integrating advanced data analytics and satellite imagery to improve the accuracy and reliability of index-based insurance products. The forthcoming study by [12] presents a neural network-based approach to managing weather index insurance by providing more sophisticated tools for insurers to reduce the basis risk. Finally, [13] propose an improved index insurance design and yield estimation method using a dynamic factor forecasting approach. Their paper highlights the importance of dynamic modeling techniques in enhancing the precision of yield predictions, which is crucial for setting fair and effective insurance premiums.

Credibility theory also has a developed literature for agricultural application [14]. By combining historical loss experience with industry-wide data, credibility theory helps insurers strike a balance between the reliability of individual claims experience and the broader risk profile of the agricultural sector. This allows insurance companies to develop fair and financially sustainable pricing structures. Particularly, credibility theory has been considered in agricultural insurances as a powerful tool that can ultimately benefit both farmers and insurers in better estimating the insurance prices. Porth et al. [15] used a modified credibility approach incorporating the Erlang mixture distribution and liability weighted LCR improves reinsurance pricing by accurately modeling data tails and providing a more conservative and scientific approach; [16] enhanced our understanding of livestock insurance for mortality risk and improved modeling and computation methods through credibility analysis, providing improved estimates for livestock mortality insurance premiums and [17] improved reinsurance pricing framework integrating crop yield forecasting with credibility estimation.

The assumption in all papers is that the data have been validated for its quality, however, we have realized that in most of the cases particularly in agricultural insurances this cannot be the case. This needs to be addressed and for that reason agricultural insurance is a fruitful area of research to propose new methods of credibility.

In study we deal with a heterogeneous data set of a worldwide avocado production data. The data has covered production data of avocado for different countries and some different areas, covering different years. However, the range of the years coverage are very different; some can span over 60 years and some just for 10 years. This needed to be addressed by developing a modification of the Bühlmann's credibility theory that allows for such a heterogeneity. This is essentially done for the purpose of obtaining the average frequency and intensity of losses.

4.2 Overview of the Methodology and Data

In this chapter we follow a methodology that can be used for any other index insurance that is defined on the production level of a commodity.

4.2.1 Methodology

Here we outline the methodology and then throughout the chapter we will use

- **Actuarial Methodology:** We consider an intensity/frequency (a.k.a. severity/frequency) approach. Both the intensity (magnitude of production loss) and frequency (occurrence rate of such events) are modeled using exponential distributions. This choice is driven by the distribution's simplicity and empirical relevance, which facilitates parameter estimation and subsequent risk analyses. Notably for exponential distribution we need only to obtain the mean of the distribution.
- **Empirical Analysis:** Initial steps involve calculating the mean of the intensity and frequency of avocado production losses across all reporting countries. Given the diversity in agricultural conditions, we have a variety of means for both intensity and frequency that needs modeling. Modeling the means ensures a more accurate reflection of the data. The models are optimized by selecting the highest R^2 values for a model fitted to the survival functions of the means.
- **Extension of Bühlmann's Credibility Theory:** To better handle data heterogeneity, we extend Bühlmann's credibility theory to heterogeneous data. This method adjusts the means based on the variability and volume of data available for each country, refining our risk estimates.
- **Model Fitting to Adjusted Means**: After adjust the intensity and frequency means, we fit a suitable statistical models to these new data sets. These models help define the distribution of the modified expected losses, which is crucial for calculating accurate insurance premiums.
- **Credibility Price Method:** We calculate the adjusted rates for each country and compare it with their initial rate.

- **Uncertainty Premium Calculation:** Given the distribution of adjusted means for intensity and frequency, we adopt a conservative approach for premium calculation due to model uncertainty. We use higher quantile values from the distributions of intensity and frequency means. The premium is derived by multiplying these quantile values, reflecting a higher risk scenario and ensuring adequate coverage for intense and frequent loss events.

4.2.2 The Countries

In the following, we analyze the data of the annual avocado production reported to FAO for 65 countries all across the globe from 1960 to 2022, in Table 4.1. The number of the years that the data is available for each country is also indicated in brackets next to the name of the countries. As you can see the challenge of this data set is that, it does not form a tabular table and the length of the data for each country is different. The reason is that the consumption of avocado has just in the recent decade increased in countries that usually would not use them as an ingredient of their traditional food.

4.3 Insurance Designing and Pricing

The insurance design can be in different forms, e.g., on the crop losses (peril insurance), on the revenue losses (revenue insurance) or on prices (e.g., derivatives). In this chapter due to the existing data we have decided to design an index insurance based on the production.

4.3.1 Insurance Coverage and Retention Level

Suppose we take a retention level ρ in the range $(0, 1)$. A farmer aims to ensure that their production is at least $\rho \times 100\%$ of the previous year's production. Due to inherent risks, achieving this target is technically unattainable. Therefore, the farmer must procure insurance that guarantees revenue to meet this threshold. Motivated by this an index insurance is structured to cover losses up to $\rho \times 100\%$ of the shortfall in the next year's market production compared to the current year's production. In essence, the insurance is designed as follows:

$$Z_{t+1}^{\rho} = P_{t+1} \times \begin{cases} 0, & Production_{t+1} > \rho \times Production_t \\ \rho \times Production_t - Productin_{t+1}, & Production_{t+1} \leq \rho \times Production_t \end{cases},$$

Table 4.1 Name of the countries and the number of available data for each (in brackets)

Argentina (59)	Australia (55)	Bahamas (12)	Barbados (49)	Bolivia (57)
Bosnia and Herzegovina (15)	Brazil (61)	Cameroon (55)	Central African Republic (52)	Chile (59)
China (29)	Colombia (60)	Congo (42)	Cook Islands (43)	Costa Rica (53)
Cote d'Ivoire (43)	Cuba (56)	Cyprus (43)	Democratic Republic of Congo (57)	Dominica (44)
Dominican Republic (61)	East Timor (28)	Ecuador (61)	El Salvador (54)	Eswatini (25)
Ethiopia (22)	France (34)	French Polynesia (33)	Ghana (45)	Greece (38)
Grenada (57)	Guatemala (58)	Guyana (57)	Haiti (56)	Honduras (53)
Indonesia (56)	Israel (60)	Jamaica (47)	Kenya (38)	Lebanon (24)
Madagascar (50)	Mexico (61)	Morocco (37)	New Zealand (48)	Palestine (26)
Panama (60)	Paraguay (60)	Peru (61)	Philippines (60)	Polynesia (60)
Puerto Rico (59)	Rwanda (19)	Saint Lucia (60)	Samoa (53)	Seychelles (35)
South Africa (61)	South America (61)	Spain (61)	Sri Lanka (32)	Trinidad and Tobago (60)
Tunisia (23)	Turkey (33)	United States (61)	Venezuela (61)	Zimbabwe (30)

or,

$$Z_{t+1}^{\rho} = P_{t+1} \max\left\{\rho \times Production_t - Productin_{t+1}, 0\right\},$$

where P_{t+1} is the next period prices. Production here acts as in index. In summary the insurance pays if the production of this year is less than $\rho \times 100$ percent of the last year production.

As a result, the revenue pay-off is

$$\begin{aligned} pay-off_{t+1} &= P_{t+1} \times Production_{t+1} + Z_{t+1}^{\rho} \\ &= P_{t+1} \max\left\{\rho \times Production_t, Productin_{t+1}\right\}. \end{aligned}$$

In this case, the farmers pay-off is not less than $\rho \times 100$ production.

As such the premium of the policy is

$$Premium = E_t\left(Z_{t+1}^{\rho}\right),$$

where E_t is the conditional expectation on time t. Note that for simplicity without confusion we may drop t later. In terms of pricing what we need to know about this insurance is the discounted expected value:

$$E_t\left(Z_{t+1}^{\rho}\right) = E_t\left(P_{t+1} \max\left\{\rho \times Production_t - Productin_{t+1}, 0\right\}\right).$$

To deal with this value as the premium, we must consider a price process $P_t, t = 0, 1, 2, ...$ While we could have fitted a time series to the price process, we opt for a simpler model that enables us to price the insurance more effectively. Let us assume that the discounted process follows a martingale process, i.e.,

$$E\left(P_{t+1}\right) = e^r P_t,$$

where r is the interest rate. This is in line with the well-known martingale process (such as risk-free geometric Brownian motion) prices model (which is nothing but the Black-Scholes model). Using this in our pricing formula we get

$$\begin{aligned} E_t(Z_{t+1}^{\rho}) &= E_t\left(P_{t+1} \max\left\{\rho \times Production_t - Production_{t+1}, 0\right\}\right) \\ &= e^r P_t \times E_t\left(\max\left\{\rho \times Production_t - Production_{t+1}, 0\right\}\right) \\ &= e^r P_t \times Production_t \times E_t\left(\max\left\{\rho - \frac{Production_{t+1}}{Production_t}, 0\right\}\right) \\ &= e^r P_t \times Production_t \times E_t\left(\max\left\{\rho - (1 + Return_t), 0\right\}\right). \end{aligned}$$

Here $Return_t = \frac{Production_{t+1} - Productin_t}{Production_t}$. As a result, we get:

$$\frac{E_t(Z_{t+1}^{\rho})}{e^r P_t \times Production_t} = E_t\left(\max\{\rho - (1 + Return_t), 0\}\right).$$

Therefore, the production can be written in terms of the revenue:

$$\frac{E_t(Z_{t+1}^{\rho})}{e^r Revenue_t} = E_t\left(\max\{\rho - (1 + Return_t), 0\}\right).$$

As one can see if $\rho < 1 + Return_t$ there would be no compensation. For instance, if the coverage is for 70% then if the return loss is not less than 30% the product would not cover anything. Let us denote $I_{t+1} = \max\{\rho - (1 + Return_t), 0\}$, $i_t = E_t\left(I_{t+1}\right)$, then the premium is given by

$$Premium = e^r Revenue_t \times i_t.$$

Of course, we always can apply some safety loading factor LR:

$$\textit{Premium modified by loss ratio} = e^r Revenue_t \times \frac{i_t}{LR}.$$

For simplicity we always assume $LR = 1$ in the sequel.

4.3.2 Stop and Two Layer Policy

To mitigate the over-hedging, it is essential to incorporate a stop level. The insurance product should be structured in a way that the farmer's payoff is not unconditionally guaranteed by the insurance for every loss; rather, the insurance coverage ceases at a certain level. Let $0 < \sigma < \rho$, and consider the following product: if today's production falls below ρ times 100% of the last year's production, the insurance covers all losses below ρ times 100% minus σ times 100%. This two layer insurance contract, that we will denote by $Z_{t+1}^{\rho,\sigma}$, is given by:

$$Z_{t+1}^{\rho,\sigma} = P_{t+1} \times \begin{cases} 0, & \rho \times Production_t < Production_{t+1} \\ \rho \times Production_t - Production_{t+1}, & \sigma \times Production_t \leq Production_{t+1} \\ & \leq \rho \times Production_t \\ \rho \times Production_t - \sigma \times Production, & \sigma \times Production < Production_{t+1} \end{cases}.$$

It is clearly the case that $Z_{t+1}^{\rho,\sigma} = Z_{t+1}^{\rho} - Z_{t+1}^{\sigma}$. The pay-ff of the farmers income is that any production loss between $\sigma \times 100$ and $\rho \times 100$ percent of the last year product is guaranteed but below $\sigma \times 100$ the risk is shared by farmer.

In terms of the premium then it is very easy as one needs to note that

$$Premium = E\left(Z_{t+1}^{\rho,\sigma}\right) = E\left(Z_{t+1}^{\rho}\right) - E\left(Z_{t+1}^{\sigma}\right) = Premium^{\rho} - Premium^{\sigma}.$$

4.3.3 Actuarial Analysis on Production Data

In this chapter, we explore a standard actuarial analysis including two key aspects:

1. Intensity: Assessing the intensity of the damages.
2. Frequency: Examining the frequency at which damages occur.

To enhance the robustness of our analysis, we employ credibility theory to refine the outcomes related to average intensity and frequency. However, the enhancement of premiums through credibility theory necessitates a tabular format of losses, which is currently absent in our data-set. Consequently, constrained by the available data, we opt to model and design the insurance based on avocado production return. Moreover, confronted with a heterogeneous and limited data-set, we choose to anchor our analysis on model uncertainty within credibility theory, presenting a heterogeneous adaptation of the theory in this chapter.

Hence, the main points of this paper are the production return and model uncertainty. The data utilized comprises total production figures for various countries, as reported by the FAO in tons. Given the heterogeneous nature of the data, traditional statistical analyses are not feasible, and there exists significant model uncertainty. Nevertheless, we must still consider models that can assist us in risk analysis in the face of this uncertainty. An often-utilized model for intensity and frequency is the exponential distribution:

$$I \sim \exp\left(\frac{1}{\theta_I}\right),$$

$$F \sim \exp\left(\frac{1}{\theta_F}\right).$$

This implies that, for modeling purposes, there is only one parameter of interest - the mean of the data. Consequently, the data mean is the essential parameter considered in this analysis. However, alternative models can be considered, where the mean suffices for fitting the model. Specifically, the risk analysis will be centered on the means obtained from various countries/regions.

Let's denote two independent random variables, I and F, representing the intensity and frequency of losses, respectively. Since we are focusing on the return of the production, we can easily consider the following:

$$I_t = -Return_t^{\rho} \Big| \, Return_t^{\rho} \leq 0,$$

$$F_t = 1_{\{Return_t^{\rho} \leq 0\}},$$

where $Return_t^{\rho} = (1 + Return_t) - \rho$. As such we expect every year to see the expected losses be given by,

$$E(L_t) = E(I_t) \times E(F_t) = \theta_I \times \theta_F.$$

The values for the premiums which is the production of the expected intensity and expected frequency for all countries are calculated and reported in Table 4.2.

As it is not very clear how the exact value for the θ_I, θ_F would be calibrated, we treat θ_I, θ_F as random variables (similar to Bayesian approach) and then will use a model uncertainty approach on this model. More precisely, we can consider all the values of the form $F_{\theta_I}^{-1}(\beta) F_{\theta_F}^{-1}(\gamma)$, where q is the quantile and $(1-\beta)(1-\gamma) = (1-\alpha)$, (or $\alpha = \beta + \gamma - \beta\gamma$) where α is the level of risk tolerance (e.g., $\alpha = 0.90$). This is closely related to the concept of multivariate quantiles (see [18])

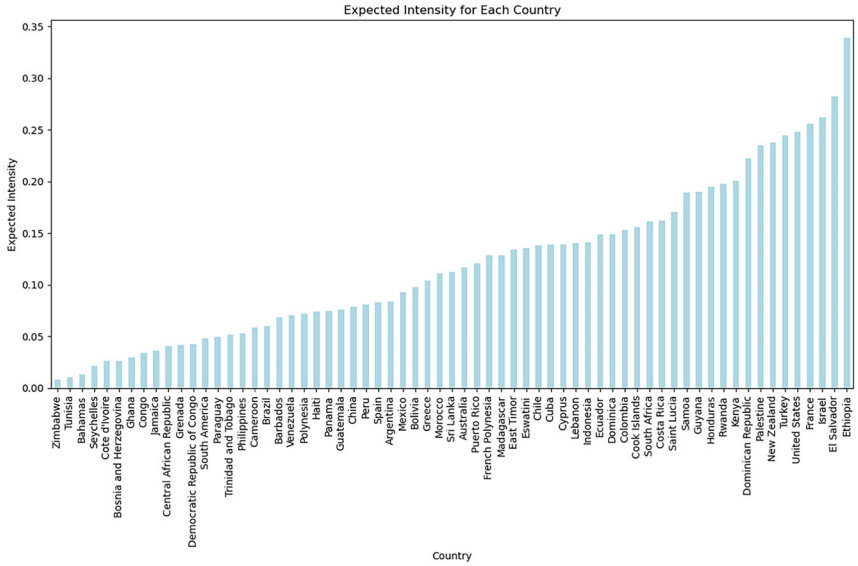
4.3.4 Modeling the Expected Intensity and Frequency

The estimation based on what is formulated above is done for all countries. However, we are interested in knowing the distribution of $E(I) = \theta_I$ and $E(F) = \theta_F$ for the modeling purpose. Here we depict the expected intensity and the survival functions of the estimated means of the countries in Fig. 4.1.

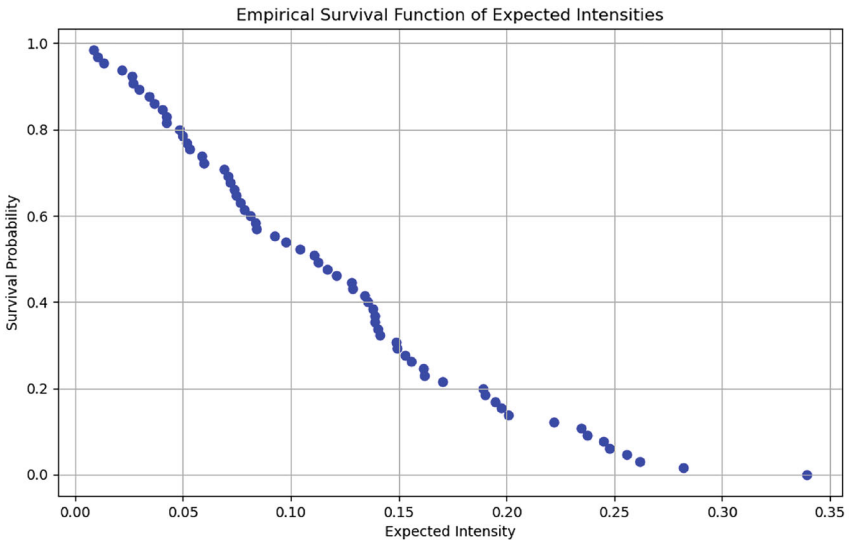
The same has been plotted for the frequency presented in Fig. 4.2.

Our model selection criteria are as follows:

1. Consider three models, being prioritized as Linear, Exponential, and Power distribution.
2. We fit these models to the survival functions and calculate their fitted R^2.
3. If the best model (with largest R^2) is at least 10% greater than the second best model (with second largest R^2), we pick the best model.
4. Otherwise, among the top 10% of R^2 use the priority list provided in 1.

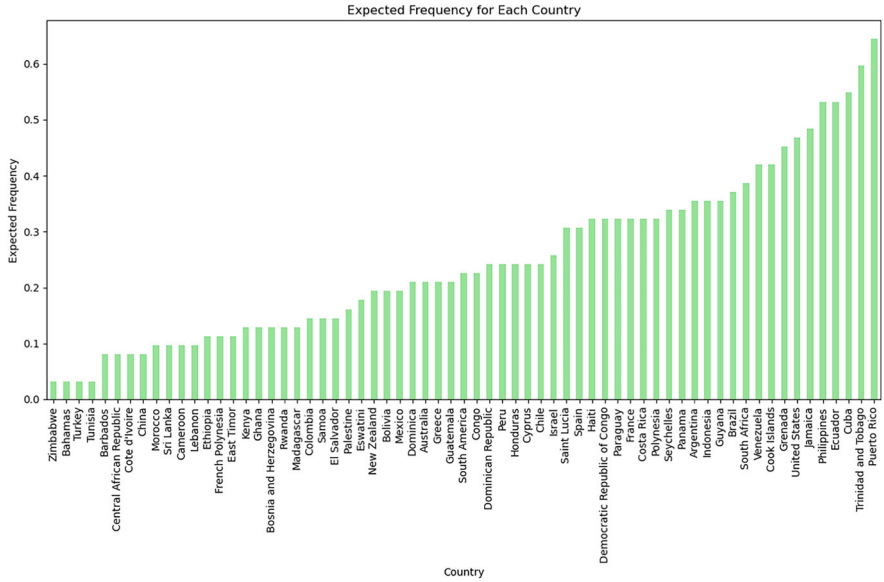


(a)

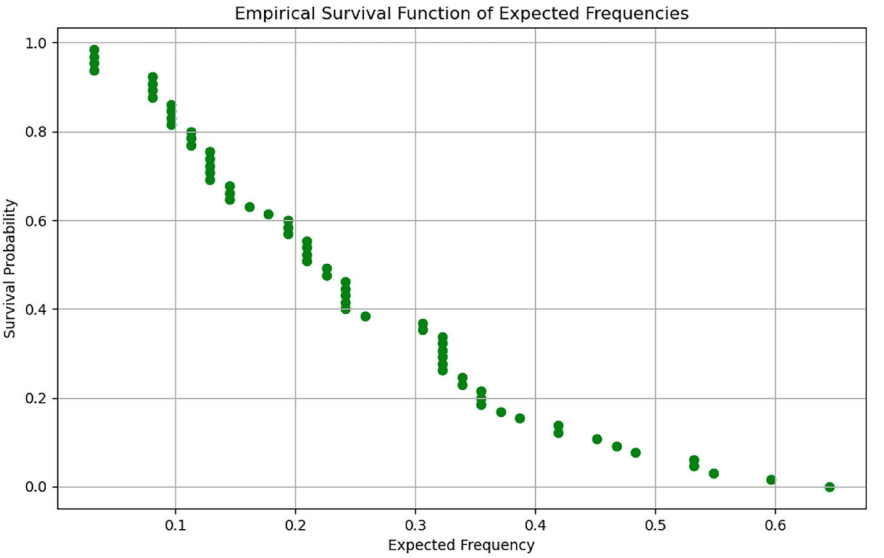


(b)

Fig. 4.1 The expected frequencies and the associated survival fit for case $\rho = 1$. (**a**) The empirical expected frequencies for the case $\rho = 1$. (**b**) The empirical survival function of the expected intensity for $\rho = 1$



(a)



(b)

Fig. 4.2 The expected frequencies and the associated survival fit for case $\rho = 1$. (**a**) The empirical expected frequencies for the case $\rho = 1$. (**b**) The empirical survival function of expected frequencies for the case $\rho = 1$

Based on the criteria we explained above, we fitted four models to the EI survival function, and we report the following R^2:

- Intensity Models:
  - Linear: $R^2 = 0.9643$, $f(x) = -3.92x + 0.96$
  - Exponential: $R^2 = 0.8254$, $f(x) = \exp(-12.10x + 0.44)$
  - Power: $R^2 = -1.0273$, $f(x) = x^{-0.90} + -3.13$
- Frequency Models:
  - Linear: $R^2 = 0.9499$, $f(x) = -1.96x + 0.97$
  - Exponential: $R^2 = 0.8229$, $f(x) = \exp(-6.15x + 0.52)$
  - Power: $R^2 = -0.8510$, $f(x) = x^{-1.03} + -2.66$

As one can see, based on the R^2, the linear models have the best fit i.e., $\theta_I \sim U(0, 2 \times EI)$ and $\theta_F \sim U(0, 2 \times EF)$ for some EI and EF.

4.3.5 *Modified Expected Intensity and Frequency and Their Modified Models*

Let us consider having two expected intensity values: one for the entire data-set including the data of all the countries, and another for single countries with significantly less data. The question arises: which value carries more credibility? On one hand, the region's expected loss appears more authentic as it relies solely on regional data. On the other hand, the limited number of data points for that region might lead to a less accurate estimation. Bühlmann suggests finding a compromise that better represents the expected loss, striking a balance between regional authenticity and the potential for estimation errors due to the smaller data-set.

However, Bühlmann's theory requires that the data-set for each mean estimation be equal, a condition not met in our case. Consequently, we need to adopt a theory to ensure that this data-set heterogeneous doesn't pose any issues. This adoption is detailed in the Appendix. If we denote the region by i and its expected intensity by EI_i and expected frequency by EF_i, we propose the following modifications to the region's expected intensity and frequency where the fitted models for $\theta_I \sim U(0, 2 \times EI)$ and $\theta_F \sim U(0, 2 \times EF)$ hold (see Appendix for more details):

$$MEI_i = \frac{n_i}{n_i + 4} EI_i + \frac{4}{n_i + 4} EI_T,$$

$$MEF_i = \frac{n_i}{n_i + 4} EF_i + \frac{4}{n_i + 4} EF_T.$$

Here, n_i is the number of the sample from region i and EI_T , EF_T are the expected intensity and frequency, respectively, of the whole data. The values for the modified premiums which is the production of the modified expected intensity and modified expected frequency for all countries are calculated and reported in Table 4.2.

As it is evident, the outcomes exhibit a considerable degree of smoothness; the smaller values are slightly elevated, while the larger ones are marginally reduced. This adjustment is influenced by both the volume of data and the aggregate means.To achieve this, we apply the same methodology used for fitting the original distributions. In Figs. 4.3, and 4.4 we plotted the modified expected intensity and frequency and their survival functions.

4.3.6 Model Uncertainty Premium

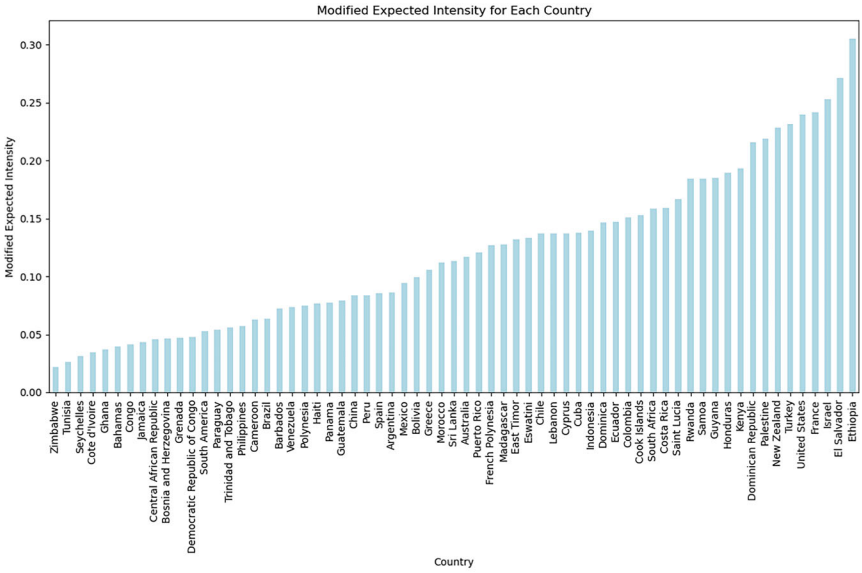
There is an issue when we want to look at the premium from the uncertainty perspective, that we really do not know which premium is the correct one. For that we look at the quantiles of the of the intensity and frequency. So let us consider the inverse CDF of the modified expected intensity and modified expected frequency modified $F^{-1}_{MEI}(\beta)$, $F^{-1}_{MEF}(\gamma)$, for $\beta, \gamma \in (0, 1)$. So if we believe that the correct values of the expected intensity and frequency are given by β, γ then the premium is nothing but the multiplication of the two given by:

$$F^{-1}_{MEI}(\beta) \times F^{-1}_{MEF}(\gamma).$$

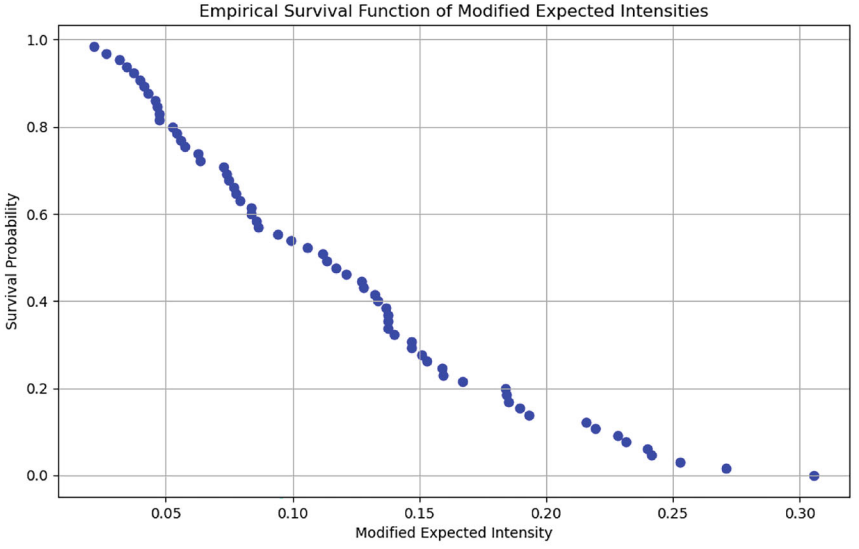
To address the uncertainty, we can consider a solution by looking at a scenario at which the prices is generated at a particular level of risk aversion. Let us consider we want to make sure that the premium that we propose with a high probability covers all the risk at α percent level, for instance 90% level. In other words, we want to have a premium that only can incorporate 10% possible under-pricing. Given this and that we are dealing with two inverse CDF the natural way to look at this is to consider the quantile space β , γ, therefore, we want to consider all the cases where so that $1 - (1 - \beta)(1 - \gamma) \leq \alpha$. As the two functions $F^{-1}_{MEI}(\beta)$, $F^{-1}_{MEF}(\gamma)$ are non-decreasing it make sense then to reduce the possibilites for β, γ to only $1 - (1 - \beta)(1 - \gamma) = \alpha$. Therefore, we introduce an uncertainty premium as follows:

$$Premium_{uncertaity}(\alpha) = \max_{1-(1-\beta)(1-\gamma)=\alpha} F^{-1}_{MEI}(\beta) \times F^{-1}_{MEF}(\gamma).$$

Now we fit a model to the modified expected intensity and frequency as it will be used in model uncertainty premium calculation.

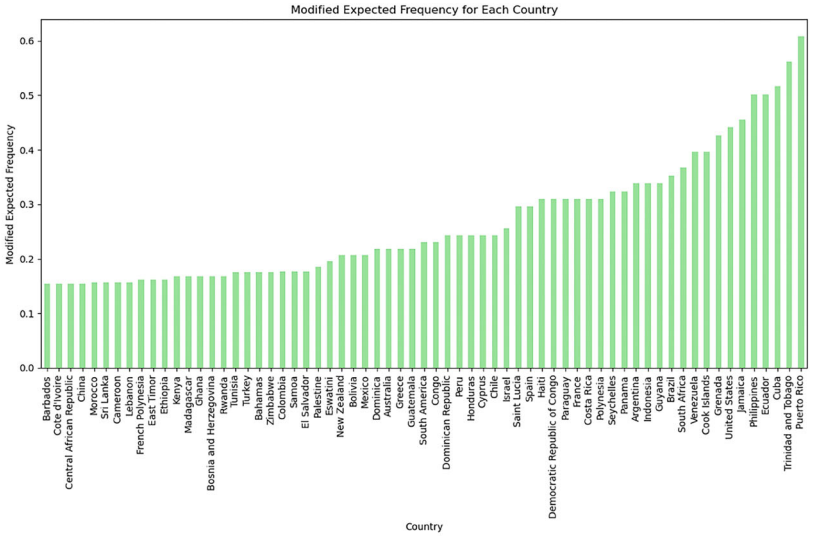


(a)

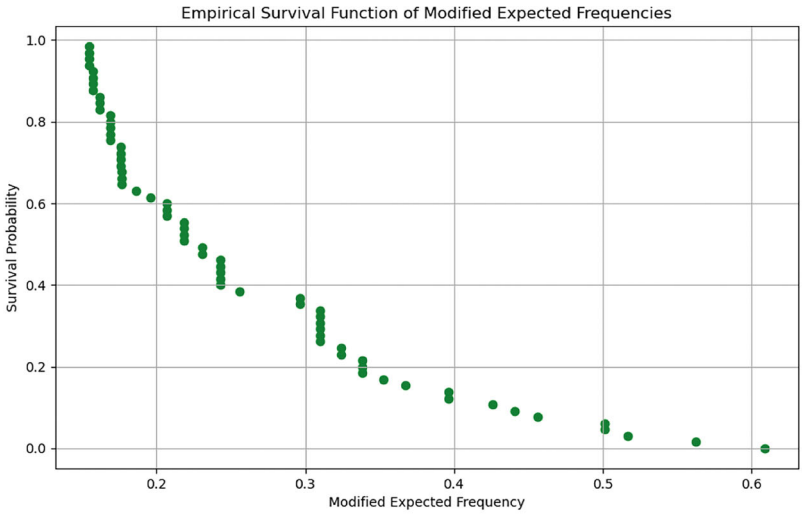


(b)

Fig. 4.3 The modified expected intensities and the associated survival fit for case $\rho = 1$. (**a**) The modified empirical expected intensities for the case $\rho = 1$. (**b**) The empirical survival function of the modified expected intensity for the case $\rho = 1$



(a)



(b)

Fig. 4.4 The modified expected frequencies and the associated survival fit for case $\rho = 1$. (**a**) The modified empirical expected frequencies for the case $\rho = 1$. (**b**) The empirical survival function of the modified expected frequencies for the case $\rho = 1$

- Modified Intensity Models:
  - Linear: $R^2 = 0.9595$, $f(x) = -4.31x + 1.00$
  - Exponential: $R^2 = 0.8667$, $f(x) = \exp(-13.39x + 0.60)$
  - Power: $R^2 = 0.0886$, $f(x) = x^{-1.24} - 3.85$
- Modified Frequency Models:
  - Linear: $R^2 = 0.8639$, $f(x) = -2.51x + 1.15$
  - Exponential: $R^2 = 0.9775$, $f(x) = \exp(-8.42x + 1.22)$
  - Power: $R^2 = 0.9184$, $f(x) = x^{-2.32} - 4.26$

The optimal model, according to our criteria, is the exponential model. We also need the inverse of the CDF

$$0 < \beta < 1, F^{-1}_{MEI}(\beta) = \frac{\beta}{4.31},$$

$$0 < \gamma < 1, F^{-1}_{MEF}(\gamma) = \frac{0.15 + \gamma}{2.51}.$$

So we introduce

$$Premium_{uncertaity}(\alpha) = \max_{1-(1-\beta)(1-\gamma)=\alpha} \frac{\beta}{4.31} \times \frac{0.15 + \gamma}{2.51}.$$

The following Fig. 4.5 shows the premium at 100α percent level, for $\alpha \in (0.95, 0.99)$.

4.3.7 Case $\rho = 70\%$

While in the previous analysis we considered to cover any loss below the last year production (i.e., $\rho = 1$), here we present the heat map of an example we have chosen where $\rho = 70\%$. Taking similar steps we find the empirical expected intensity and frequency and their survival functions plotted in Figs. 4.6 and 4.7

Consequently we fit the following models to the survival with their R^2 calculated

- Intensity Models:
  - Linear: $R^2 = 0.9140$, $f(x) = -1.53x + 0.57$
  - Exponential: $R^2 = 0.8867$, $f(x) = exp(-7.61x + -0.14)$
  - Power: $R^2 = -1.6790$, $f(x) = x^{-}0.77 + -3.03$

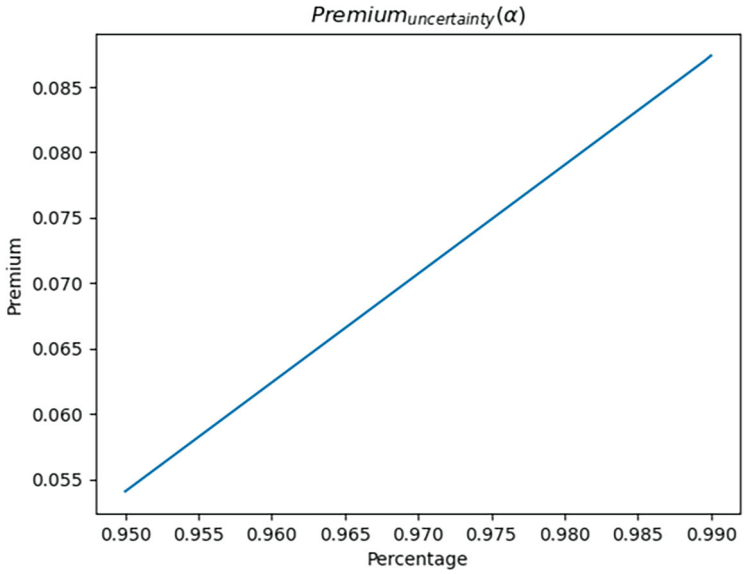


Fig. 4.5 Premium based on model uncertainty for all ranges of risk aversion based on the modified model for the case $\rho = 1$

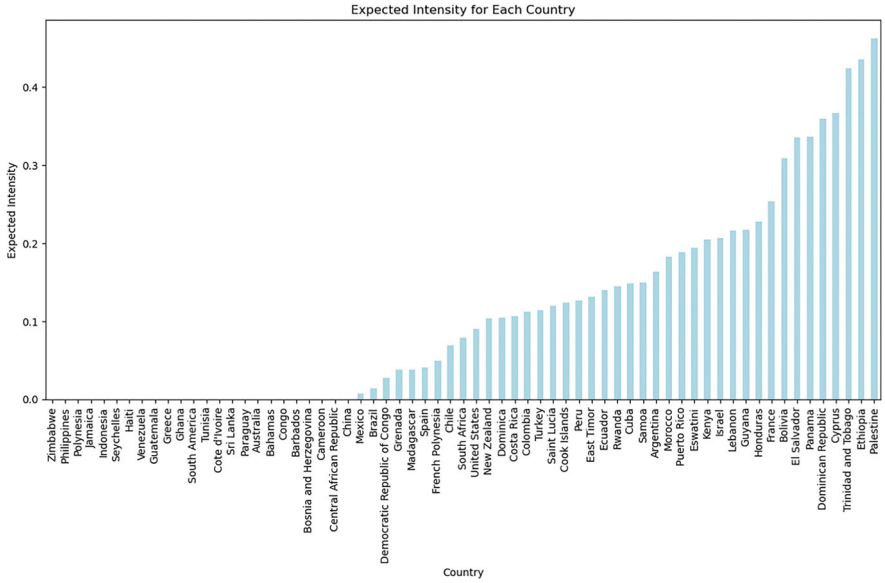
- Frequency Models:
  - Linear: $R^2 = 0.7674$, $f(x) = -5.94x + 0.54$
  - Exponential: $R^2 = 0.9157$, $f(x) = exp(-32.28x + -0.18)$
  - Power: $R^2 = 0.8005$, $f(x) = x^{-1.32} + -5.99$

As one can see unlike the case for $\rho = 1$, which was the original one, for the frequencies, the exponential model emerges as the best fit. Therefore, for modifying the data we need to consider a different approach. Using the exponential model for the frequency we will come up with the following

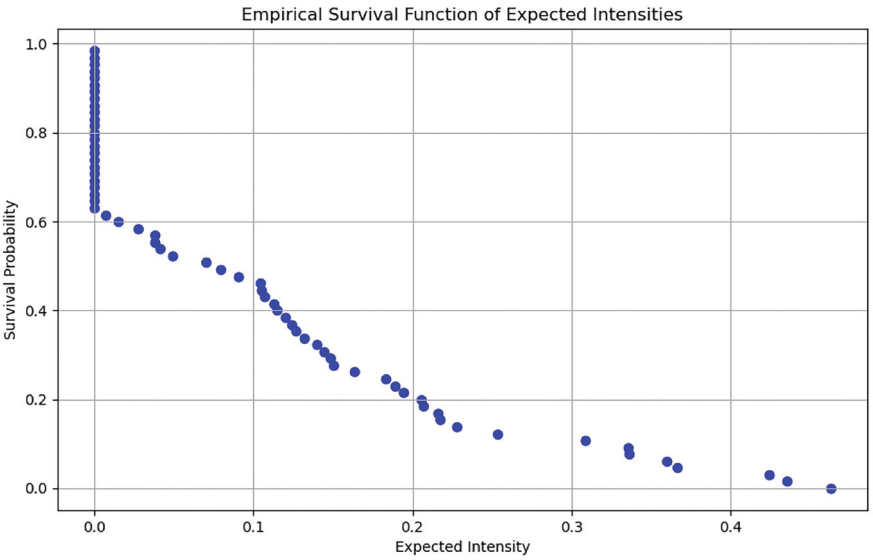
$$MEI_i = \frac{n_i}{n_i + 4} EI_i + \frac{4}{n_i + 4} EI_T,$$

$$MEF_i = \frac{n_i}{n_i + 2} EF_i + \frac{2}{n_i + 2} EF_T.$$

For more details see the Appendix. Figures 4.8 and 4.9 present the modified expected intensity and frequency, and their survivals.

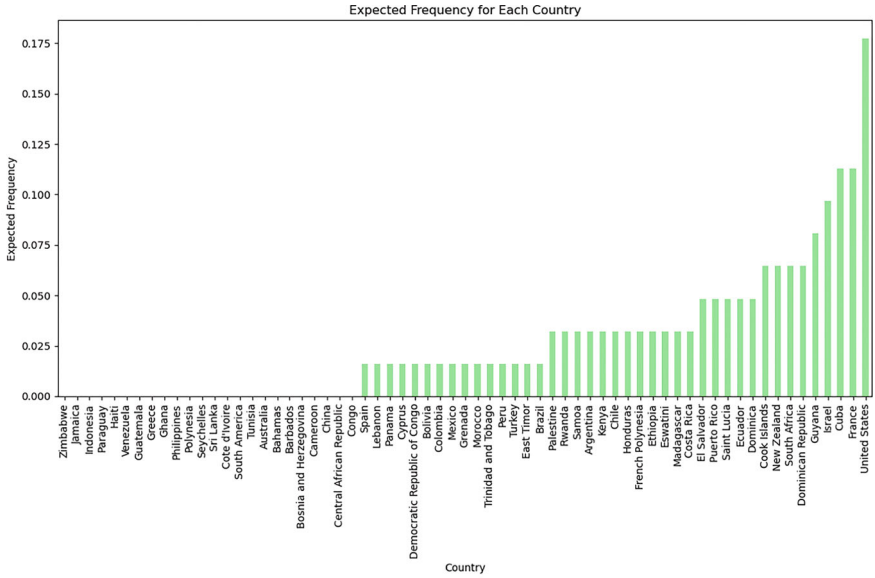


(a)

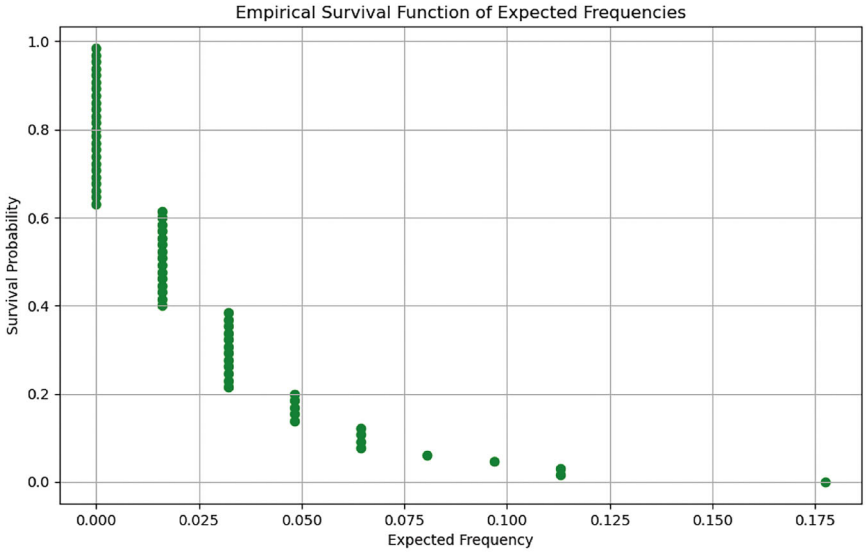


(b)

Fig. 4.6 The expected intensity and the associated survival fit for case $\rho = 0.7$. (**a**) The empirical expected intensity for the case $\rho = 0.7$. (**b**) The empirical survival function of the expected intensities for the case $\rho = 0.7$

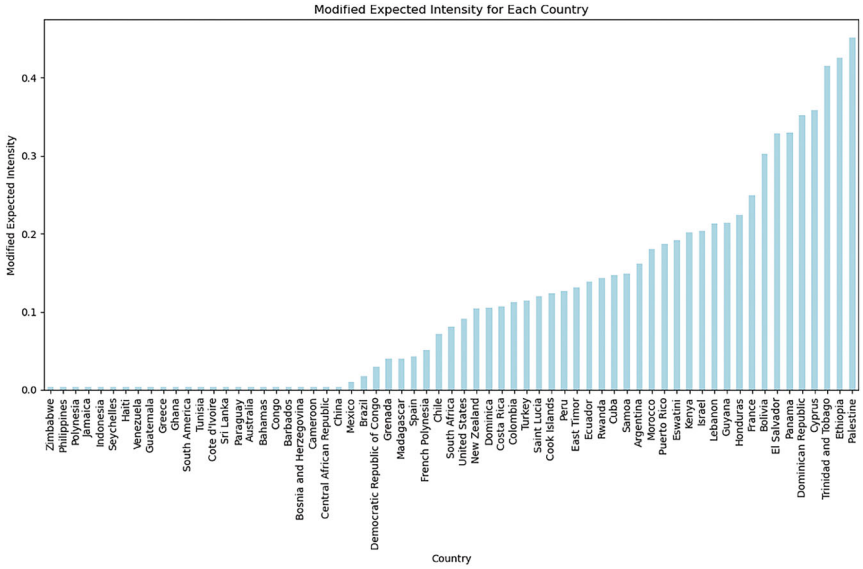


(a)

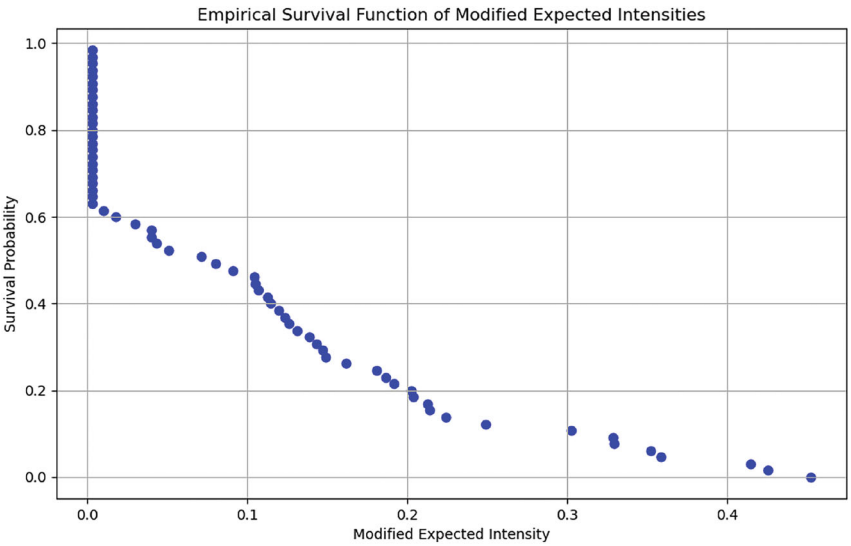


(b)

Fig. 4.7 The expected frequencies and the associated survival fit for case $\rho = 0.7$. (**a**) The empirical expected frequencies for the case $\rho = 0.7$. (**b**) The empirical survival function of the expected frequencies for the case$\rho = 0.7$

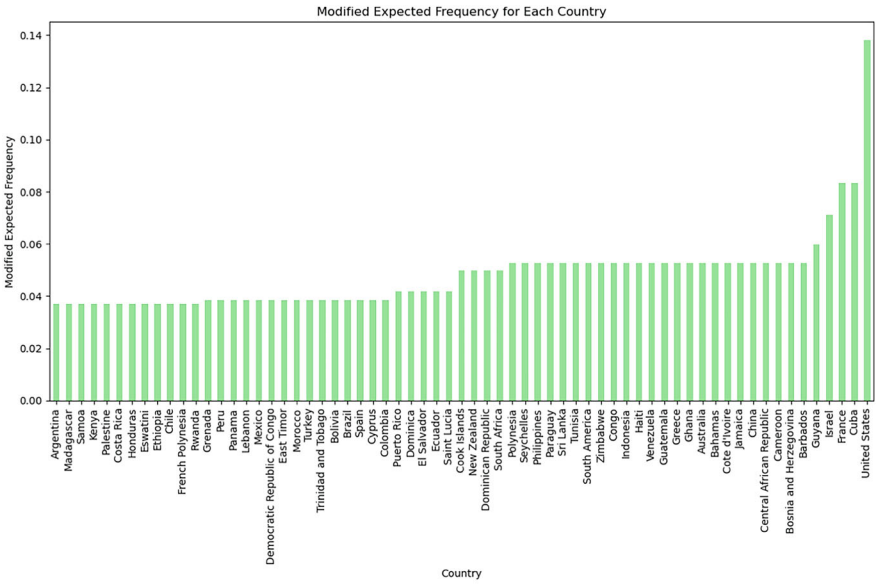


(a)

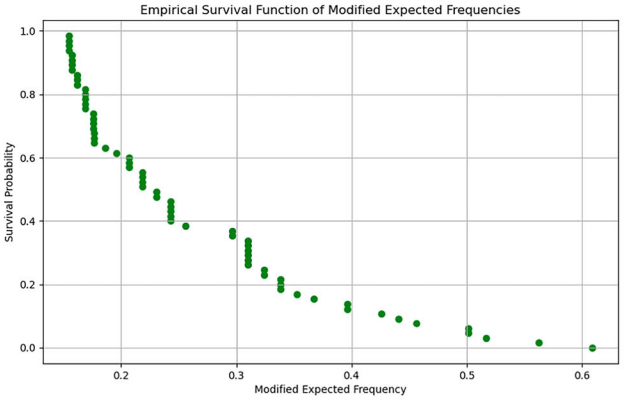


(b)

Fig. 4.8 The modified expected intensities and the associated survival fit for case $\rho = 0.7$. (**a**) The modified expected intensities for the case $\rho = 0.7$. (**b**) The empirical survival function of the modified expected intensities for $\rho = 0.7$



(a)



(b)

Fig. 4.9 The modified expected frequencies and the associated survival fit for case $\rho = 0.7$. (**a**) The modified expected frequencies for the case $\rho = 0.7$. (**b**) The survival empirical for the modified expected frequencies for case $\rho = 0.7$

Finally, we will modify the data based on this new rule and present the following modified returns and loss model.

- Modified Intensity Models:
  - Linear: $R^2 = 0.8393$, $f(x) = -2.25x + 0.74$
  - Exponential: $R^2 = 0.9261$, $f(x) = \exp(-7.62x + -0.17)$
  - Power: $R^2 = 0.7403$, $f(x) = x^{-0.40} + -2.32$

- Modified Frequency Models:
  - Linear: $R^2 = 0.7323$, $f(x) = -23.84x + 1.61$
  - Exponential: $R^2 = 0.8792$, $f(x) = \exp(-80.97x + 2.81)$
  - Power: $R^2 = 0.8617$, $f(x) = x^{-4.11} + -13.64$
- Inverse CDF of the modified expected intensity linear model: $F^{-1}_{MEI}(\alpha) = \frac{\alpha - 0.26}{2.25}$.
- Inverse CDF of the modified expected frequency linear model: $F^{-1}_{MEF}(\alpha) = \frac{2.81 - \ln(1-\alpha)}{80.97}$.

Similar to what we have done above we get the following premium that is concerned with the model uncertainty:

$$Premium_{uncertainty}(\alpha) = \max_{1-(1-\beta)(1-\gamma)=\alpha} \frac{\beta - 0.26}{2.25} \times \frac{2.81 - \ln(1-\gamma)}{80.97},$$

In Fig. 4.10 we have plotted $Premium_{uncertainty}(\alpha)$.

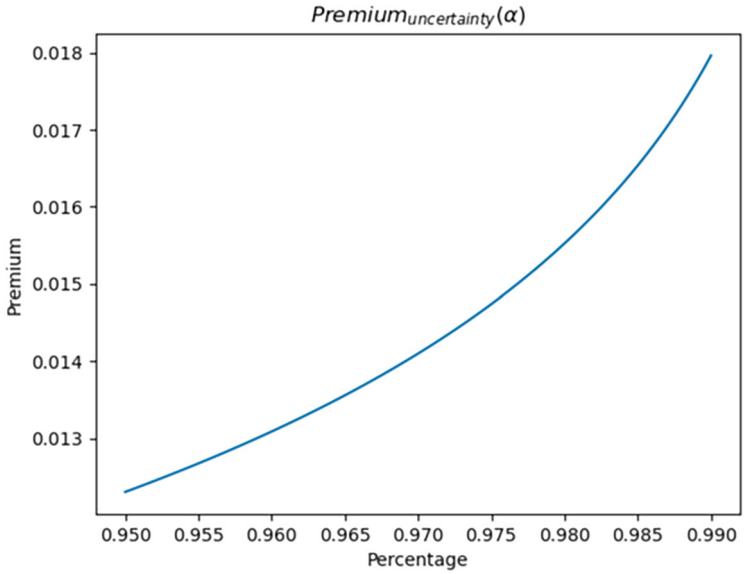


Fig. 4.10 Premium based on model uncertainty for all ranges of risk aversion based on the modified model for the case $\rho = 0.7$

4.4 Conclusion

This study has undertaken a risk analysis of avocado production across 65 countries, in the world. By introducing a modified credibility theory for insurance pricing on a heterogeneous data-set, our methodology addresses challenges arising from varying data-set characteristics, modifying Bühmann's credibility theory for increased reliability. The derived premium rate for coverage aligns with industry expectations and global trends, showing rather moderate rates based on historical trends. In addition to evaluating premium rates, our research proposes a two-layer insurance policy covering production return risks, incorporating both a standard deductible and a preventive measure. Our framework also discusses the uncertainty pricing of insurances as an option for the risk averse insurance companies.

Appendix

Mathematical Theorems

Proposition 4.1 *If we have m samples* $\{x_{i1}\}_{i=1,\ldots,n_1}, \ldots, \{x_{im}\}_{i=1,\ldots,n_m}$ *of m risks, and assume that the samples are identically distributed and independent given a parameter* ϑ *is known. Then, the best sample estimation of the conditional expectation*

$$\Pi_i = E\left(E\left[x_{ij} \mid \Theta_i = \vartheta\right] \mid x_{i1}, x_{i2} \ldots x_{in_i}\right) = E\left(m(\vartheta) \mid x_{i1}, x_{i2}, \ldots x_{i\,n_i}\right),$$

is

$$z_i \bar{x}_i + (1 - z_i)\,\mu,$$

where

$$x_i = \frac{x_{i1}+x_{i2}+\ldots+x_{in_i}}{n_i}, \quad z_i = \frac{1}{1+\frac{\sigma^2}{v^2}n_i},$$

and

$$\begin{aligned}\mu &= E\left(m_i(\vartheta)\right), m(\vartheta) = E\left[x_{ij}|\theta_i = \vartheta\right], s^2(\vartheta)\\ &= Var\left[x_{ij}|\theta_i = \vartheta\right], \sigma^2 = E\left[s^2(\vartheta)\right], v^2 = Var\left[m(\vartheta)\right].\end{aligned}$$

Proof Let us compute the quadratic error as follows:

$$
\begin{aligned}
&E\left[\left(a_{i0}+\sum_{j=1}^{n_i} a_{i,}X_{ij}-m(\vartheta)\right)^2\right]\\
&=E\left[\left(a_{i0}+\sum_{i=1}^{n_i} a_{ij}X_{i,}-\Pi_i\right)\right]+E\left[(m(\vartheta)-\Pi_i)^2\right]\\
&\quad+2E\left[\left(a_{i0}+\sum_{j=1}^{n_i} a_{ij}X_{ij}-\Pi_i\right)(m(\vartheta)-\Pi_i)\right]\\
&=E\left[\left(a_{i0}+\sum_{i=1}^{n_i} a_{ij}x_{ij}-\Pi_i\right)^2\right]+E\left[(m,(\vartheta)-\Pi_i)^2\right]
\end{aligned}
$$

The last equation follows from the fact that:

$$
\begin{aligned}
&E\left[\left(a_{i0}+\sum_{j=1}^{n_i} a_{ij}x_{ij}-\Pi_i\right)(m(\vartheta)-\Pi_i)\right]\\
&=E\left[E\left[\left(a_{i0}+\sum_{j=1}^{n_i} a_{ij}x_{ij}-\Pi_i\right)(m(\vartheta)-\Pi_i)\mid x_{i1},\ldots,x_{i\pi_i}\right]\right]\\
&=E\left[\left(a_{i0}+\sum_{j=1}^{n_i} a_{ij}x_{ij}-\Pi_i\right)E\left[(m(\vartheta)-\Pi_i)\mid x_{i1},\ldots,x_{in_i}\right]\right]=0.
\end{aligned}
$$

Let us find critical points of the function:

$$
\begin{aligned}
\tfrac{1}{2}\tfrac{\partial f}{\partial a_{i0}}=E\left[a_{i0}+\textstyle\sum_{j=1}^{n_i} a_{ij}x_{ij}-m(\vartheta)\right]&=a_{i0}+\textstyle\sum_{j=1}^{n_i} a_{ij}E\left(x_{ij}\right)-E(m(\vartheta))\\
&=a_{i0}-\left(\textstyle\sum_{j=1}^{n_i} a_{ij}-1\right)\mu.
\end{aligned}
$$

As a result, we get that $a_{i0}=\left(\sum_{j=1}^{n_i} a_{ij}-1\right)\mu$. For$k\neq 0$ we have:

$$
\frac{1}{2}\frac{\partial f}{\partial a_{ik}}=E\left[x_{ik}\left(a_{i0}+\sum_{j=1}^{n_i} a_{ij}x_{ij}-m(\vartheta)\right)\right]
$$

$$
=E\left[x_{ik}\right]a_{i0}+\sum_{i=1,k}^{n_1}\left(a_{ij}E\left[x_{ik}x_{ij}\right]\right)+a_{ik}E\left[x_{ik}^2\right]-E\left[x_{ik}m_i(\vartheta)\right]=0.
$$

We can simplify derivative, noting that:

$$\begin{aligned} E\left[x_{ij}x_{ik}\right] &= E\left[E\left[x_{ij}x_{ik} \mid \vartheta\right]\right] = E\left[cov\left(x_{ij}x_{ik} \mid \vartheta\right) + E\left(x_{ij} \mid \vartheta\right) E\left(x_{ik} \mid \vartheta\right)\right] \\ &= E\left[(m(\vartheta))^2\right] = v^2 + \mu^2, \\ E\left[x_{ik}^2\right] &= E\left[E\left[x_{ik}^2 \mid \vartheta\right]\right] = E\left[s^2(\vartheta) + (m(\vartheta))^2\right] = \sigma^2 + v^2 + \mu^2, \\ E\left[x_{ik}m(\vartheta)\right] &= E\left[E\left[x_{ik}m(\vartheta) \mid \vartheta\right]\right] = E\left[m(\vartheta)E\left[x_{ik} \mid \vartheta\right]\right] \\ &= E\left[(m(\vartheta))^2\right] = v^2 + \mu^2. \end{aligned}$$

Taking above equations and inserting into derivative, we have:

$$\begin{aligned} \tfrac{1}{2}\tfrac{\partial f}{\partial a_{ik}} &= \left(1 - \textstyle\sum_{j=1}^{n_i} a_{ij}\right)\mu^2 + \textstyle\sum_{j=1,kk}^{n_i} a_{ij}\left(v^2 + \mu^2\right) \\ &\quad + a_{ik}\left(\sigma^2 + v^2 + \mu^2\right) - \left(v^2 + \mu^2\right) \\ &= a_{ik}\sigma^2 - \left(1 - \textstyle\sum_{j=1}^{n_i} a_{ij}\right)v^2 = 0, \\ &\Rightarrow \sigma^2 a_{ik} = v^2\left(1 - \textstyle\sum_{i=1}^{n_i} a_{ij}\right). \end{aligned}$$

Right side doesn't depend on k. Therefore, all a_{ik} are constant:

$$a_{i1} = \ldots = a_{in_i} = \frac{v^2}{\sigma^2 + n_i v^2}.$$

From the solution for a_{i0} we have:

$$a_{i0} = (1 - n_i a_{ik})\,\mu = \left(1 - \frac{n_i v^2}{\sigma^2 + n_i v^2}\right)\mu.$$

Finally, the best estimator is

$$a_{i0} + \sum_{j=1}^{n_i} a_{ij}x_{ij} = \frac{n_i v^2}{\sigma^2 + n_i v^2}\bar{x}_i + \left(1 - \frac{n_i v^2}{\sigma^2 + n_i v^2}\right)\mu = z_i\bar{x}_i + (1 - z_i)\ \mu.$$

□

Proposition 4.2 *If* $\theta \sim \exp(\lambda)$ *and* $x_{ij} \sim \exp\left(\frac{1}{\theta}\right)$ *then we have* $z_i = \frac{n_i}{n_i+2}$.

Proof First, we have the following quantities:

$$\begin{aligned} m(\theta) &= E\left[x_{ij} \mid \theta\right] = \theta. \\ \mu &= E(m(\theta)) = E(\theta) = \tfrac{1}{\lambda} \\ s^2(\theta) &= Var\left[x_{ij} \mid \theta\right] = \theta^2. \\ \sigma^2 &= E\left(s^2(\theta)\right) = E\left(\theta^2\right) = 2\left(\tfrac{1}{\lambda}\right)^2. \\ v^2 &= Var(m(\theta)) = Var(\theta) = \left(\tfrac{1}{\lambda}\right)^2. \end{aligned}$$

From these we get:$z_i = \frac{1}{1+\frac{\sigma^2}{n_i v^2}} = \frac{1}{1+\frac{2\left(\frac{1}{\lambda}\right)^2}{n_i\left(\frac{1}{\lambda}\right)^2}} = \frac{n_i}{n_i+2}$. □

Proposition 4.3 *If* $\theta \sim U(0, b)$ *and* $x_{ij} \sim \exp\left(\frac{1}{\theta}\right)$ *then we have* $z_i = \frac{n_i}{n_i+4}$.

Proof First, we have the following quantities:

$$\begin{gathered} m(\theta) = E\left[x_{ij} \mid \theta\right] = \theta. \\ \mu = E(m(\theta)) = E(\theta) = \frac{b}{2} \\ s^2(\theta) = Var\left[x_{ij} \mid \theta\right] = \theta^2. \\ \sigma^2 = E\left(s^2(\theta)\right) = E\left(\theta^2\right) = \frac{b^2}{3}. \\ v^2 = Var(m(\theta)) = Var(\theta) = \frac{b^2}{12}. \end{gathered}$$

From these we get:$z_i = \frac{n_i}{n_i+4}$. □

Premiums, and Adjusted Premiums

Table 4.2 Premiums for the case $\rho = 0.7$ and $\rho = 1$

	$\rho = 1$		$\rho = 0.7$			$\rho = 1$		$\rho = 0.7$	
Country	Premium	Modified premium	Premium	Modified premium	Country	Premium	Modified premium	Premium	Modified premium
Argentina	0.53	0.45	0.53	0.45	Haiti	0.0	0.02	0.0	0.02
Australia	0.0	0.02	0.0	0.02	Honduras	0.73	0.62	0.73	0.62
Bahamas	0.0	0.02	0.0	0.02	Indonesia	0.0	0.02	0.0	0.02
Barbados	0.0	0.02	0.0	0.02	Israel	2.0	1.38	2.0	1.38
Bolivia	0.5	0.72	0.5	0.72	Jamaica	0.0	0.02	0.0	0.02
Bosnia and Herzegovina	0.0	0.02	0.0	0.02	Kenya	0.66	0.56	0.66	0.56
Brazil	0.02	0.05	0.02	0.05	Lebanon	0.35	0.51	0.35	0.51
Cameroon	0.0	0.02	0.0	0.02	Madagascar	0.12	0.12	0.12	0.12
Central African Republic	0.0	0.02	0.0	0.02	Mexico	0.01	0.03	0.01	0.03
Chile	0.23	0.2	0.23	0.2	Morocco	0.3	0.43	0.3	0.43
China	0.0	0.02	0.0	0.02	New Zealand	0.67	0.47	0.67	0.47
Colombia	0.18	0.27	0.18	0.27	Palestine	1.49	1.25	1.49	1.25
Congo	0.0	0.02	0.0	0.02	Panama	0.54	0.78	0.54	0.78
Cook Islands	0.8	0.56	0.8	0.56	Paraguay	0.0	0.02	0.0	0.02
Costa Rica	0.34	0.3	0.34	0.3	Peru	0.2	0.31	0.2	0.31
Cote d'Ivoire	0.0	0.02	0.0	0.02	Philippines	-0.0	0.02	0.0	0.02
Cuba	1.68	1.19	1.68	1.19	Polynesia	-0.0	0.02	0.0	0.02
Cyprus	0.59	0.85	0.59	0.85	Puerto Rico	0.92	0.66	0.92	0.66
Democratic Republic of Congo	0.04	0.08	0.04	0.08	Rwanda	0.47	0.4	0.47	0.4
Dominica	0.51	0.38	0.51	0.38	Saint Lucia	0.58	0.43	0.58	0.43
Dominican Republic	2.32	1.56	2.32	1.56	Samoa	0.48	0.42	0.48	0.42

East Timor	0.21	0.32	0.21	0.32	Seychelles	0.0	0.02	0.0	0.02
Ecuador	0.68	0.49	0.68	0.49	South Africa	0.51	0.37	0.51	0.37
El Salvador	1.62	1.15	1.62	1.15	South America	0.0	0.02	0.0	0.02
Eswatini	0.63	0.53	0.63	0.53	Spain	0.07	0.11	0.07	0.11
Ethiopia	1.41	1.18	1.41	1.18	Sri Lanka	0.0	0.02	0.0	0.02
France	2.86	1.99	2.86	1.99	Trinidad and Tobago	0.68	0.98	0.68	0.98
French Polynesia	0.16	0.15	0.16	0.15	Tunisia	0.0	0.02	0.0	0.02
Ghana	0.0	0.02	0.0	0.02	Turkey	0.18	0.28	0.18	0.28
Greece	0.0	0.02	-0.0	0.02	United States	1.61	1.26	1.61	1.26
Grenada	0.06	0.1	0.06	0.1	Venezuela	0.0	0.02	0.0	0.02
Guatemala	0.0	0.02	0.0	0.02	Zimbabwe	0.0	0.02	0.0	0.02
Guyana	1.75	1.19	1.75	1.19					

References

1. Ahsan, S. M., Ali, A. A. G., & Kurian, N. J. (1982).Toward a theory of agricultural insurance. *American Journal of Agricultural Economics, 64*(3), 520–529.
2. Mahul, O. (1999). Optimum area yield crop insurance. *American Journal of Agricultural Economics, 81*(1), 75–82.
3. Miranda, M. J. (1991). Area-yield crop insurance reconsidered. *American Journal of Agricultural Economics, 73*(2), 233–242.
4. Miranda, M. J., & Glauber, J. W. (1997). Systemic risk, reinsurance, and the failure of crop insurance markets. *American Journal of Agricultural Economics, 79*(1), 206–215.
5. Turvey, C. G., & Amanor-Boadu, V. (1989). Evaluating premiums for a farm income insurance policy. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie, 37*(2), 233–247.
6. Stokes, J. R., Nayda, W. I., & English, B. C. (1997). The pricing of revenue assurance. *American Journal of Agricultural Economics, 79*(2), 439–451.
7. Stokes, J. R. (2000). A derivative security approach to setting crop revenue coverage insurance premiums. *Journal of Agricultural and Resource Economics, 25*(01).
8. Turvey, C. G., & Stokes, J. R. (2008). Market structure and the value of agricultural contingent claims. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie, 56*(1), 79–94.
9. Tan, K. S., & Zhang, J. (2023). Flexible weather index insurance design with penalized splines. *North American Actuarial Journal, 0*(0), 1–26.
10. Assa, H., & Wang, M. (Simon) (2021). Price index insurances in the agriculture markets. *North American Actuarial Journal, 25*(2), 286–311.
11. Brock Porth, C., Porth, L., Zhu, W., Boyd, M., Tan, K. S., & Liu, K. (2020). Remote sensing applications for insurance: A predictive model for pasture yield in the presence of systemic weather. *North American Actuarial Journal, 24*(2), 333–354.
12. Chen, Z., Lu, Y., Zhang, J., & Zhu, W. (2024). Managing weather risk with a neural network-based index insurance. *Management Science, 70*(7), 4306–4327.
13. Li, H., Porth, L., Tan, K. S., & Zhu, W. (2021). Improved index insurance design and yield estimation using a dynamic factor forecasting approach. *Insurance: Mathematics and Economics, 96*, 208–221.
14. Bühlmann, H., & Gisler, A. (2005). *A course in credibility theory and its applications*. Springer.
15. Porth, L., Zhu, W., & Seng, T. K. (2014). A credibility-based Erlang mixture model for pricing crop reinsurance. *Agricultural Finance Review*, *74*(2), 162–187.
16. Pai, J., Boyd, M., & Porth, L. (2015). Insurance premium calculation using credibility analysis: An example from livestock mortality insurance. *The Journal of Risk and Insurance*, *82*(2), 341–357.
17. Wenjun, Z., Ken, S. T., & Porth, L. (2019). Agricultural insurance ratemaking: Development of a new premium principle. *North American Actuarial Journal*, *23*(4), 512–534.
18. Embrechts, P., & Puccetti, G. (2006). Bounds for functions of multivariate risks. *Journal of Multivariate Analysis, 97*(2), 526–547.

Chapter 5
How Do Economic Variables Affect the Pricing of Commodity Derivatives and Insurance?

Hirbod Assa, Philippe Grégoire, Gabriel J. Power, and Djerry Charli Tandja-M.

Abstract This paper focuses on designing and pricing commodity derivatives and insurance within a novel financial engineering framework that can be subsequently tested empirically using commodity price data. Optimal contract solutions are obtained and interpreted. We quantify explicitly how derivative prices and insurance premiums are affected by economic variables linked to commodity supply and demand. Our results generalize some existing commodity derivative pricing models and further show under which conditions there will be no trading of derivative instruments and insurance. We report GMM estimates of the model parameters for a large dataset of commodity futures. These results also contribute to a better understanding of the "financialization" of commodities.

5.1 Introduction

The pricing of insurance contracts and financial derivatives on commodities such as crude oil or wheat are methodologically linked by their use of risk-neutral valuation methods such as the seminal Black-Scholes formula. Unlike financial security prices, which are driven by priced equity risk factors, commodity prices are mainly influenced by commodity-specific economic variables, which depend on fundamental weather, production, and storage variables [1, 2]. Commodity prices thus result from demand, supply, inventory, and economic risk factors. Although a large literature exists that takes a reduced-form approach to model the prices of commodity derivatives and insurance, a deeper understanding of their

H. Assa
Model Library, London, UK

P. Grégoire · G. J. Power (✉)
Université Laval, Quebec City, QC, Canada
e-mail: gabriel.power@fsa.ulaval.ca

D. C. Tandja-M.
Université du Québec en Outaouais, Gatineau, QC, Canada

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_5

pricing requires a richer model of these economic fundamentals. Indeed, financial engineering methods are often silent when it comes to quantifying the exact role of economic variables on these contingent claim prices. This paper aims to make a contribution to the literatures on commodity derivative and insurance contract pricing by proposing an economically motivated model, which allows us to generate new insights on pricing and risk management.

Generally speaking, there are two approaches to modelling commodity prices prior to modelling their contingent claims. The first approach belongs to the literature on finance and financial engineering, which uses models based on diffusion processes, and begins with [3]. The multifactor models of [4, 5], the CEV model of [6], and the mean reverting models of [7] are just a few examples. The second approach belongs to the economics literature and is based on rational commodity storage. In particular, this paper aims to adapt to derivative and insurance pricing the framework found in the Deaton and Laroque models [8–10]. By making explicit the price elasticity parameter, this paper also relates to the CEV option pricing model in [11].

This paper therefore aims to combine the two approaches mentioned above by taking an established methodology from the rational expectations theory of storage, and then modifying and applying it to a financial engineering framework. As a result, it develops a new methodology that directly quantifies the impact of economic variables on commodity insurance and derivative prices. To achieve this goal requires tackling several mathematical and technical problems, which we solve in this paper. The findings described in this paper are also of interest for researchers working on the "financialization" of commodities [12–14]. Indeed there is great interest in understanding the closer link between commodity and financial markets facilitated by exchange-traded derivatives. The methods and results in this paper should be useful to academics, traders, and practitioners in finance and financial engineering, as well as those in insurance, reinsurance, and risk management.

The remainder of the paper is as follows. In Sect. 5.2, we develop a representative agent model. Section 5.3 presents the dynamics for demand and price, defines the loss function and solves for the premium. In Sect. 5.4, optimal contracts are solved in the general case, and an application is provided for the special case where Value-at-risk is used as the risk measure. Section 5.5 presents empirical estimates of the parameters of the model, using a large dataset of commodity futures contract prices and GMM estimation. We conclude in Sect. 5.6.

5.2 Representative Agent Model

Let us consider a representative agent with an isoelastic or power utility function given by:

$$u(x) = \frac{x^{1-\phi} - 1}{1 - \phi}, \text{ if } \phi \neq 1 \text{ and } u(x) = \log(x) \text{ otherwise, with } \phi > 0.$$

In this model the parameter ϕ represents the coefficient of relative risk aversion (RRA) for the representative consumer/investor.[1] In this paper, we focus our attention on a single good, and treat spending on the other good as a residual that can be added to the agent's utility by using a quasi-linear utility function. Therefore, we consider the agent will be solving the following problem:

$$\max_{x,m} \left[ku(x) + m\right], \quad \text{s.t.} \quad px + m = B.$$

where m is residual income for all other goods, k is a constant, p is the price of good x and B is the budget constraint. Since the budget constraint is given by $B = px + m$, we can solve for m in terms of p using that equation. This yields $m = B - px$, which we substitute into our objective function to solve the following problem:

$$\max_{x} \left[ku(x) + B - px\right].$$

Since the first difference in terms of x is

$$ku'\left(x^*\right) - p = k\left(x^*\right)^{-\phi} - p = 0 \Rightarrow x^* = \left(\frac{k}{p}\right)^{1/\phi}.$$

If $m = 0$, then

$$x^* = \frac{B}{p}$$

Therefore,

$$x^* = \min\left\{\frac{B}{p}, \left(\frac{k}{p}\right)^{1/\phi}\right\}.$$

The inverse demand function is then $p(x) = kx^{-\phi}$. To simplify the notation, denote $\gamma = -\phi$, then we can consider the following inverse demand function:

$$p(x) = kx^{\gamma}$$

The form of the demand function for different values of k and γ are depicted in Fig. 5.1:

[1] Other functional forms are considered in the appendix.

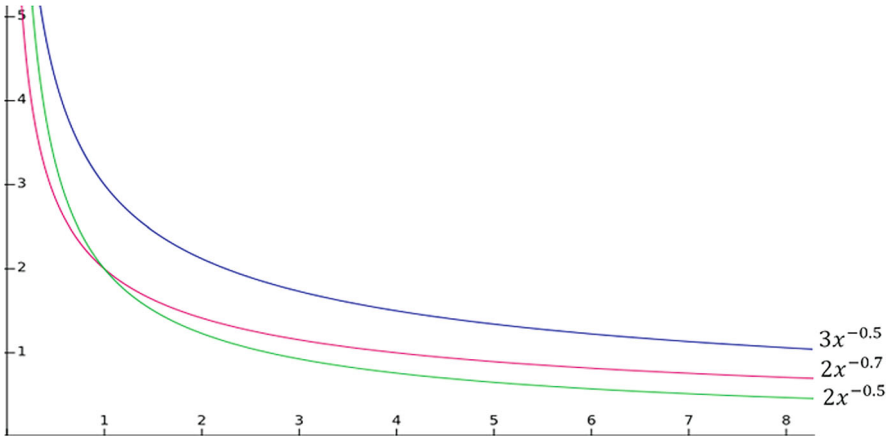


Fig. 5.1 Demand functions for different values of k and γ

5.3 Economic Model, Loss Distribution, and Premium

5.3.1 The Demand and Price Process

Let us consider a stochastic demand process following geometric Brownian motion (gBm) dynamics as follows:

$$\frac{dx_t}{x_t} = \mu dt + \sigma dw_t,$$

where in this case $(w_t)_{0 \le t \le T}$ is a standard Brownian motion, μ is the drift term representing the rate of growth in consumption, and σ represents the magnitude of the demand volatility. As a result, the dynamics of the demand process given above can be written as follows:

$$x_t = x_0 e^{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma w_t}, \quad t \ge 0$$

As the demand functions are allowed to vary, we can study different markets according to their elasticity of demand (e.g., different agricultural commodities). Therefore, we can study the effect of economic and financial market variables on the market demand, and on the resulting derivatives and insurance contracts. Considering the isoelastic demand function, combining the inverse demand function $p(x)$ with the demand process dynamics yields the following price dynamics for an inverse demand function p:

$$p_t = p\left(x_t\right) = k x_t^{\gamma} = p_0 \exp\left(\gamma\left(\mu - \frac{1}{2}\sigma^2\right)t + \gamma\sigma w_t\right)$$

If we consider a new Brownian motion $B_t = -w_t$, one can rewrite the price dynamics as follows:

$$p_t = p_0 \exp\left(\gamma\left(\mu - \frac{1}{2}\sigma^2\right)t - \gamma\sigma B_t\right)$$

This change is necessary because $\gamma = -\phi < 0$, and as a result $-\gamma\sigma > 0$. Using the Ito formula for the function $f(x,t) = ke^{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)t-\gamma\sigma x}$ gives:

$$\frac{\partial f}{\partial t} = \gamma\left(\mu - \frac{1}{2}\sigma^2\right)f, \quad \frac{\partial f}{\partial x} = (-\gamma\sigma)\,f, \quad \frac{\partial^2 f}{\partial x^2} = \gamma^2\sigma^2 f,$$

resulting in:

$$dp_t = \gamma\left(\mu + \frac{1}{2}(\gamma - 1)\sigma^2\right)p_t dt - \gamma\sigma p_t dB_t.$$

For simplicity, we can write the stochastic differential equation (SDE) of the price process as follows:

$$\frac{dp_t}{p_t} = \nu dt + \eta dB_t,\ p_0 > 0$$

where $\nu = \gamma\left(\mu + \frac{1}{2}(\gamma - 1)\sigma^2\right)$ and $\eta = -\gamma\sigma$.

It is worth considering some conditions under which the model makes greater economic sense. The first condition is that the drift term of the price, i.e., ν, must be non-negative. Since $\gamma \leq 0$ then this is equivalent to checking that:

$$\mu + \frac{1}{2}(\gamma - 1)\sigma^2 \leq 0.$$

However, on the other hand, the market price of risk needs to be non-negative to make sure that market participants will be involved. For that reason, it is necessary to check whether $\nu - r > 0$. This condition will certainly yield the previous one. The two conditions are economically sensible, but in general they are not necessary to obtain solutions.

5.3.2 Loss Distribution

Let us consider a time horizon T at which we want to introduce a loss variable and write an insurance contract to hedge against the risk of loss. We consider the following non-negative random variable as the loss

$$L = \left(p_0 - e^{-rT} p_T\right)_+.$$

To motivate this definition, note that if $p_0 - e^{-rT} p_T < 0$ then the excess return i.e., $\log\left(\frac{p_T}{p_0}\right) - rT$, is also negative. Next, we wish to find out the distribution of the loss variable L. First, it is not difficult to see that:

for $x < 0$, we have ,	$F_L(x) = 0$,
for $x = 0$ we have	$F_L(0) = P\left(p_0 \le e^{-rT} p_T\right) = N\left(\frac{\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - rT}{-\gamma\sigma\sqrt{T}}\right)$
and for $x > p_0$, we have	$F_L(x) = 1$

Now let us consider $p_0 \ge x > 0$. In this case, we have:

$$\begin{aligned}
F_L(x) = 1 - P\left(L > x\right) &= 1 - P\left(\left(p_0 - e^{-rT} p_T\right)_+ > x\right) \\
&= 1 - P\left(p_0 - e^{-rT} p_T > x\right) \\
&= 1 - P\left(p_0\left(1 - e^{-rT} e^{\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - \gamma\sigma B_T}\right) > x\right) \\
&= 1 - P\left(e^{rT}\left(1 - \frac{x}{p_0}\right) > e^{\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - \gamma\sigma B_T}\right) \\
&= 1 - P\left(\frac{\log e^{rT}\left(1 - \frac{x}{p_0}\right) - \gamma\left(\mu - \frac{1}{2}\sigma^2\right)T}{-\gamma\sigma\sqrt{T}} > B_1\right) \\
&= 1 - N\left(\frac{rT + \log\left(1 - \frac{x}{p_0}\right) - \gamma\left(\mu - \frac{1}{2}\sigma^2\right)T}{-\gamma\sigma\sqrt{T}}\right) \\
&= N\left(\frac{\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - rT - \log\left(1 - \frac{x}{p_0}\right)}{-\gamma\sigma\sqrt{T}}\right).
\end{aligned}$$

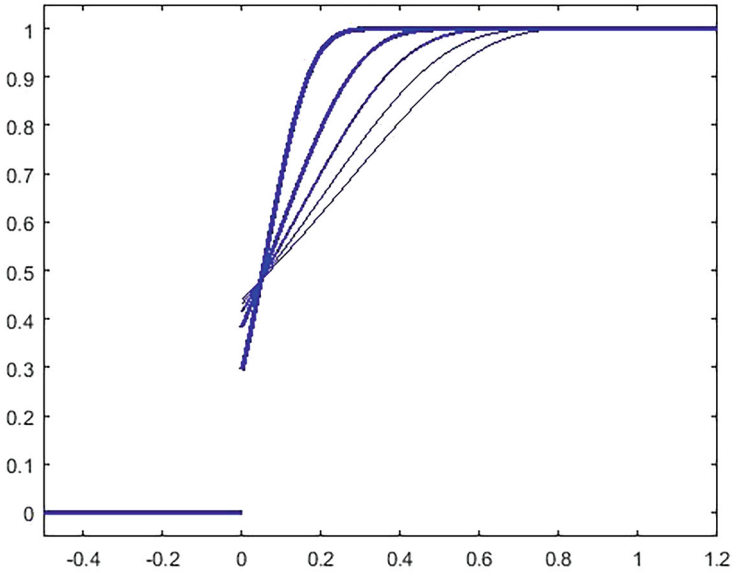


Fig. 5.2 Optimal solution in terms of $F_L(x)$ for different values of γ

In sum, we get:

$$F_L(x) = \begin{cases} 0, & x < 0 \\ N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT-\log\left(1-\frac{x}{p_0}\right)}{-\gamma\sigma\sqrt{T}}\right), & 0 \le x < p_0. \\ 1, & x \ge p_0 \end{cases}$$

The graph of $F_L(x)$ is depicted in Fig. 5.2.

5.3.3 *Premium*

Moreover, we can use risk-neutral valuation principles to price any contract $H = h(p_T)$ by using the risk-free probability measure Q as follows:

$$Price = e^{-rT} E^Q \left(h\left(p_T\right)\right) = e^{-rT} E\left(\frac{dQ}{dP} h\left(p_T\right)\right),$$

where

$$\frac{dQ}{dP} = e^{\left(\frac{m^2}{2\eta^2} - \frac{m}{2}\right)T} \left(\frac{e^{-rT} p_T}{p_0}\right)^{-\frac{m}{\eta^2}},$$

and $m = \nu - r$. From this, we can show that:

$$\pi(L) = E\left(\frac{dQ}{dP} h\left(P_T\right)\right) = \int_0^1 \mathrm{VaR}_t \ \left(h\left(p_T\right)\right) d\Gamma(t), \tag{5.1}$$

where $\Gamma(t) = N\left(N^{-1}(t) - \frac{|m|\sqrt{T}}{\eta}\right)$.

This representation (5.1) will subsequently help us, along with the risk measures that we use, to find the optimal solutions in a useful way. The above formula can also be used to price options. To do so, first, we define the contract H based on its payoff profile. Then, we compute the Radon-Nikodym derivative using the estimated underlying price parameters. This is equivalently a risk-neutralization of the process. Finally, we use either analytical formulae or Monte Carlo simulation to compute the price.[2]

5.4 Designing Optimal Insurance and Pricing Derivatives

5.4.1 Ill-Posed Hedging Issue

In this section, we consider how to use the above model for purposes of pricing optimal insurance contracts as well as options. In this paper, we consider the contracts in the form of $X = k(L)$, where k is called the indemnity function and $i(x) = x - k(x)$ is called the retained loss function. Inspired by [15] and Cong et al. [16], to avoid ill-posed hedging, we impose some conditions on the insurance contracts. First, we assume zero loss needs no indemnity and no retained loss, i.e., $k(0) = i(0) = 0$. Second, we assume that the indemnity is compatible with the loss increase; meaning that, larger losses need larger indemnity. This assumption implies that k is a non-decreasing function. Third, we assume that the insurance company will not over-hedge the losses by assuming that i is non-decreasing which can be justified since larger risk cannot imply smaller retained losses.

[2] This approach is however not directly applicable to path-dependent options such as American options. For these we would have to consider alternative methods, such as Least-Squares Monte Carlo, which is beyond the scope of this paper.

Summarizing all the assumptions, we can list them as follows:

1. Zero risk assumption: $k(0) = i(0) = 0$;
2. Risk compatibility: $x_1 \leq x_2 \Rightarrow k(x_1) \leq k(x_2)$;
3. No over-hedging: $x_1 \leq x_2 \Rightarrow i(x_1) \leq i(x_2)$.

The conditions above can be summarized in the following assumption: *We consider contracts $X = k(L)$, where k belongs to the following set:*

$$C = \left\{ k : R_+ \to R_+ | \begin{array}{c} k(x) \text{ and } x - k(x) \\ \text{are nonnegative, non} - \text{decreasing} \end{array} \right\}.$$

The following lemma can be found in [75].

Lemma 1 *For any $k \in C$, the derivative of k and i exists a.s., and we have $0 \leq k^{'}$, $i^{'} \leq 1$ a.s.*

5.4.2 Optimal Solution

Next, we set up an optimal insurance problem and try to find an optimal solution. For that, we assume that the insuree is a risk-averse agent whose risk is measured according to a distortion risk measure ρ on the set of non-negative random variables defined as follows:

$$\rho(X) = \int_0^1 \text{VaR}_t(X) d\Pi(t).$$

Here $\Pi : [0, 1] \to [0, 1]$ is a non-decreasing function so that $\Pi(0) = 0$ and $\Pi(1) = 1$. This family of risk measures includes very important examples, e.g., Value-at-Risk with:

$$\Pi(t) = 1_{[\alpha,1]}$$

or Conditional Value at Risk with:

$$\Pi(t) = \frac{t - \alpha}{1 - \alpha} 1_{[\alpha,1]}.$$

A distortion risk measure is a way to better capture the risk by distorting the loss distribution. For instance, some risk measures (e.g., CVaR), distort the distribution by taking a pessimistic point of view towards the risk. Thus, note that the pricing method we propose earlier for the contracts in (5.1) is also a distortion risk measure, where the prices are distorted according to the pricing kernel distribution and the level of distortion depends on market price of risk. The insuree's global loss is the

part of the loss that is not covered by insurer, added up to the amount that is paid for the premium, i.e.,

$$Global\ loss = L - X + \pi(X).$$

Since distortion risk measures are cash-invariant, the risk of the global loss is $\rho(L - X) + \pi(X)$. In order to study insurance premiums, we consider an optimal insurance design problem as proposed in [17, 18] (or similarly with a budget constraint in [19]):

$$\min_{k \in C} \rho\left(L - k(L)\right) + \delta\pi\left(k(L)\right),$$

for a risk loading factor $\delta \geq 1$ that is used by the insurance company. Using the marginal indemnification function method (MIF) introduced by [17] and developed in [18–20], this problem can be rewritten as follows:

$$\min_{0 \leq k' \leq 1} \int_0^1 \left(\delta\left(1 - \Gamma\left(F_L(t)\right)\right) - \left(1 - \Pi\left(F_L(t)\right)\right)\right) k'(t) dt,$$

where $k^{'}$ is the derivative of k. The optimal solution is then given by $X = k(L)$, where:

$$k'(t) = \begin{cases} 1, & 1 - \Pi\left(F_L(t)\right) > \delta\left(1 - \Gamma\left(F_L(t)\right)\right) \\ 0, & 1 - \Pi\left(F_L(t)\right) \leq \delta\left(1 - \Gamma\left(F_L(t)\right)\right) \end{cases}. \tag{5.2}$$

5.4.3 Solving for the Contracts

In this section, we consider a technical assumption, namely that there are values a, $b \in (0, 1)$ such that:

$$1 - \Pi(x) > \delta\left(1 - \Gamma(x)\right) \text{ on } (a, b.)$$

and everywhere else than the interval (a, b):

$$1 - \Pi(x) < \delta\left(1 - \Gamma(x)\right) \text{ on } (0, a) \cup (b, 1).$$

This assumption holds for many interesting cases including $\rho = \text{VaR}$ and CVaR (see Fig. 5.3).

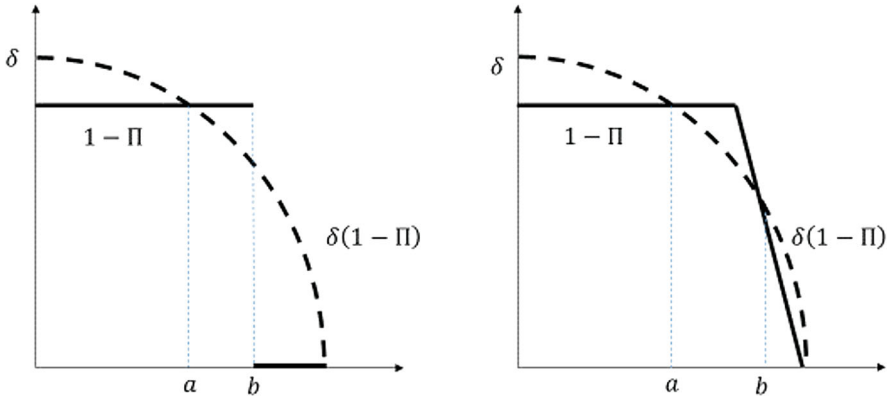


Fig. 5.3 Representing the optimal solution: cases of VaR and CVaR

The existence of the optimal solution (and its form) depends on $F_L(0)$. However, we know that:

$$F_L(0) = N\left(\frac{\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - rT}{-\gamma\sigma\sqrt{T}}\right) = N\left(\frac{\left(\mu - \frac{1}{2}\sigma^2\right)\sqrt{T}}{-\sigma} - \frac{r\sqrt{T}}{|\gamma|\,\sigma}\right).$$

Therefore, increasing the absolute value of γ will decrease the value of $F_L(0)$. The optimal solution in this case either: (i) does not exist, (ii) is a stop loss policy, or (iii) is a two-layer policy. This result can be shown as in Fig. 5.3 by depicting F_L for different values of γ.

Figure 5.3 Optimal solution in terms of $F_L(x)$ for different values of γ

We have three cases:

1. If $F_L(0) > b$, or $\frac{r\sqrt{T}}{\sigma\left(N^{-1}(b) - \frac{\left(\mu - \frac{1}{2}\sigma^2\right)\sqrt{T}}{-\sigma}\right)} < \gamma$, then $k' = 0$ and there is no contract.
2. If $a < F_L(0) < b$, or $\frac{r\sqrt{T}}{\sigma\left(N^{-1}(b) - \frac{\left(\mu - \frac{1}{2}\sigma^2\right)\sqrt{T}}{-\sigma}\right)} < \gamma < \frac{r\sqrt{T}}{\sigma\left(N^{-1}(a) - \frac{\left(\mu - \frac{1}{2}\sigma^2\right)\sqrt{T}}{-\sigma}\right)}$, then there is a stop loss policy with retention level that solves

$$N\left(\frac{\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - rT - \log\left(1 - \frac{b^*}{p_0}\right)}{-\gamma\sigma\sqrt{T}}\right) = b.$$

This results in

$$b^* = p_0\left(1 - \exp\left(\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - rT + \gamma\sigma\sqrt{T}N^{-1}(b)\right)\right).$$

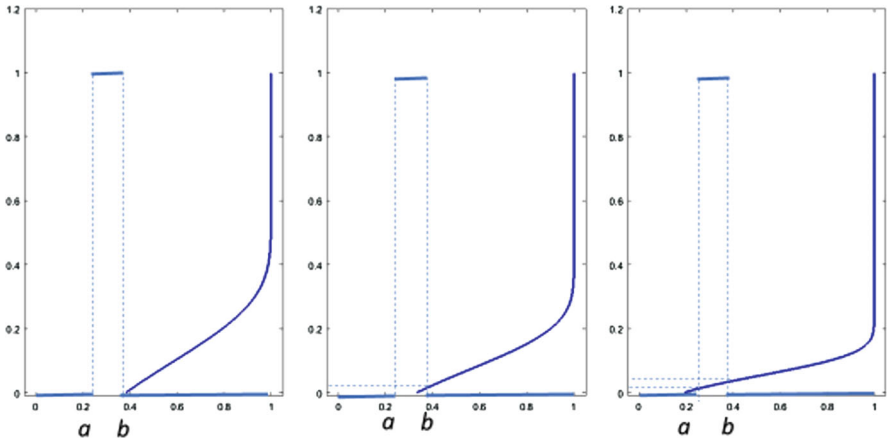


Fig. 5.4 Cases 1 through 3, in terms of values of a and b

3. Finally, if, $F_L(0) < a$ or $\gamma < \frac{r\sqrt{T}}{\sigma\left(N^{-1}(a) - \frac{\left(\mu - \frac{1}{2}\sigma^2\right)\sqrt{T}}{-\sigma}\right)}$, then the contract is a two-layer policy with upper retention level b^* given above and lower retention level a^* given as

$$a^* = p_0\left(1 - \exp\left(\gamma\left(\mu - \frac{1}{2}\sigma^2\right)T - rT + \gamma\sigma\sqrt{T}N^{-1}(a)\right)\right).$$

These three cases are represented in Fig. 5.4**.**

Remark 1 Some explanation is warranted here. In case 1, one can see that the probability of no loss, i.e., $F_L(0)$, happens to be large enough ($F_L(0) > b$), and for that reason the insuree does not look at the insurance as a necessary risk management tool, whereas in case 2 there is a demand for insurance from the insuree side. In the most extreme case 3, since the probability of no loss is small ($F_L(0) < a$), then the insurance company needs to include a stop-loss policy to contract to limit its exposure to large losses.

Taking the integral from the marginal indemnity function if (5.2), we get the indemnity as follows:

$$k(t) = \begin{cases} 0, & t \le a^* \\ t - a^*, & a^* \le t \le b^* \\ b - a^*, & t \ge b^* \end{cases}. \tag{5.3}$$

from which we can get the following contract:

$$X = \begin{cases} 0, & L \le a^* \\ L - a^*, & a^* \le L \le b^* \\ b - a^*, & L \ge b^* \end{cases}. \tag{5.4}$$

Now using the particular loss variable we have introduced in this paper (Sect. 5.3.2), we find that:

$$X = \begin{cases} 0, & p_0 - e^{-rT} p_T < 0 \text{ or} \\ & \left(p_0 - e^{-rT} p_T \ge 0 \text{ and } p_0 - e^{-rT} p_T \le a^*\right) \\ p_0 - e^{-rT} p_T - a^*, & a^* \le p_0 - e^{-rT} p_T \le b^* \\ b^* - a^*, & p_0 - e^{-rT} p_T \ge b^* \end{cases}. \tag{5.5}$$

Furthermore, one can easily see that this results in:

$$X = e^{-rT}\left(e^{rT}\left(p_0 - a^*\right) - p_T\right)_+ - e^{-rT}\left(e^{rT}\left(p_0 - b^*\right) - p_T\right)_+ \tag{5.6}$$

Finally, using the pricing kernel of the Black-Scholes model (Eq. 5.1) we get:

$$Price = P\left(p_0, e^{rT}\left(p_0 - a^*\right), r, T, -\gamma\sigma\right) - P\left(p_0, e^{rT}\left(p_0 - b^*\right), r, T, -\gamma\sigma\right) \tag{5.7}$$

where $P(p_0, K, r, T, \sigma)$ is the price of a put option with risk-free r, volatility σ, expiration T and strike price K.

Remark 2 As one can see, the price of the optimal product is a function of multiple parameters, including γ, the demand elasticity parameter, and σ, the demand volatility. It is clear that the prices of both put options in (5.7) increase if the absolute value of γ and σ increases. However, the retention levels, a^*, b^* are also functions of these two parameters, which makes it ultimately unclear how the increase in γ and σ will affect the optimal price. We will discuss it within an example when we use Value-at-Risk as the risk measure.

5.4.4 Value at Risk (VaR)

In the following two sections we focus our attention to a particular risk measure VaR. However, everything that we explore here is true for CVaR as well, since based on Fig. 5.2, the only difference between VaR and CVaR contracts is that the second layer for CVaR is greater than the second layer for VaR i.e., $b_{CVaR} > b_{VaR}$. This means considering CVaR does not essentially provide new information about the

behavior of the contracts except for VaR we also can find all explicit solutions for the the layers and prices.

Based on general case that we discussed above, let a be the solution to $\delta(1-\Gamma(a))=1$ or $a=\Gamma^{-1}\left(1-\frac{1}{\delta}\right)$. This means:

$$\delta\left(1-N\left(N^{-1}(a)-\frac{|m|\sqrt{T}}{\eta}\right)\right)=1$$

or

$$a=N\left(\frac{|m|\sqrt{T}}{\eta}+N^{-1}\left(1-\frac{1}{\delta}\right)\right).$$

It is also clear that in this case $b=\alpha$.

There are three cases:

1. If $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right)\geq\alpha$.

   This is equivalent to $\gamma\left(\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha)\right)\geq rT$. Since $\gamma\leq 0$, therefore, $\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha)\leq 0$, and as a result we have to check if

$$\gamma\leq\frac{rT}{\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha)}.$$

   In this case, $k'=0$ and there is no insurance contract.

2. If $a\leq N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right)<\alpha$.

   This is equivalent to $N\left(\frac{|m|\sqrt{T}}{\eta}+N^{-1}\left(1-\frac{1}{\delta}\right)\right)\leq N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right)<\alpha$.

   First, let us look to the right inequality.

   (a) If $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right)<\alpha$ and $\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha)\leq 0$ we get $\gamma>\frac{rT}{\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha)}$.

(b) If $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right) < \alpha$ and $\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha) > 0$ we get $\gamma < \frac{rT}{\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha)}$.

Second, the left inequality results in:

$$\frac{\left(\gamma\left(\mu+\frac{1}{2}(\gamma-1)\sigma^2\right)\right)\sqrt{T}}{-\gamma\sigma}-\frac{r\sqrt{T}}{-\gamma\sigma}+N^{-1}\left(1-\frac{1}{\delta}\right)$$
$$\leq \frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}=\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)\sqrt{T}}{-\gamma\sigma}+\frac{-r\sqrt{T}}{-\gamma\sigma},$$

which implies:

$$\frac{2N^{-1}\left(1-\frac{1}{\delta}\right)}{\sigma\sqrt{T}} \leq \gamma.$$

In this case, the contract is a stop-loss policy with retention level b^* that solves $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT-\log\left(1-\frac{b^*}{p_0}\right)}{-\gamma\sigma\sqrt{T}}\right)=\alpha$.

If we solve for b^* we find

$$b^* = p_0\left(1-\exp\left(\left(\left(\mu-\frac{1}{2}\sigma^2\right)T+\sigma\sqrt{T}N^{-1}(\alpha)\right)\gamma-rT\right)\right) \tag{5.8}$$

3. If $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right) < a$ or $\frac{2N^{-1}\left(1-\frac{1}{\delta}\right)}{\sigma\sqrt{T}} > \gamma$.

In this case, the contract is a two-layer contract with lower and upper retention levels a^*, b^*, where b^* is as in case 2 and a^* solves $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT-\log\left(1-\frac{a^*}{p_0}\right)}{-\gamma\sigma\sqrt{T}}\right)=a$, which similarly gives:

$$a^* = p_0\left(1-\exp\left(\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT+\gamma\sigma\sqrt{T}N^{-1}(a)\right)\right).$$

Note however that, $N^{-1}(a) = N^{-1}\left(N\left(\frac{|m|\sqrt{T}}{\eta} + N^{-1}\left(1-\frac{1}{\delta}\right)\right)\right) = \frac{|m|\sqrt{T}}{\eta} + N^{-1}\left(1-\frac{1}{\delta}\right)$. Therefore,

$$
\begin{aligned}
a^* &= p_0\left(1-\exp\left(\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T - rT\right.\right.\\
&\quad \left.\left.+\gamma\sigma\sqrt{T}\left(\frac{\left(\gamma\left(\mu+\frac{1}{2}(\gamma-1)\sigma^2\right)-r\right)\sqrt{T}}{-\gamma\sigma} + N^{-1}\left(1-\frac{1}{\delta}\right)\right)\right)\right)\\
&= p_0\left(1-\exp\left(\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T - rT \pm \gamma\left(\mu-\frac{1}{2}\sigma^2+\frac{1}{2}\gamma\sigma^2\right)T\right.\right.\\
&\quad \left.\left.+rT+\gamma\sigma\sqrt{T}N^{-1}\left(1-\frac{1}{\delta}\right)\right)\right)
\end{aligned}
$$

So finally, we get:

$$
a^* = p_0\left(1-\exp\left(-\frac{1}{2}\gamma^2\sigma^2 T + \gamma\sigma\sqrt{T}N^{-1}\left(1-\frac{1}{\delta}\right)\right)\right).
$$

5.4.5 *Calibration and Simulation for VaR*

Let us use the following calibration for the parameters: $=p_0 = 1$, $\mu = 0.01$, $r = 0.05$, $\alpha = 0.95$, $\delta = 1.1$, and let us consider $\gamma \in [-0.9, 0.1]$ and $\sigma \in [0.2, 0.9]$.

First, if we check $\left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma\sqrt{T}N^{-1}(\alpha) \geq 0$ then case 1 and case 2.a do not happen. From our observation in the parameter area that we study, this inequality holds; see the following Fig. 5.5. As a result, the lower retention level is always zero, i.e., $a^* = 0$.

Second, we find the upper retention level b^* using Eq. (5.8) and the result is graphed in the following Fig. 5.6. As one can see, increases in σ or $|\gamma|$ increase the retention level b^*.

Third, as shown in Fig. 5.7, the contract prices will also increase with an increase in either σ or $|\gamma|$. This may appear counter-intuitive since for a constant σ, if $|\gamma|$ increases then ϕ gets closer to zero suggesting that a risk neutral producer is willing to pay more for a risk management tool designed in this framework. This effect warrants some explanation. In designing the contracts, the risk aversion parameter ϕ (or γ) does not only affect the price volatility (i.e., $-\gamma\sigma$), but also it will have an impact on the lower and the upper retention levels. This means that, even though a larger $|\gamma|$ will result in greater volatility, it also changes the optimal contract. Ultimately, the impact from changing a contract will dominate the volatility effect

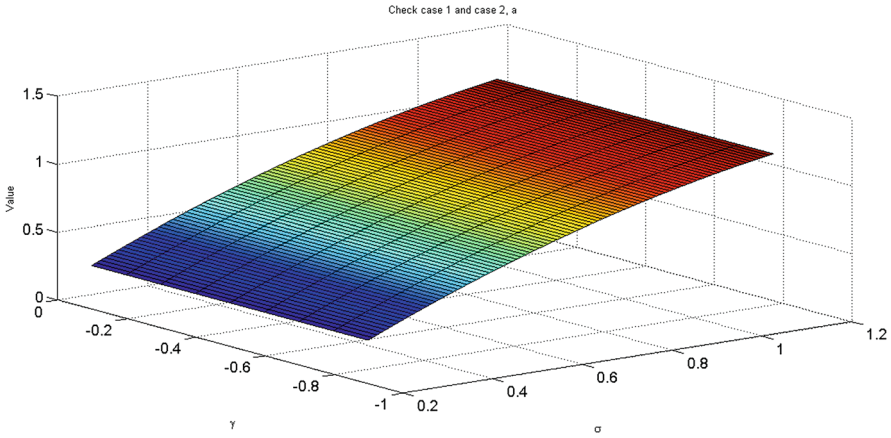


Fig. 5.5 Verifying cases 1 and 2.a in the parameter space σ-γ

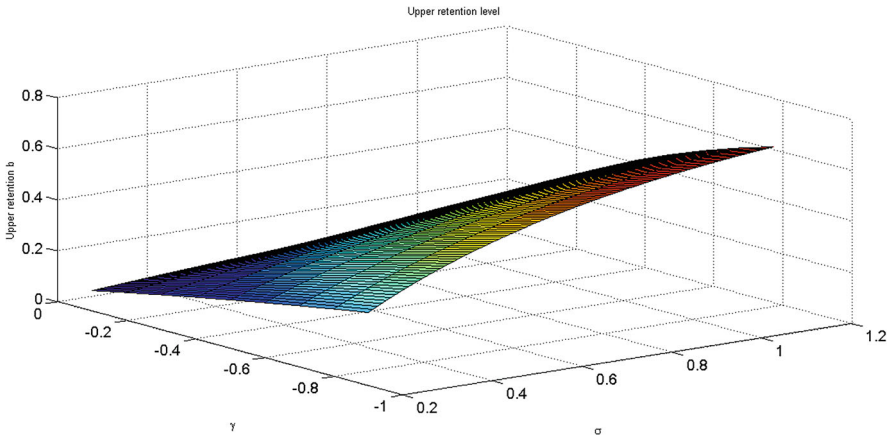


Fig. 5.6 Upper retention level b* in terms of parameters σ-γ

for a more risk-averse producer with larger ϕ and this results in a cheaper optimal contract.

Further, for the premium we need to discuss what parameters can be admissible. That means determining which parameters can generate a nonnegative market price of risk. For that, we need to check for different parameters the nonnegativity of $(\nu - r)$, namely:

$$\gamma\left(\mu + \frac{1}{2}(\gamma - 1)\sigma^2\right) - r = \frac{1}{2}\gamma^2\sigma^2 + \gamma\left(\mu - \frac{1}{2}\sigma^2\right) - r \geq 0$$

Verifying this condition, we report in the following Fig. 5.8 a graph which represents the area for which the market price of risk is nonnegative.

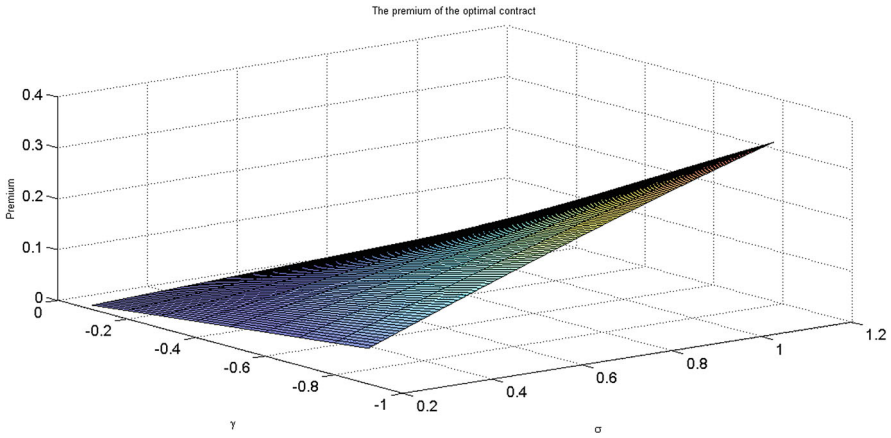


Fig. 5.7 Premium of the optimal contract in σ-γ space

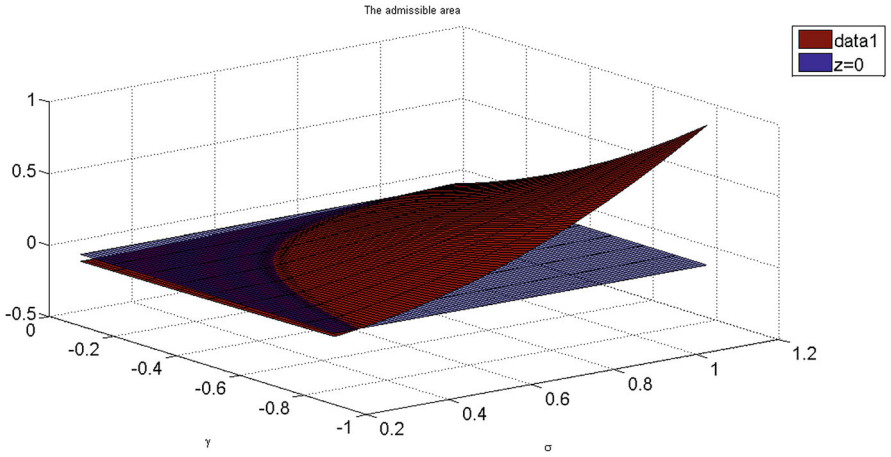


Fig. 5.8 Admissible area to ensure a nonnegative market price of risk, σ-γ space

5.4.6 Case for CVaR

As it has been mentioned earlier in Sect. 5.4.4, considering CVaR instead on VaR would not provide more insight into the problem as the only difference between the cases is that the second layer for CVaR is larger than VaR. However, one thing that is worth paying attention is that in both cases there is no contract if $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right) \geq \alpha$. This will be very useful since then one can assess this condition as a universal condition for the worthiness of considering a risk management tool. We will be using this condition in the empirical assessment to

see if there is a need to introduce risk management contracts, like insurances and derivatives, for the commodities we study. This can to some extent validate our theory, as all commodities we know have tradable derivatives.

To empirically assess the condition we have presented above (to determine whether a contract is justified for risk management), we compute for each commodity the value of N(.) using the estimated commodity-specific parameters for γ, μ and σ, as well as the risk-free rates r over the sample period. The details of the estimation procedure to obtain these parameters from the data are presented in the following section. The rate r is annualized and represents the 3-month US Treasury bill rate. For robustness, we compute N(.) using the mean, minimum, and maximum risk-free rate values over the sample period for each commodity. Since the sample period varies by commodity, the values of r also differ by commodity.

The results show that for all commodities, $N(.) < \alpha$ so the theoretical condition is empirically validated. $N(.)$ is always close to 0.5, and decreasing in the risk-free rate r. The result holds whether the commodities are more or less price-inelastic (based on the estimated parameter γ). Therefore, the result means that it is useful to introduce a contract, as the risk management tools are justified by the model developed in this paper. This conclusion holds whether we use α values of 0.9, 0.95, 0.99 or another plausible value for α. The results are also robust to whether we use mean, minimum or maximum values of r for each commodity. Overall, the conclusion is unambiguous that for all commodities, the risk management tools are relevant and should be introduced if they do not already exist.

5.5 Empirical Estimation of the Model Using Commodity Futures Price Data

5.5.1 Discretization of the Process and Estimation Procedure

This section reports empirical estimates of the model parameters, obtained using a large dataset of commodity futures prices and a GMM estimation approach. Recall that we have shown that the price process has the following SDF:

$$\frac{dp_t}{p_t} = \nu dt + \eta dB_t,\ p_0 > 0$$

where $\nu = \gamma\left(\mu + \frac{1}{2}(\gamma - 1)\sigma^2\right)$ and $\eta = -\gamma\sigma$. Following standard procedures in the literature (e.g., [21]), we estimate the parameters of this continuous-time model using a discrete-time econometric specification. We first need to derive the discrete-time version of the model, which is

$$\frac{\Delta p_t}{p_t} = \nu\Delta t + \eta\epsilon_{t+1}\sqrt{\Delta t}$$

Moreover, since $\Delta t = 1$,

$$\frac{p_{t+1} - p_t}{p_t} = \nu + \eta\epsilon_{t+1} \Longrightarrow p_{t+1} - p_t = \nu p_t + \eta\epsilon_{t+1} p_t$$

Then, considering a change of variable $\epsilon_{t+1} \leftarrow \eta\epsilon_{t+1} p_t$, we can write

$$p_{t+1} - p_t = \nu p_t + \epsilon_{t+1}$$

$$E\left[\epsilon_{t+1}\right] = 0, \quad E\left[\epsilon_{t+1}^2\right] = \eta^2 p_t^2 \tag{5.9}$$

using $E\left[\epsilon_{t+1}^2\right] = 1$. Now, define $\theta = (\theta_1, \theta_2)$ to be the vector of parameters with elements $\theta_1 = (\nu, \eta^2)$, $\theta_2 = (\gamma, \mu, \sigma)^{'}$. Then, we let the vector $f_t(\theta_1)$ be:

$$f_t\left(\theta_1\right) = \begin{bmatrix} \epsilon_{t+1} \\ \epsilon_{t+1} p_t \\ \epsilon_{t+1}^2 - \eta^2 p_t^2 \\ \left(\epsilon_{t+1}^2 - \eta^2 p_t^2\right) p_t \end{bmatrix}$$

Then, if substituting from (5.9), i.e., $p_{t+1} - p_t = \nu p_t + \epsilon_{t+1}$, we get:

$$f_t\left(\theta_1\right) = \begin{bmatrix} p_{t+1} - (1+\nu)\, p_t \\ \left(p_{t+1} - (1+\nu)\, p_t\right) p_t \\ \left(p_{t+1} - (1+\nu)\, p_t\right)^2 - \eta^2 p_t^2 \\ \left(\left(p_{t+1} - (1+\nu)\, p_t\right)^2 - \eta^2 p_t^2\right) p_t \end{bmatrix}$$

Under the null hypothesis implied by restrictions from equation above, $E[f_t(\theta_1)] = 0$. Then, the GMM procedure consists in replacing $E[f_t(\theta_1)]$ with $g_T(\theta_1)$, and using the T sample observations where

$$g_T\left(\theta_1\right) = T^{-1} \sum\nolimits_{t=1}^{T} f_t\left(\theta_1\right)$$

and then choosing the parameters in θ that minimize

$$J_T\left(\theta_1\right) = g_T'\left(\theta_1\right) W_T\left(\theta_1\right) g_T\left(\theta_1\right) \tag{5.10}$$

where $W_T(\theta_1)$ is a positive-definite symmetric weighting matrix. To test the suitability of the model (i.e., the null hypothesis of correct model specification and the orthogonality conditions required for using GMM estimation), we compute Hansen's J-test (Sargan test of overidentification). A higher p-value for the J-test means that the instruments satisfy the orthogonality conditions and that the model specification is correct.

The GMM estimation method we use has two stages. First, we estimate the parameter values θ_1 that minimize Eq. (5.10). Second, we replace the parameters $\theta_1 = (\nu, \eta^2)$ in Eq. (5.9) with the first-stage parameter estimates and rewrite it as follows:

$$p_{t+1} - p_t = \left(\gamma\mu + \frac{\hat{\eta}^2}{2} - \frac{\hat{\eta}}{2\gamma}\right) p_t + \epsilon_{t+1}$$

The resulting equation is then estimated by GMM to obtain θ_2.

5.5.2 *Description of the Data and Empirical Results*

The dataset consists of daily settlement prices for 19 commodity futures contracts obtained from Thomson Datastream, generally ending in late May 2018. At date t for a given commodity, there are a number M_t of contracts traded, where the first nearby is the nearest to maturity (expiry). Thus, the dataset is an unbalanced panel. For our purposes, only the nearby futures contract is needed. Rather than use the continuous series provided by Datastream for a given commodity futures contract (which introduces a splicing bias), we construct each time series of observations by rolling over from the first to second nearby contract on the 15th day of the month preceding maturity. This rollover method is standard and avoids including observations for dates near contract maturity, as the latter may not be reliable prices. All series use at least 10 years of daily observations. The specific length of the time series depends on data availability in Datastream.

Tables 5.1 and 5.2 present descriptive statistics for commodity futures prices and returns, respectively, for all contracts used in the analysis. Table 5.1 shows that most commodity price series are right-skewed and display negative excess kurtosis. Table 5.2 shows that for a majority of commodities, price log returns are right-skewed, while all display positive excess kurtosis.

Results for the estimated model, using GMM and applied to each of the 19 commodity series, are presented in Table 5.3. First, for all 19 series, we fail to reject the null hypothesis of the Sargan J-test. Thus, the overidentifying restrictions are valid. Second, the results show that the main parameter of interest, γ, is around -1 for all series, implying roughly unit elasticity. For more than half of the commodities in our sample, the estimated γ is greater than 1 in absolute value. This result implies that quantity responds strongly to price changes. These commodities are Chicago ethanol, Oman crude oil, cocoa, coffee, corn, soybean oil, oats, lean hogs, live cattle, feeder cattle, and gold. On the other hand, the following commodities have an estimated γ less than 1 in absolute value: WTI crude oil, Dubai crude, wheat, sugar, rough rice, soybean meal, orange juice, and lumber. These results are consistent with prior empirical evidence on the price-elasticity or price-inelasticity of these commodities. For example crude oil is well known to be price-inelastic. The parameter estimates suggest that prices of gold, ethanol, and soybean oil are the

Table 5.1 Descriptive statistics for daily settlement prices of 19 commodity futures contract time series

Futures	Mean	Med.	Std	Q25	Q75	Skew	Exc. Kurt.	Nb.obs	Start	End
Energy										
Chicago ethanol	1.92	1.79	0.43	1.55	2.27	0.48	−1.02	2984	12/15/06	5/23/18
WTI crude oil	76.46	84.24	23.68	54.44	96	0.02	−1.38	1958	11/29/10	5/30/18
Dubai crude oil	63.36	61.23	39.71	40.27	103.7	−0.33	−1.3	2608	1/2/08	12/29/17
Oman crude oil	79.67	79.1	25.12	55.62	104.3	0	−1.27	2870	6/1/07	5/31/18
Storable agricultural										
Wheat	449.1	449.1	192.7	316.8	563.6	0.54	−0.18	6428	10/4/99	5/23/18
Sugar	13.75	15.13	9.17	5.31	19.55	−0.06	−0.77	9865	8/1/80	5/24/18
Cocoa	1937	1791	615.6	1.09	2192	0.85	−0.21	2920	3/16/07	5/24/18
Coffee	145.71	136.5	43.03	120.1	162	1.2	1.5	3819	10/6/03	5/24/18
Corn	468.55	421.1	115.6	384.5	550.4	1.03	0.19	2622	8/8/08	5/25/18
Rough rice	1189	1192	306.6	994	1438	−0.11	−0.5	6016	5/10/95	5/25/18
Soybean oil	41.51	38.81	9.68	33.32	49.84	0.7	−0.67	2984	12/15/06	5/23/18
Oats	298	292	69.11	239.1	352	0.36	−0.65	2878	5/15/07	5/23/18
Soybean meal	325.3	320	55.07	289.3	358.7	0.61	0.99	2984	12/15/06	5/23/18
Orange juice	140.7	141	28.27	124.4	155.8	−0.02	0.38	2822	8/1/07	5/24/18
Non-storable agricultural										
Lean hogs	78.23	79.25	14.9	68.16	86.68	0.54	1.21	2679	2/18/08	5/24/18
Live cattle	118.5	119.4	21.23	102.7	132.7	0.15	−0.62	2789	9/17/07	5/24/18
Feeder cattle	144.4	143.4	35.84	115.2	157.7	0.72	0.05	2654	3/24/08	5/24/18
Other										
RL lumber	291.1	295	53.78	252.2	333.1	−0.31	−0.54	1746	10/24/08	7/3/15
Gold	690.6	690.6	571.48	70.02	1106	0.38	−1.11	3628	6/30/04	5/25/18

Notes: Time series are constructed from daily settlement price observations. The first nearby futures contract is used, except near maturity when the second nearby is used. The series is rolled over from the first to second nearby contract on the 15th day of the month preceding maturity, to avoid including observations near contract expiry. The statistics are, in order: mean, median, standard deviation, 25th quantile, 75th quantile, skewness, excess kurtosis, number of daily observations for the commodity, sample start date, sample end date

Table 5.2 Descriptive statistics for daily price log returns of 19 commodity futures contract time series

Futures	Mean	Med.	Std	Q25	Q75	Skew	Exc. kurt.	Nb.obs	Start	End
Energy										
Chicago ethanol	0	0	0.02	−0.004	0.007	−3.33	33.56	2983	12/15/06	5/23/18
Crude oil	0	0	0.02	−0.008	0.008	0.32	14.11	1957	11/29/10	5/30/18
Dubai crude oil	−0.02	0	1.34	−0.01	0.009	−12.99	595.6	2607	1/2/08	12/29/17
Oman crude oil	0	0	0.02	−0.009	0.009	0.12	5.71	2869	6/1/07	5/31/18
Storable agricultural										
Wheat	0	0	0.07	−0.008	0.008	23	715.56	6427	10/4/93	5/23/18
Sugar	0.02	0	1.47	−0.01	0.01	60.5	3660.3	9864	8/1/80	5/24/18
Cocoa	0	0	0.03	−0.008	0.008	4.79	131.24	2919	3/16/07	5/24/18
Coffee	0	0	0.02	−0.01	0.01	1.64	24.99	3818	10/6/03	5/24/18
Corn	0	0	0.02	−0.009	0.008	0.5	9.01	2621	8/8/08	5/25/18
Rough rice	0	0	0.02	−0.006	0.006	25.48	1150.4	6015	5/10/95	5/25/18
Soybean oil	0	0	0.01	−0.008	0.008	0.22	2.74	2983	12/15/06	5/23/18
Soybean meal	0	0	0.02	−0.008	0.008	−1.5	20.24	2983	12/15/06	5/23/18
Orange juice	0	0	0.02	−0.01	0.01	0.27	3.88	2821	8/1/07	5/24/18
Oats	0	0	0.02	−0.009	0.009	0.08	11.43	2877	5/15/07	5/23/18
Non-storable agricultural										
Lean hogs	0	0	0.02	−0.008	0.008	1.58	36.13	2678	2/18/08	5/24/18
Live cattle	0	0	0.01	−0.004	0.005	−1.96	37.1	2788	9/17/07	5/24/18
Feeder cattle	0	0	0.01	−0.005	0.005	−0.25	3.36	2653	3/24/08	5/24/18
Other										
RL lumber	0	0	0.02	−0.008	0.006	4.36	74.26	1745	10/24/08	7/3/15
Gold	0	0	0.02	−0.006	0.008	−21.06	899.3	3627	6/30/04	5/25/18

Notes: Time series are constructed from daily settlement price observations. The first nearby futures contract is used, except near maturity when the second nearby is used. The series is rolled over from the first to second nearby contract on the 15th day of the month preceding maturity, to avoid including observations near contract expiry. The statistics are, in order: mean, median, standard deviation, 25th quantile, 75th quantile, skewness, excess kurtosis, number of daily observations for the commodity, sample start date, sample end date

most sensitive to changing supply conditions, while wheat and soybean meal are the least sensitive.

Our theory can properly specify if one needs to introduce a contract on a commodity index price (e.g., insurance, futures or option). As discussed earlier, for both cases of VaR and CVaR we need to introduce a contract if $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right) < \alpha$. This can be regarded as a universal condition for both risk measures, VaR and CVaR. We can also look at this condition from a different perspective, namely that there would be no need to introduce a contract if the risk aversion confidence parameter is smaller than $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)T-rT}{-\gamma\sigma\sqrt{T}}\right)$. However, before measuring this quantity, it is important to understand how it will change for different time intervals. To see this, let us consider the annual based volatility and drift σ and μ. Thinking of contracts with a one-year expiration date, i.e., $T = 1$, this quantity is equal to $N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)-r}{-\gamma\sigma}\right)$. Given the Black-Scholes model that we already considered for the index prices, for any interval Δt we have that

$$\sigma_{\Delta t} = \sqrt{\Delta t}\sigma \quad \text{and} \quad \mu_{\Delta t} = \Delta t\ \mu,$$

where, $\sigma_{\Delta t}$ and $\mu_{\Delta t}$, are the volatility and drift for the time interval Δt. Replacing them above we get

$$N\left(\frac{\gamma\left(\mu-\frac{1}{2}\sigma^2\right)-r}{-\gamma\sigma}\right) = N\left(\frac{\gamma\left(\frac{\mu_{\Delta t}}{\Delta t}-\frac{\frac{1}{2}{\sigma_{\Delta t}}^2}{\Delta t}\right)-r}{-\frac{\gamma\sigma_{\Delta t}}{\sqrt{\Delta t}}}\right)$$

$$= N\left(\frac{r}{\sigma_{\Delta t}}\frac{\sqrt{\Delta t}}{\gamma}-\frac{1}{\sqrt{\Delta t}}\left(\frac{\mu_{\Delta t}}{\sigma_{\Delta t}}-\frac{1}{2}\sigma_{\Delta t}\right)\right). \tag{5.11}$$

To empirically assess the condition we have presented above (determining whether a contract should be introduced), we compute for each commodity the value in the Eq. (5.11) using the estimated commodity-specific parameters for γ, μ and σ, as well as the risk-free rates over the sample period. Note that since the sample period varies by commodity, we also use matched risk-free rates. Specifically, we compute the value in (5.11) for each of the three time intervals 0.25, 0.5 and 1, using the mean, minimum, and maximum risk-free rate values during the sample period for each commodity. The results presented in Tables 5.4, 5.5 and 5.6 show that for all commodities, the condition is empirically validated: Thus, it is useful to introduce a contract, as these risk management tools are justified by the model developed in the paper. This finding is robust to whether we use mean, minimum or maximum risk-free rates. Indeed, if one looks more carefully, the value of the argument in (5.11) is

Table 5.3 Estimates of the model parameters for 19 commodity futures, using GMM

	First-stage			Second-stage				N.obs.
Futures	θ_1			θ_2				
Contract	ν (s.e.)	η (s.e.)	J-test (p-value)	γ (s.e.)	μ (s.e.)	σ	J-test (p-value)	
Energy								
Chicago ethanol	0.0002(0.0004)	0.0173**(0.0063)	4.515(0.105)	−1.368**(0.06)	0.0007*(0.0003)	0.0141	1.347(0.51)	2984
WTI crude oil	−0.0002(0.0003)	0.0173**(0.0044)	2.551(0.279)	−0.94**(0.041)	0.0005(0.0003)	0.0173	2.551(0.279)	1958
Dubai crude	−0.0002(0.0003)	0.0141**(0.001)	1.36(0.507)	−0.977**(0.033)	0.0002(0.0003)	0.0141	1.426(0.49)	2608
Oman crude	0.001(0.0003)	0.0173**(0.0045)	2.348(0.309)	−1.152**(0.03)	0.0002(0.0003)	0.0173	5.416(0.07)	2870
Storable agricultural								
Wheat	0.007**(0.0005)	0.0447**(0.02)	0.645(0.71)	−0.846**(0.105)	0.0029**(0.0006)	0.0548	3.971(0.137)	6428
Sugar	−0.0008(0.0004)	0.0548**(0.0071)	1.969(0.374)	−0.955**(0.043)	0.001*(0.0005)	0.0265	1.728(0.422)	9865
Cocoa	0.0008°(0.0005)	0.0316**(0.0141)	0.013(0.993)	−1.202**(0.082)	0.0224(0.02)	0.0265	3.983(0.137)	2920
Coffee	0.0002(0.0004)	0.02**(0.0045)	3.308(0.191)	−1.007**(0.025)	0.00003(0.096)	0.02	4.199(0.123)	3819
Corn	0.0001(0.0003)	0.0173**(0.0044)	2.944(0.229)	−1.013**(0.0003)	0.0002(0.0003)	0.0173	2.91(0.233)	2622
Rough rice	0.00001(0.0002)	0.0141**(0.0032)	2.135(0.344)	−0.89**(0.026)	0.0003(0.0002)	0.0173	2.329(0.312)	6016
Soybean meal	0.025**(0.0003)	0.0316**(0.0045)	3.475(0.176)	−0.521**(0.014)	0.002**(0.0006)	0.0632	5.642(0.06)	2984
Soybean oil	0.0004(0.0003)	0.0141**(0.0032)	3.702(0.157)	−1.665**(0.074)	0.0005*(0.0002)	0.00836	1.328(0.515)	2984
FC orange juice	0.007**(0.0004)	0.0224**(0.0045)	1.775(0.412)	−0.875**(0.024)	0.0003(0.0005)	0.0264	1.198(0.549)	2822
Oats	0.0001(0.0004)	0.0173**(0.0055)	3.429(0.181)	−1.06**(0.034)	0.0002(0.0003)	0.0173	3.324(0.19)	2878
Nonstorable agricultural								
Lean hogs	−0.0004(0.0003)	0.0173**(0.0055)	2.68(0.262)	−1.13**(0.05)	0.0003(0.0003)	0.0141	2.661(0.264)	2679
Live cattle	0.0006**(0.0002)	0.01**(0.0032)	1.331(0.514)	−1.064**(0.042)	−0.00001(0.0002)	0.01	2.745(0.254)	2789
Feeder cattle	0.0001(0.0002)	0.01**(0.00224)	2.15(0.341)	−1.044**(0.026)	−0.00001(0.0002)	0.01	2.15(0.341)	2654
Other								
RL lumber	−0.00004(0.0004)	0.0141**(0.00316)	4.532(0.104)	−0.989**(0.031)	0.0002(0.0004)	0.0141	4.532(0.104)	1746
Gold	0.0004(0.0003)	0.0141**(0.0089)	3.35(0.187)	−1.35**(0.19)	−0.0001(0.0002)	0.01	3.111(0.211)	3628

Notes: This table presents estimates of the parameters of the price process described in this paper, for a range of different futures contracts. The data are obtained from Thomson Datastream. Prices are daily settlement. The following discrete-time equation is estimated using the Generalized Method of Moments (see the paper for the derivation of this equation, and see the algorithm described in appendix A). The null hypothesis of the Hansen J-Test is that the overidentification restrictions are valid, so the instruments are valid. Statistical significance is denoted using ** (1% level), * (5% level), and ° (10% level)

$p_{t+1} - p_t = \nu p_t + \epsilon_{t+1}, E[\epsilon_{t+1}] = 0, E\left[\epsilon_{t+1}^2\right] = \eta_t^2 p_t^2, \nu = \gamma\left(\mu + \frac{1}{2}(\gamma - 1)\sigma^2\right)$, and $\eta = -\gamma\sigma$

Table 5.4 Estimates of the quantity N for 19 commodity futures, $T = 0.25$

					r			$N(.)$	
Futures	γ	μ	σ	min	avg	max	min	avg	max
Energy									
Chicago ethanol	−1.368	0.0007	0.0141	0.01	0.76	5.03	0.5	0.5	0.497
Crude oil	−0.94	0.0005	0.0173	0.01	0.31	1.90	0.5	0.5	0.498
Dubai crude oil	−0.977	0.0002	0.0141	0.01	0.38	2.75	0.5	0.5	0.498
Oman crude oil	−1.152	0.0002	0.0173	0.01	0.57	4.82	0.5	0.5	0.496
Storable agricultural									
Wheat	−0.846	0.0029	0.0548	0.01	2.40	6.17	0.5	0.492	0.480
Sugar	−0.955	0.001	0.0265	0.01	4.21	16.30	0.5	0.494	0.477
Cocoa	−1.202	0.0224	0.0265	0.01	0.67	4.94	0.5	0.499	0.495
Coffee	−1.007	0.00003	0.02	0.01	1.21	5.03	0.5	0.499	0.495
Corn	−1.013	0.0002	0.0173	0.01	0.51	4.20	0.5	0.5	0.496
Rough rice	−0.89	0.0003	0.0173	0.01	2.27	6.17	0.5	0.498	0.494
Soybean oil	−0.521	0.002	0.0632	0.01	0.76	5.03	0.5	0.463	0.271
Soybean meal	−1.665	0.0005	0.00836	0.01	0.76	5.03	0.5	0.5	0.499
Orange juice	−0.875	0.0003	0.0264	0.01	0.51	4.20	0.5	0.5	0.494
Oats	−1.06	0.0002	0.0173	0.01	0.60	4.82	0.5	0.5	0.496
Non-storable agricultural									
Lean hogs	−1.13	0.0003	0.0141	0.01	0.37	2.12	0.5	0.5	0.498
Live cattle	−1.064	−0.00001	0.01	0.01	0.48	3.90	0.5	0.5	0.498
Feeder cattle	−1.044	−0.00001	0.01	0.01	0.35	1.90	0.5	0.5	0.5
Other									
RL lumber	−0.989	0.0002	0.0141	0.01	0.28	1.90	0.5	0.5	0.499
Gold	−1.35	−0.0001	0.01	0.01	1.22	5.03	0.5	0.498	0.494

almost zero for all cases and that is why the value of the Eq. (5.11) is a very close neighbor of 0.5. This result shows the condition under which we need to introduce a contract holds whether we use 0.9, 0.95, 0.99 or another plausible value for α.

These new results can be used to improve risk management practice and derivative pricing, but such applications are beyond the scope of this paper.

5.6 Conclusion

The financialization of commodities has brought renewed interest in finance and risk management research to this asset class. Black's model [1] remains the standard for pricing commodity derivatives, and most models are said to be reduced-form in the style of [2]. To gain a deeper understanding of these markets, both for exchange-traded derivatives as well as insurance instruments, it is important to explicitly model the economic variables that determine the stochastic price process. To obtain explicit solutions to this problem, this paper. The contingent claim methodology that

Table 5.5 Estimates of the quantity N for 19 commodity futures, $T = 0.5$

					r			$N(.)$	
Futures	γ	μ	σ	min	avg	max	min	avg	max
Energy									
Chicago ethanol	−1.368	0.0007	0.0141	0.01	0.76	5.03	0.5	0.499	0.493
Crude oil	−0.94	0.0005	0.0173	0.01	0.31	1.90	0.5	0.5	0.495
Dubai crude oil	−0.977	0.0002	0.0141	0.01	0.38	2.75	0.5	0.5	0.494
Oman crude oil	−1.152	0.0002	0.0173	0.01	0.57	4.82	0.5	0.499	0.490
Storable agricultural									
Wheat	−0.846	0.0029	0.0548	0.01	2.40	6.17	0.5	0.478	0.444
Sugar	−0.955	0.001	0.0265	0.01	4.21	16.30	0.5	0.483	0.436
Cocoa	−1.202	0.0224	0.0265	0.01	0.67	4.94	0.5	0.498	0.485
Coffee	−1.007	0.00003	0.02	0.01	1.21	5.03	0.5	0.497	0.486
Corn	−1.013	0.0002	0.0173	0.01	0.51	4.20	0.5	0.499	0.490
Rough rice	−0.89	0.0003	0.0173	0.01	2.27	6.17	0.5	0.494	0.483
Soybean oil	−0.521	0.002	0.0632	0.01	0.76	5.03	0.5	0.495	0.470
Soybean meal	−1.665	0.0005	0.00836	0.01	0.76	5.03	0.5	0.5	0.496
Orange juice	−0.875	0.0003	0.0264	0.01	0.51	4.20	0.5	0.498	0.482
Oats	−1.06	0.0002	0.0173	0.01	0.60	4.82	0.5	0.499	0.489
Non-storable agricultural									
Lean hogs	−1.13	0.0003	0.0141	0.01	0.37	2.12	0.5	0.5	0.496
Live cattle	−1.064	−0.00001	0.01	0.01	0.48	3.90	0.5	0.5	0.495
Feeder cattle	−1.044	−0.00001	0.01	0.01	0.35	1.90	0.5	0.5	0.497
Other									
RL lumber	−0.989	0.0002	0.0141	0.01	0.28	1.90	0.5	0.5	0.496
Gold	−1.35	−0.0001	0.01	0.01	1.22	5.03	0.5	0.499	0.495

is proposed here is inspired by the rational storage models of Deaton and Larocque [8–10] and based on standard risk-neutral valuation arguments. Therefore, this paper develops a framework to price commodity derivatives and optimal insurance contracts that has more structural features than typically found in the literature. The framework can be applied to exchange-traded or OTC derivatives, or to insurance instruments.

In this paper, we show how to obtain commodity-specific pricing solutions in terms of deeper economic parameters such as the price elasticity of demand for a given commodity, as well as the loss function that best describes the trader or hedger (e.g. Value-at-risk, or conditional Value-at-risk). We also consider the role played by the risk specification among a class of distortion risk measures. Results are presented for some risk management applications, where optimal contract types are obtained in terms of the parameter space. In some cases, no contract is optimal. These findings should be useful for academics and practitioners in commodity finance, derivatives, risk management and insurance.

The analysis described in the paper also suggests some potentially useful avenues for further research. One is to take the model to data on commodity futures

Table 5.6 Estimates of the quantity N for 19 commodity futures, $T = 1$

					r			$N(.)$	
Futures	γ	μ	σ	min	avg	max	min	avg	max
Energy									
Chicago ethanol	−1.368	0.0007	0.0141	0.01	0.76	5.03	0.5	0.497	0.479
Crude oil	−0.94	0.0005	0.0173	0.01	0.31	1.90	0.5	0.498	0.486
Dubai crude oil	−0.977	0.0002	0.0141	0.01	0.38	2.75	0.5	0.498	0.484
Oman crude oil	−1.152	0.0002	0.0173	0.01	0.57	4.82	0.5	0.497	0.471
Storable agricultural									
Wheat	−0.846	0.0029	0.0548	0.01	2.40	6.17	0.5	0.438	0.345
Sugar	−0.955	0.001	0.0265	0.01	4.21	16.30	0.5	0.453	0.325
Cocoa	−1.202	0.0224	0.0265	0.01	0.67	4.94	0.5	0.494	0.456
Coffee	−1.007	0.00003	0.02	0.01	1.21	5.03	0.5	0.490	0.460
Corn	−1.013	0.0002	0.0173	0.01	0.51	4.20	0.5	0.497	0.471
Rough rice	−0.89	0.0003	0.0173	0.01	2.27	6.17	0.5	0.482	0.452
Soybean oil	−0.521	0.002	0.0632	0.01	0.76	5.03	0.5	0.463	0.271
Soybean meal	−1.665	0.0005	0.00836	0.01	0.76	5.03	0.5	0.498	0.490
Orange juice	−.0875	0.0003	0.0264	0.01	0.51	4.20	0.5	0.494	0.450
Oats	−1.06	0.0002	0.0173	0.01	0.60	4.82	0.5	0.496	0.469
Non-storable agricultural									
Lean hogs	−1.13	0.0003	0.0141	0.01	0.37	2.12	0.5	0.498	0.489
Live cattle	−1.064	−0.00001	0.01	0.01	0.48	3.90	0.5	0.498	0.485
Feeder cattle	−1.044	−0.00001	0.01	0.01	0.35	1.90	0.5	0.499	0.493
Other									
RL lumber	−0.989	0.0002	0.0141	0.01	0.28	1.90	0.5	0.498	0.489
Gold	−1.35	−0.0001	0.01	0.01	1.22	5.03	0.5	0.496	0.485

and options contracts to recover estimates of the parameters and compare pricing accuracy relative to commonly used methods. A second would be to investigate empirically the no-optimal-contract case by relating the model's prediction to data on contract trading volume and interest. We saw that all the commodities need some type of risk management tool. Furthermore, the numbers that we have found for the values of $N(.)$ in Eq. (5.11) show this is almost always true, as the risk aversion parameters are way above the point that would necessitate the introduction of derivatives.

References

1. Assa, H. (2016). Financial engineering in pricing agricultural derivatives based on demand and volatility. *Agricultural Finance Review, 76*(1), 42–53.
2. Geman, H. (2015). *Agricultural finance: From crops to land, water and infrastructure*. Wiley Finance Series.

3. Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics, 3*(1–2), 167–179.
4. Gibson, R., & Schwartz, E. S. (1990). Stochastic convenience yield and the pricing of oil contingent claims. *Journal of Finance, 45*(3), 959–976.
5. Nielsen, M. J., & Schwartz, E. S. (2004). Theory of storage and the pricing of commodity claims. *Review of Derivatives Research, 7*(1), 5–24.
6. Geman, H., & Shih, Y. F. (2009). Modelling commodity prices under the CEV model. *Journal of Alternative Investments, 11*(3), 65–84.
7. Ribeiro, D. R., & Hodges, S. D. (2004). *A two-factor model for commodity prices and futures valuation*. EFMA 2004 Basel meetings paper.
8. Deaton, A., & Laroque, G. (1992). On the behaviour of commodity prices. *Review of Economic Studies, 59*(1), 1–23.
9. Deaton, A., & Laroque, G. (1995). Estimating a nonlinear rational expectations commodity price model with unobservable state variables. *Journal of Applied Econometrics, 10*(Suppl), S9–S40.
10. Deaton, A., & Laroque, G. (1996). Competitive storage and commodity price dynamics. *Journal of Political Economy, 104*(5), 896–923.
11. Cox, J. (1975). *Notes on option pricing I: Constant elasticity of variance diffusions*. Working paper, Stanford University.
12. Basak, S., & Pavlova, A. (2016). A model of financialization of commodities. *Journal of Finance, 71*(4), 1511–1556.
13. Cheng, I.-H., & Xiong, W. (2014). The Financialization of commodity markets. *Annual Review of Financial Economics, 6*, 419–441.
14. Tang, K., & Xiong, W. (2012). Index investment and financialization of commodities. *Financial Analysts Journal, 68*(6), 54–74.
15 Cong, J., Tan, K. S., & Weng, C. (2012). VaR-based optimal partial hedging. *Astin Bulletin, 43*(3), 271–299.
16. Cong, J., Tan, K. S., & Weng, C. (2014). Conditional value-at-risk-based optimal partial hedging. *Journal of Risk, 16*(3), 49–83.
17. Assa, H. (2015). On optimal reinsurance policy with distortion risk measures and premiums. *Insurance Mathematics and Economics, 61*, 70–75.
18. Assa, H. (2015). Risk management under a prudential policy. *Decisions in Economics and Finance, 38*(2), 217–230.
19. Zhuang, S. C., Weng, C., Tan, K. S., & Assa, H. (2016). Marginal indemnification function formulation for optimal reinsurance. *Insurance: Mathematics and Economics, 67*, 65–76.
20. Assa, H., Wang, M., & Pantelous, A. A. (2018). Modeling frost losses: Application to pricing frost insurance. *North American Actuarial Journal, 22*(1), 137–159.
21. Chan, K. C., Karolyi, G. A., Longstaff, F. A., & Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance, 47*(3), 1209–1227.

Chapter 6
Empirical Results for Cross-Hedging in the Incomplete Market

Jess Carr and Simon Wang

Abstract This paper examines different hedging techniques for options written on non-exchange-traded agricultural commodities using the futures markets to hedge, and evaluates the performance using statistical measures. The paper applies the hedging methods to real agricultural commodity data from the USDA. In these markets there are a number of commercial risks, such as weather and supply-chain disruption, which need to be managed by both producers and consumers. Typically, there is no perfectly correlated hedging instrument available for the product being traded, and as such there is basis risk present when trying to find a hedging solution. This highlights the need for empirical studies which address the problem of how to hedge in this environment. We evaluate static and dynamic hedging strategies for European options written on livestock indices using live cattle futures to hedge. Hedging methods based on delta, minimum variance, value-at-risk (VaR), and conditional VaR (C-VaR) are tested. Hedging performance is examined by hedging effectiveness (i.e calculating risk reduction versus an unhedged portfolio) and distribution statistics. Overall, we found that the static minimum-variance technique provided the best hedging performance in terms of risk reduction versus the unhedged portfolio.

6.1 Introduction

Commodity markets have undergone a number of significant changes in the past few decades. In particular, the proportion of commodities traded over-the-counter has grown substantially compared to exchange-traded volumes, and the market has expanded to include more sophisticated financial instruments [1], including indexes and derivatives. This has enabled the development of more complex hedging and risk management strategies. Furthermore, the increased volatility of the underlying

J. Carr (✉) · S. Wang
Stable Group Ltd., London, UK

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_6

commodity prices has made the commercial requirement for risk management solutions clear.

In this paper we focus on agricultural commodity markets. In these markets there are a number of risks which highlight the need for empirical studies demonstrating how these can be effectively managed even when no direct hedging strategy is feasible. In particular, both producers and consumers face risks of:

- Weather—crop yields and production are heavily dependent on rainfall, temperature and other factors, with climate change making these conditions less dependable year-on-year
- Disease and pests
- Energy price volatility—the energy required for production, transportation and the increased use of bio-diesel
- Government policy and regulation changes
- Geopolitical risk and exchange rate exposure

Typically, in this industry the primary risk management tools, aside from futures-based hedging, are physical hedging, swaps and insurance. Physical hedging is closely related to futures-based hedging, except involves storing or being willing to store the physical good, which can be effective when the infrastructure for storage is already available within the context of the business. However, this can be a challenging strategy to execute effectively, given that in agricultural commodity markets certain goods have a short storage life, or the cost of storage and transportation can be high, limiting the usefulness of the hedge. Additionally, given these physical constraints and the niche nature of these markets, it can be difficult to adjust positions quickly or find counterparties with favourable pricing. Also, this strategy may not protect against all types of price risk. Swaps are derivative contracts where two parties exchange cash flows based on the value of the underlying asset. These can be individually tailored since they are over-the-counter, so the contract specifications can be closely aligned to the risk faced. They also do not require an upfront premium payment. However, there is significant counterparty risk, liquidity risk and credit risk. Both physical hedging and swaps require comprehensive market knowledge and significant investment into employee time and expertise in order to execute the strategies that provide cost-effective and comprehensive coverage.

The agricultural insurance industry is a large market, with a broad range of available of products, from crop and livestock insurances to revenue insurance, multi-peril and in more recent years, CAT (catastrophic) risk products. Crop insurance is typically linked to yield or whole-farm revenue protection, and similarly livestock insurances provide coverage for death or disease of the livestock. In the cases of the beef markets that we discuss in this paper, these insurances can be useful for specific risks but may lack coverage in the general case of price movement. Multi-peril or CAT insurance has broader utility but since it includes many high-risk events it can come with higher premiums and these events may not necessarily result in equivalent price movement. Commodity price insurance can aid with this, and we explore some examples of this in the literature review, but for the companies providing the insurance, the problem of hedging the risk remains.

For all participants, whether it be companies selling risk management strategies who need to offset the risk they onboard or direct producers/consumers, derivatives can be used to facilitate their hedging strategies. Derivatives enable this by offsetting potential losses from the underlying asset by allowing counterbalancing positions in related securities to be taken. For many commodities which have these exposures, the desired derivative for hedging is illiquid or non-existent, and in such incomplete markets, actors must rely on cross-hedging, which introduces basis risk as the cross-hedge fails to fully offset losses from the underlying asset due to imperfect correlation.

There are a number of constraints which make the hedging of non-exchange-traded commodities particularly challenging. There is a lack of liquidity as many agricultural commodities do not have active trading and derivatives markets, making price discovery difficult, with larger bid-ask spreads and limited hedging opportunities. Estimating the implied volatility for such markets can be challenging given this lack of actively traded options. Furthermore, historical price and volatility data can be sparse for these commodities, making it difficult to model, price derivatives, backtest and evaluate potential hedging strategies. Given these constraints, it can be hard to determine the best set of hedging instruments for a given commodity, and after these are determined we then need to estimate the optimal hedging ratio for the frequency of hedging we wish to (and feasibly can) execute.

Historically, we have seen many instances where despite these difficulties hedging on the futures would have potentially provided significant commercial value, and we provide an example here that is specific to the beef market, which was the main commodity we focused on in this study. In 2003, Mad Cow Disease severely disrupted the US beef market as other countries imposed bans on importing US beef, with exports from the US dropping by approximately 80% [2] after 2003. Through beef producers obtaining short positions in futures contracts, there would have been some revenue protection despite the sharp drop in prices [3]. Conversely, in 2020 due to the Covid-19 pandemic, beef prices increased significantly [4], resulting in insurance companies having to provide large payouts which could have been hedged through the futures market. The supply chain disruption affected every participant in the supply chain, with many meat processing plants closing and supply chain bottlenecks, which led to higher prices for food, with retail beef prices increasing 25% in June 2020 year-on-year [5]. This highlighted the need for protection across the entire supply chain. In our paper we examine beef, cattle and beef trim indexes, which all play a key role in the US beef markets. Whilst the findings in this paper are intended to be applicable to broader agricultural commodity markets, the beef industry alone plays a vital role in the US agricultural economy, with the total value of the US beef cattle industry estimated to be worth $101.6bn billion dollars in 2024 [6]. Despite the historical evidence that beef and other agricultural commodity markets are very susceptible to both global and regional disruption of all kinds from weather to geopolitics, there have still been very few empirical studies done to test and develop hedging strategies where the futures can provide some protection with some remaining basis risk. This is particularly relevant to companies in the middle of the supply chain who have risk in both directions and handle many different

closely related products, such as different cuts of beef, and insurers who many be exposed on multiple fronts. This study empirically highlights the challenge of hedging options using correlated futures and includes commercial considerations such as transaction cost, making it a key contribution to a broadly unsolved problem that could impact a large swathe of the US agricultural and insurance economies. This is particularly the case as commodity price insurance becomes more widely adopted and insurance companies need to offset their risk.

The focus of this paper is—given a non-tradable commodity, on which we will write a given portfolio of options, and a single commodity futures—to determine the optimal hedge ratio and frequency of hedging given certain commercial constraints, such as transaction cost. We will price the options using the Black-Scholes model, and then calculating the hedge ratio using methods based on: delta, min-variance, VaR and C-VaR. We will evaluate the performance using hedging effectiveness, ECDF and KDE profiles and the distribution statistics.

We will draw conclusions about which strategies might be the most appropriate under certain market conditions, and note what needs to be further investigated and the next steps.

6.2 Literature Review

Academic research into hedging methods and strategies in markets has seen substantial attention over the past decades. Earlier studies predominantly concentrated on the motivation for hedging and hedging techniques in complete markets, utilizing exchange-traded instruments. In contrast, recent research, particularly in more specialised markets such as agricultural commodities, delves into cross-hedging techniques, insurance, and other risk management tools. As the demand for cross-hedging non-exchange-traded commodities grows, this becomes a key research area, particularly given the practical challenges associated with the lack of available direct hedging instruments. As discussed in the introduction, this problem is relevant to a number of commercial players, including insurance companies, producers and consumers (both retail and food service businesses). In this literature review, we hope to give a broad overview of the risk management strategies employed in agricultural commodity markets, as well as, the historical development of hedging strategies in broader markets. We examine how researchers have handled basis risk, the development of the minimum variance and related techniques, the value-at-risk approach and we introduce hedging effectiveness. We also introduce the use of the Black-Scholes model as a pricing and hedging benchmark.

In the introduction, we highlighted agricultural insurance as one form of risk management for producers and consumers, with the caveat that these insurance companies must then manage the risk that they onboard. Assa and Wang outline in their 2020 paper [7] three main categories of agricultural insurance: crop insurances, revenue insurances and derivatives. These insurances form the basis of risk management for participants with physical risk in this industry. Whilst these

insurances have utility for certain types of risk that participants may be facing, such as crop insurance providing financial support when the harvest is damaged or revenue insurance guaranteeing a minimum income, they also have a number of disadvantages. These are discussed extensively by Goodwin, in 'Problems with market insurance in agriculture' [8], where he outlines issues such as moral hazard, adverse selection and high cost due to the degree of systemic risk inherent in the crop insurance market. In this paper, we examine options written on beef commodities, however, this can easily be generalised to other commodities, and the discussion of crop insurance is an important aspect of risk management in the agriculture industry. Goodwin and Mishra, in their 2003 paper [9], outline the difficulties of revenue insurance and basis risk, suggesting that the basis risk causes the payouts to not match the buyers' actual exposure resulting in less future participation, and even with government subsidisation of these insurance programs, participants may struggle to find a plan that covers their risks adequately at an affordable price.

Assa and Wang [7] suggest that one solution to this could be price index insurances on agricultural goods; these insurances would be sold in the form of options written on these price indexes. In their paper they outline the exact optimal structure of these option policies. Whilst purchasing these policies may be a good solution for participants with physical risk, particularly if the price index is a non-exchanged traded commodity, since in that case the participant can themselves choose a contract which is directly related to the price risk they are facing, it still leaves the option provider with a significant risk exposure. Cross-hedging these options with futures can provide a solution. The options together with the hedging strategies that we test could enable the majority of the physical risk to be offloaded.

Other techniques, such as self-insurance and marketing flexibility are outlined by the Kansas State University Department of Agricultural Economics [10]. Typically self-insurance works by income diversification and marketing arrangements with breeders, processors etc., and can also help with risk management, but these techniques have limited benefit, particularly in more extreme scenarios. Schoeder and Yang [11] in their 2001 paper show that live cattle futures do not offer much opportunity for effective hedging of wholesale beef cuts, particularly because the correlation is not strong enough, and suggest that a Choice-to-Select spread futures contract with a boxed beef futures contract would increase the opportunity for hedging. Similarly in a 2010 paper, Bieroth [12] suggests that changes in the way that beef is marketed has led to poor performance in cross-hedging. They also highlight the difficulty of basis risk and test the impact of bundling, but find that it does not reduce basis risk. A more recent paper looking at cross-hedging in the beef industry was published by Alcoforado et al. [13] in Brazil where they were able to use cattle spot prices to estimate futures prices using a GARMA model, which suggests that the basis in that case could be modelled. Nonetheless, the majority of the available research suggests that cross-hedging effectively in these markets is still a challenge.

Another large market is pork, and lean hog futures have been studied as a possible hedging vehicle. For instance, Ditsch and Leuthold [14] evaluate the usage of lean hog futures and others to hedge cash live hogs and cash meats using a number of

different hedging techniques including the minimum-variance one that we study here. Overall, they found that neither contract that they tested showed significant hedging opportunities for pork trimmings or hams implying that there is more research to be done here as well.

We now outline the historical background and basis for the option pricing and hedging strategies which we use to model and evaluate. The Black-Scholes model, developed by Fischer Black and Myron Scholes in 1973, provides a mathematical framework that is often used for option pricing and hedging. This model was the first widely adopted option pricing model and remains a cornerstone of the industry for many reasons, despite a number of significant assumptions and limitations underlying the model. Firstly, it highlights the core variables, such as stock price, volatility, time-to-maturity and interest rates, with many more advanced models, with more realistic assumptions, still building off the Black-Scholes framework. It also provides a lower bound on prices, with the assumption of perfectly liquid markets, no arbitrage opportunities and negligible transaction costs reflecting a lower premium than would typically be required. The assumptions of the model also enable closed-form solutions for hedge ratios and option prices. In the Black-Scholes model section, we explain how the model works and include the relevant equations. Since the focus of our paper is not the option pricing, but testing the different hedging techniques, we simply use the Black-Scholes model for the option pricing.

In agricultural commodities a number of more advanced approaches have been proposed, particularly when using derivatives in place of insurance, for instance, Assa in his 2015 [15] and 2016 [16] papers determines the dynamics of the derivative prices and proposes a financial engineering framework. This framework enables modelling of commodity prices and explains the relationship between key variables such as demand, volatility and the leverage effect of commodities, so that derivative prices can be more accurately determined. Furthermore, Ye et al. [17] developed an improved Black-Scholes model for the calculation of crop price insurance premiums.

Prior to the publication of Black-Scholes and other models, Johnson [18] in 1960 developed the foundational minimum-variance approach to hedging. In his paper he considers a cash position coupled with a short hedge using futures contracts, and the basis risk is between the cash and futures prices. He showed that this is optimal for any investor seeking to minimise risk, regardless of expected returns. This minimum-variance approach can be combined with the Black-Scholes model, and in particular, using the Black-Scholes delta to model the change in option value with respect to the change in the underlying asset's price, to generate a minimum-variance cross-hedging ratio which accounts for the fact that we are holding a position in the option, rather than just the underlying asset. In 1984 Frechette [19] outlined the commercial problem that in the agricultural insurance basis risk can be a significant barrier to the implementation of hedging techniques commercially, particularly spatial basis risk. He highlights previous studies [20–23] which also discuss the fact that including hedging costs can change optimal hedge ratios and

influence whether participants choose to hedge. Whilst we included a transaction cost in our study, further sensitivity analysis could be done.

In the 1990s many of these ideas were combined to form blended hedging models. A number of approaches were introduced including lower partial moments, mean-Gini, and stochastic dominance, to name a few, alongside more theoretical models, such as, time-varying hedge ratios estimated using GARCH and stochastic volatility models. These approaches were all reviewed in the 2002 paper by Lien and Tse [24], who provided a comprehensive overview of the state of research on this topic. In their paper they conclude that minimum variance is a useful benchmark, but models such as VaR, mean-Gini and LPM can better address tail-risk. They also concluded that stochastic models have mixed empirical performance despite the theoretical appeal. In our study we focus on testing tail-risk based models with historical realised volatility (see implied and realised volatility section for more detail).

In 2003 Harris and Shen introduce a value-at-risk-based hedging approach which estimates VaR from historical simulation and GARCH and we consider a similar approach in our empirical study. They demonstrate that minimising variance can reduce the standard deviation of portfolio returns, but may increase skewness and kurtosis, whereas minimising VaR or C-VaR instead can reduce extremes.

Contrastingly, Hung et al. [25] developed a parametric approach to estimate the minimum VaR hedging ratios. This method relies on the assumption that if the returns of the index and the futures are normally distributed, the VaR of the hedged portfolio can be expressed as the difference between the potential loss due to market volatility and the expected return over a specific time horizon. While this method has the advantage of simplicity and easy interpretability, it heavily depends on the normality assumption of the returns, which may not hold true, particularly during extreme market conditions, and in the case of agricultural commodities where we see factors such as weather, supply-chain shocks and geopolitical events, changing the distribution. To compare the effectiveness of the non-parametric method and parametric approach, Cao et al. [26] developed a semi-parametric approach based on the Cornish and Fisher expansion. This method adjusts the appropriate quantile of the standard normal distribution using the higher moments such as skewness and kurtosis, thereby approximating the quantile of the probability distribution. By doing so, they express the VaR of the hedged portfolio in terms of the multiplication of the volatility of the hedged portfolio and the Cornish and Fisher expansion. This approach has shown significant improvements in VaR reduction when comparing to Harris and Shen's Kernel method. However, this method requires the portfolio return to be drawn from location-scale family distribution and have a mean of zero, which is a restrictive assumption and may limit its applicability in certain scenarios. Considering the limitations and assumptions associated with both parametric and semi-parametric methods, this study adopts the non-parametric approach to determine the optimal minimum VaR hedging ratio. While the non-parametric method may avoid the constraints imposed by distributional assumptions, it is essential to acknowledge its susceptibility to extreme quantiles, which could impact the hedging effectiveness under certain market conditions. In our study, we also use the non-

parametric historical simulation approach to determine an optimal minimum C-VaR hedging ratio, to compare to the VaR-based ratio.

Cong, Tan and Weng examine an optimal partial hedging strategy for both VaR [27] and C-VaR [28]. They find that the optimal strategy is either a knock-out call or call spread when VaR is the risk measure. When C-VaR is the risk measure they find the optimal strategy is the bull call spread hedging. In our paper we study a portfolio of vanilla call and put options, but it would be an interesting extension to test these methods on a portfolio of options with the structure outlined by Cong, Tan and Weng.

In order to evaluate and compare these strategies, we also introduce literature on the hedging effectiveness, in particular we use the measure introduced by Heifner and Ederington [29] and discuss further variations of this in the hedging effectiveness section.

6.3 Problem Definition and Method

6.3.1 Basis Risk

Basis risk is a key consideration when developing hedging strategies in the context of agricultural commodity markets. We define the basis as the price difference between the non-exchange-traded commodity and the corresponding futures price contract on the exchange for the given futures.

$$B_t = S_t - F_t \tag{6.1}$$

where, at a given time t:

- B_t is the basis
- S_t is the spot price of the non-exchange-traded commodity
- F_t is the future price on the exchange

The difference between these two prices can be significant and can vary greatly over time, making the problem of managing the basis risk crucial for those trading in the commodity space and looking to hedge price risk via futures. It is important to ensure that the futures positions are offsetting the spot price movements. In our study we are trying to offset the change in value of the option that we write on the non-exchange-traded commodity with the futures. Several factors drive this difference between the spot price and futures price, in particular, the movement and storage of physical goods can have a significant impact on the price difference. When there are supply-demand mismatches, or supply-chain disruption, the magnitude of volatility can be different in each of these markets, and this contributes to the unpredictability of the basis over time. Timing mismatches between needing to transact physical positions vs future expirations can further exacerbate the basis. Quality variation

between the non-exchanged-traded commodity and futures contract specification can also cause differences in price.

6.3.2 Black-Scholes Model

As introduced previously, the Black-Scholes model [30], developed by Fischer Black and Myron Scholes in 1973, provides the mathematical framework that we will use for option pricing and as a benchmark for comparison when determining the hedge ratio.

The Black-Scholes model considers a "risky" asset (such as a stock, or in our case a commodity price index), a "riskless asset" (e.g. cash), and an option written on the risky asset, whose value is to be determined. It is assumed that the risky asset follows Brownian motion, the volatility is constant, the future asset price at any point in time is lognormally distributed and that it pays no dividends. The riskless asset is considered an investment alternative to the risky asset or a source of financing; for instance, in the case of cash, you can receive it when selling the option or stock and deposit it to earn interest or you can borrow it to buy the option or stock and pay interest. The interest is the risk-free interest rate, and it is assumed in the Black-Scholes model that this is constant and known in advance. The option is assumed to be European, which is defined such that it can only be exercised at expiration. The model also assumes the following regarding the market: there are zero transaction costs, perfect liquidity, no arbitrage, no restrictions on trading (e.g. no restrictions on short selling), and securities are infinitely divisible.

Commodity prices tend to have skewed, fat-tailed distributions [31], due to: weather sensitivity, extreme events such as droughts, floods, heat waves etc. can cause supply shocks which result in price spikes; low storage time; seasonality; new-crop old-crop jumps; and government policies/subsidies. This causes our prices to violate the lognormality assumption and typically these challenges also tend to result in changing volatility. Additionally, whilst commodity markets have evolved towards having more liquidity they tend to be much less liquid than the stock market environment that Black-Scholes was developed for. Furthermore, the costs of transporting, storing and trading agricultural commodities can be very large, particularly when there are significant supply/demand imbalances.

Despite these shortcomings, the Black-Scholes model still holds utility as a first approximation to our option price and the delta as a simple guideline to an initial hedging strategy.

Black-Scholes Option Price The value of a call option, V_c, and a put option, V_p, are given by the following formulae;

$$V_c = S_0\Phi(d_1) - K\mathrm{e}^{-rT}\Phi(d_2) \tag{6.2}$$

$$V_p = K\mathrm{e}^{-rT}\Phi(-d_2) - S_0\Phi(-d_1), \tag{6.3}$$

where

$$d_1 = \frac{\log(S_0/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}, d_2 = \frac{\log(S_0/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}. \tag{6.4}$$

Where

- S_0 is the initial price
- K is the strike price
- r is the risk-free interest rate
- T is the time-to-maturity
- σ is the realised historical volatility
- Φ is the CDF of the normal distribution

Here we also note that the payoff for a European Call option is given as: $P_C = max(S_t - K, 0)$ and a European Put option is given as: $P_p = max(K - S_t, 0)$.

The quantity that we need for hedging is the Delta which is defined as below.

Delta, Δ, is a measure of the rate of change of the options calculated value, V, with respect to the change of the underlying assets price, S.

$$\Delta = \frac{\partial V}{\partial S} \tag{6.5}$$

In the Black-Scholes model, for a European Call or Put option it is calculated as shown:

$$\Delta_{call} = \mathrm{e}^{-rT}\Phi(d_1) \tag{6.6}$$

$$\Delta_{put} = \mathrm{e}^{-rT}(\Phi(d_1) - 1) \tag{6.7}$$

The traditional hedging strategy using the delta is known as 'Delta-neutral hedging', where the option is hedged with the underlying assets and is based on the replication of the option, where the logic is the same as option pricing using the binomial tree model.

6.3.3 Implied and Realised Volatility

Implied volatility plays an important role in the Black-Scholes option pricing model, which assumes constant volatility over the life of the option. However, calculating implied volatility in agricultural commodity markets presents challenges, especially when derivative markets are absent for certain commodities.

Estimating implied volatility (IV) using the Black-Scholes model is an iterative process that involves adjusting the volatility parameter until the calculated option price matches the market price, with the underlying assumption that there is constant volatility over the life of the option. This process, often executed through numerical methods like the Newton-Raphson method, yields an implied volatility surface or curve, providing insights into market expectations regarding future price fluctuations. However, the Black-Scholes model has inherent limitations, especially when faced with the complex dynamics of agricultural markets, such as weather-related shocks or geopolitical developments. In scenarios where no derivative markets exist, determining implied volatility is difficult because the Black-Scholes model relies on option prices, and without an options market, there is a lack of observable data to infer market expectations about future price movements. This absence hinders the accurate calculation of implied volatility, a crucial input for the model.

The limitations of the Black-Scholes model have spurred the development of more advanced volatility models. The SABR model [32], for instance, introduces stochastic volatility, allowing for greater flexibility in capturing the dynamics of the implied volatility surface. Similarly, the Heston model [33] incorporates stochastic volatility and the correlation between asset prices and volatility, making it relevant for situations where volatility exhibits mean-reverting behaviour. The Chen model [34] and the GARCH model [35, 36], while not directly option pricing models, also contribute to the discussion by addressing the limitations of the Black-Scholes model through the introduction of features such as jump diffusion processes and time-varying volatility. In some cases, these models may be used to estimate the volatility input for option pricing models, bridging the gap between historical data and market expectations.

Even in cases where derivative markets for agricultural commodities do exist, calculating implied volatility can be challenging. Issues such as liquidity constraints may result in sparse trading of options, limiting the number of data points available to gauge market sentiment. Additionally, the inherent complexities of agricultural commodities, influenced by factors like weather conditions, supply chain shocks and geopolitical events, make isolating the impact of these variables on implied volatility challenging. Therefore, realised volatility becomes a relevant method to also include, serving as an alternative when calculating implied volatility is challenging. While implied volatility captures market expectations embedded in option prices, realised volatility is derived from historical price movements. In situations where derivative markets are absent or illiquid, as is often the case with certain agricultural commodities, relying on realised volatility is the pragmatic choice. Realised volatility provides a tangible measure of historical market dynamics, offering valuable insights into how prices have behaved in the past. This historical context becomes particularly relevant when dealing with agricultural commodities, where the impact of external factors can be significant and difficult to quantify using traditional models.

The choice between implied volatility and realised volatility is context-dependent. Given that our focus of our research in this paper is not the volatility calculation, we use the historical realised volatility given below.

Let $R_t = \frac{S_t}{S_{t-1}}$, then

$$\sigma = \log \sqrt{\frac{1}{N-1} \sum_{t=1}^{N} (R_t - \overline{R})^2} \tag{6.8}$$

This is typically annualised by multiplying the square root of the number of time periods in a year.

6.3.4 Static and Dynamic Hedging

Hedging can be implemented once at the time that the option position is opened, and once at the time that the option position is closed to simultaneously close the hedging position. This is typically referred to as static hedging. This has the advantage of being straightforward to implement and only a single calculation of the hedging ratio is required. In many cases, however, particularly when there is uncertainty about the liquidity of the market at future times or the option maturity is a long time period compared with the frequency of the asset used for hedging, it may be useful to more frequently update the hedging position and this is typically referred to as dynamic hedging. Typically for dynamic hedging we would have a greater transaction cost overall, and it is important to compare the advantage of the increased flexibility in updating our futures position and the disadvantage of an increased transaction cost. In particular, our hedging effectiveness may only increase by a small amount with dynamic hedging, and in this case, it may be more profitable to take the static approach. It may also be difficult, particularly in a market, such as agricultural commodities, to obtain data that is a sufficiently high frequency with reliability, to model and perform dynamic hedging. In our empirical study, we test static and dynamic cross-hedging for the Black-Scholes approach, min-variance and value-at-risk.

6.3.5 Variance-based Hedging Strategies

Minimum-Variance Cross-Hedging Ratio In this section, we show a derivation of the optimal cross-hedge ratio formulae mathematically. This was originally derived in Ederington's 1979 paper [37]. The following notation is used:

- ΔX: the change of index spot price in the hedging period
- ΔY: the change of future contract price in the hedging period

- h: the optimal hedging ratio
- σ_X: the standard deviation of ΔX
- σ_Y: the standard deviation of ΔY
- ρ: the correlation coefficient
- $\Delta\pi$: the change in value of the hedger's portfolio

$$\Delta\pi = \Delta X - h\Delta Y \tag{6.9}$$

$$Var(\Delta\pi) = \sigma_X^2 + h^2\sigma_Y^2 - 2h\rho\sigma_X\sigma_Y \tag{6.10}$$

Then we want to minimise the variance with respect to h, so we calculate the derivative with respect to h, equate it to zero and solve for h.

$$dVar(\Delta\pi) = 2h\sigma_Y^2 - 2\rho\sigma_X\sigma_Y = 0 \tag{6.11}$$

$$h = \rho\frac{\sigma_X}{\sigma_Y} \tag{6.12}$$

We can see that this is indeed a minima by examining the second derivative and seeing that it is greater than zero.

$$d^2Var(\Delta\pi) = 2\sigma_Y^2 > 0 \tag{6.13}$$

We can improve this formula, with the additional assumption that X and Y follow geometric Brownian motions with drifts μ_X, μ_Y and volatilities σ_X, σ_Y. The increments dW_1 and dW_2 have correlation ρ. This assumption gives us the following set of equations:

$$dX = \mu_X X dt + \sigma_X X dW_1 \tag{6.14}$$

$$dY = \mu_Y Y dt + \sigma_Y Y dW_2 \tag{6.15}$$

$$\mathbf{E}[dW_1 dW_2] = \rho dt \tag{6.16}$$

Now, we consider a portfolio long in Y and short an option on X, where h is the hedge ratio we are trying to find, i.e. the amount of Y that should be in this portfolio to minimise the variance:

$$\Pi = hY - V(X, t) \tag{6.17}$$

where the change $d\Pi$ is partly due to dV and dY. Using Ito's lemma:

$$d\Pi = hdY - dV \tag{6.18}$$

$$dV = \frac{\partial V}{\partial t}dt + \frac{\partial V}{\partial X}dX + \frac{1}{2}\sigma_X^2 X^2 \frac{\partial^2 V}{\partial X^2}dt \tag{6.19}$$

Taking the variance:

$$\text{var}(d\Pi) = \text{var}(hdY - dV_t) \tag{6.20}$$

We can use the identity, the variance of sum formula:

$$\text{var}(X_1 - X_2) = \text{var}(X_1) - \text{var}(X_2) - 2\text{cov}(X_1, X_2) \tag{6.21}$$

This gives:

$$\text{var}(hdY - dV_t) = h^2\text{var}(dY_t) - \text{var}(dV_t) - 2h\text{cov}(dY_t, dV_t) \tag{6.22}$$

Taking the first derivative with respect to h (so, $\frac{d}{dh}\text{var}(dV_t) = 0$) and minimising through using the first order condition, we obtain:

$$h_{Opt} = \frac{\text{cov}(dV_t, dY_t)}{\text{var}(dY_t)} \tag{6.23}$$

To solve this we substitute in the expressions for dY and dV so:

$$\begin{aligned}\text{cov}(dV_t, dY_t) = \text{cov}(&\frac{\partial V}{\partial t}dt + \frac{\partial V}{\partial X}X\mu_X dt + \frac{\partial V}{\partial X}\sigma_X X dW_1 \\ &+ \frac{1}{2}\sigma_X^2 X^2 \frac{\partial^2 V}{\partial X^2}dt, \mu_Y Y dt + \sigma_Y Y dW_2)\end{aligned} \tag{6.24}$$

The only terms which contribute here are $\frac{\partial V}{\partial X}X\sigma_X dW_1$ and $\sigma_Y Y dW_2$, as the covariance of dt with the other terms gives 0, and since $\mathbb{E}[dW_1, dW_2] = \rho dt$, $\text{cov}(\frac{\partial V}{\partial X}X\sigma_X dW_1, \sigma_Y Y dW_2) = (\frac{\partial V}{\partial X}X\sigma_X\sigma_Y Y)\rho dt$. Combining this with $\text{var}(dY_t) = Y^2\sigma_Y^2 dt$, we obtain:

$$h_{Opt} = \rho\frac{X\sigma_X}{Y\sigma_Y}\Delta_{BS} \tag{6.25}$$

where $\Delta_{BS} = \frac{\partial V}{\partial X}$ and is the Black-Scholes delta as referenced previously.

6.3.6 Value-at-Risk-based Hedging Strategies

In this section, we consider Value-at-Risk based hedging strategies. Value-at-Risk (VaR) is a statistical measure which is used to quantify the level of financial risk

of a portfolio over a specific timeframe. It estimates the maximum loss within the time horizon at the given probability and conceptually represents the quantile of the loss distribution. This is typically calculated using past market data to model the distribution of returns, commonly this is done using either: parametric, historical simulation or Monte Carlo methods.

Let F be the CDF of X where X is the profit and loss distribution over the time period. Let $\alpha \in (0, 1)$ be the given confidence level, then,

$$VaR_\alpha(X) = -inf\{x \in \mathbb{R} | F_X(x) > \alpha\} \tag{6.26}$$

The choice of distribution for calculating VaR (Gaussian, Student's t etc.) can have a large impact on the VaR calculation. Since for our methods the VaR is calculated non-parametrically we do not need to be concerned about the choice of the distribution, but it may be worthwhile considering using a parametric method in future research, as in the non-parametric case there are very few data points.

To implement this strategy, we estimate the minimum-VaR hedge ratio by choosing an arbitrary hedge ratio, calculating hedge portfolio returns over a rolling window of historical data (this is a historical simulation approach) and estimating the VaR of this resulting hedged portfolio. Then a numerical optimization procedure is used in order to estimate the value of the hedge ratio that minimizes the hedge-portfolio VaR.

Instead of simply minimising the potential loss of the portfolio over a specified time, we may want to focus more closely on the downside tail risk, beyond the VaR threshold, and for this we consider the C-VaR (Conditional Value-at-Risk), which is given by the following formula:

$$CVaR_\alpha(X) = \mathbf{E}[X | X > VaR_\alpha] \tag{6.27}$$

where VaR_α is as above.

The C-VaR gives the average loss in the tail beyond the specified percentile of the loss distribution given by the VaR, and gives information about the magnitude of losses in the event of extreme market moves beyond the VaR. The implementation of C-VaR is the same as VaR, except minimising the C-VaR risk measure instead of the VaR.

6.3.7 Defining Hedging Performance

6.3.7.1 Hedging Efficiency

To measure the hedging strategy's performance we follow the guidance of Heifner and Ederington [29] in their 1983 paper, which outlined hedging effectiveness as

$$1 - \frac{RiskMeasure(hedged\,PnL)}{RiskMeasure(unhedged\,PnL)} \tag{6.28}$$

where the choice of risk measure is fundamental in capturing what we want to quantify in terms of the risk of PnL. This metric measures the proportion of variations in the unhedged position's PnL that is hedged off by the introduction of hedging positions. If the effectiveness of the hedging strategy is 1 then the hedging position eliminates the risk exposures completely. Otherwise, the smaller the hedging effectiveness the worse the strategy is.

There are a number of risk measures available, for example: standard deviation, downside deviation, value-at-risk and expected shortfall. The standard deviation examines how spread out the values are, ignoring profit and loss, but simply determining whether the variation of the PnL is decreased by hedging. The downside deviation examines the standard deviation of negative PnLs. Value-at-risk measures the tail loss and is the loss level that will not be exceeded with a certain confidence level during a certain period of time. Alternatively, if the extreme loss is large but has a low probability, VaR may underestimate tail risk, so instead we can look at expected shortfall, which gives the expected loss in the worst cases.

In our case we use value-at-risk (95%) and standard deviation as our risk measures, as we feel that these give a clear indication of some of the key ways that the distribution may change as a result of hedging, however, an individual business may choose different measures depending on their key performance metrics.

Using a metric such as hedging efficiency, which only considers a single distribution statistic, can be one-dimensional, particularly when we want to compare the performance across the entire distribution. For this we use ECDFs (Empirical Cumulative Density Function) and KDEs (Kernel Density Estimation). These plots provide a visual tool for examining the shape and characteristics of the PnL distributions.

6.3.7.2 Empirical Cumulative Density Function and Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric method used to estimate the probability density function (PDF) of a continuous random variable based on a set of observed data points. It provides a smoothed representation of the underlying distribution, giving insights into the shape and characteristics of the data. At each data point we place a kernel and sum these kernels to create a continuous estimate of the PDF.

The KDE at a specific point x is calculated using the following formula:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{6.29}$$

where:

- $\hat{f}(x)$ is the estimated PDF at point x
- n is the number of data points

- h is the bandwidth, controlling the width of the kernel
- x_i represents each observed data point
- $K(\cdot)$ is the kernel function

The kernel used is the Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}u^2}$ The Empirical Cumulative Distribution Function (ECDF) is a step function that represents the cumulative probability distribution of a set of observed data points and it is also a non-parametric method.

Let's consider a dataset of n observations $x_1, x_2, \ldots, x_n$. The ECDF at a specific value x is calculated as the proportion of data points less than or equal to x. Mathematically, the ECDF is defined as follows:

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \tag{6.30}$$

where:

- $F(x)$ is the value of the ECDF at x
- n is the number of data points
- $I(x_i \leq x)$ is an indicator function that equals 1 if $x_i \leq x$ and 0 otherwise.

The ECDF is constructed by sorting the data in ascending order and assigning cumulative probabilities based on the number of data points less than or equal to each value. It starts at 0 for the smallest observation and ends at 1 for the largest observation.

For a given x, the ECDF value represents the proportion of data points less than or equal to x. This step function provides a visual depiction of how the data is spread across its range and is useful for assessing percentiles of the data.

Mathematically, the ECDF and KDE encode the same information—the distribution—however, we choose to include both as it is easier to see different aspects of the distribution from the two plots. In particular we look to the ECDF for a clear comparison of the percentiles and the KDE for the mean, variance and skew.

These visual representations enable us to easily see how the distributions of the different techniques compare with one another. However, the downside of visually examining the ECDFs and KDEs is that this is does not provide us with a concrete quantitative way of comparing the distributions. For this we use the statistics of the distributions such as mean, variance, skew and kurtosis.

6.3.7.3 Distribution Characteristics

The mean, variance, skewness and kurtosis of the distribution are statistical measures that quantify important characteristics of the distribution.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p(x,t) dx \tag{6.31}$$

where $p(x, t)$ is the probability density of the random variable.

$$\mu = \mathbb{E}[X] \tag{6.32}$$

$$\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X] \tag{6.33}$$

$$skew = \mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^3] \tag{6.34}$$

$$kurtosis = \mathbb{E}[\left(\frac{X-\mu}{\sigma}\right)^4] \tag{6.35}$$

Here, X is a random variable, $E[\cdot]$ denotes the expected value.

- Mean: Represents the average or expected value.
- Variance: This measures the spread or dispersion of the distribution, we can also examine the square root, i.e. the standard deviation.
- Skewness: The skewness indicates the asymmetry in the distribution, a skewness of zero suggests a symmetric distribution.
- Kurtosis: This helps to measure the shape of the tails of the distribution.

For comparing the hedging strategy performances the characteristics can help to discriminate different properties of interest. For example, differences in skewness or kurtosis might indicate variations in the risk profiles of different strategies. By calculating and comparing these characteristics, we can gain a more nuanced understanding of the shape and characteristics of PnL distributions, offering insights into the comparative performance of various hedging strategies.

6.4 Data and Portfolio

6.4.1 Indices and Futures

6.4.1.1 Description

In this study, we aim to hedge: Beef Index (all beef type, steers and heifers, daily direct slaughter cattle); Cattle Index (all steers and heifers, total all grades, domestic, weighted average price); Beef Trim Index (chemical lean, fresh, national 50%) using the live cattle futures (Fig. 6.1).

The index prices are obtained from USDA, and are available on a daily basis for Beef Index and Cattle Index and a weekly basis for Beef Trim Index. The methodology for how these prices are collected can be found on the USDA website [38]. The future prices are obtained from Bloomberg, and are available on a daily basis (Tables 6.1 and 6.2).

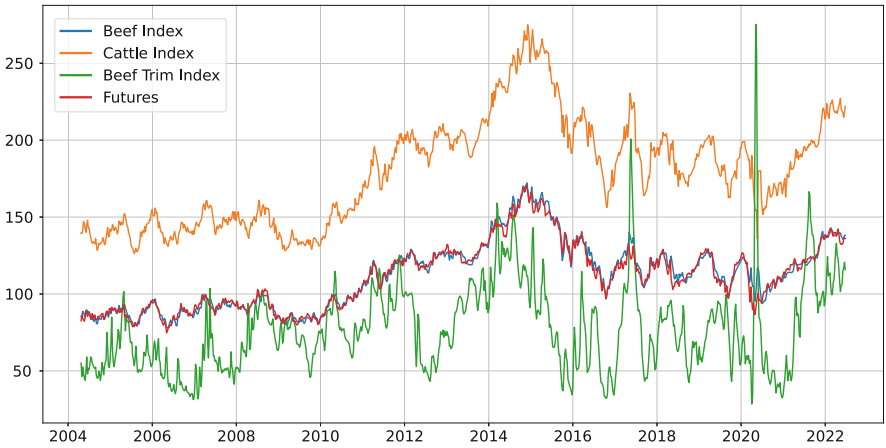


Fig. 6.1 Beef index, cattle index, beef trim index, futures

Table 6.1 Basic data information and symbols

Asset name	Description	Data availability range
Futures	Live cattle futures	2000-01-03 ∼ 2023-07-18
Beef index	All beef type steers and heifers daily direct slaughter cattle negotiated purchases, live FOB	2002-11-22 ∼ 2022-06-24
Cattle index	All steers and heifers, total all grades domestic, weighted avg price	2004-04-16 ∼ 2022-07-01
Beef trim index	Chemical lean, fresh 50%, National	2003-01-03 ∼ 2023-06-30

Table 6.2 Summary statistics for the returns of indices and future returns

Security	Mean	Standard deviation	Skewness	Kurtosis	Correlation with futures
Futures	0.0042	0.0415	−0.6307	1.3059	–
Beef index	0.0009	0.0224	0.0155	5.5722	0.785
Cattle index	0.0008	0.0245	2.25	34.7158	0.545
Beef trim index	0.0077	0.1204	2.0338	13.543	0.258

6.4.1.2 Data Processing

We cut all of the index and futures data to the range 2004-06-30 to 2022-06-30. We converted these datasets to weekly, using the average weekly price for both the index and futures. For the futures, we use the continuous curve provided by Bloomberg, as opposed to working directly with the individual contracts. This continuous curve captures the price trends of the live cattle future contracts over time, and we treat the continuous curve as a synthetic price series for the futures. This provides a useful tool for the analysis, mitigating the need to select a tenor at each time step.

6.4.2 Portfolio Selection

6.4.2.1 Portfolio Structure

We consider the portfolio structures outlined below (Tables 6.3 and 6.4).

We assume that we are always selling and settling at the end of a month. The notation T+3 means that, if the settlement price is the latest price (i.e current month end price), the strike is calculated based on the price 3 months ago. Similarly for T+6 and T+9. All of the options are at-the-money, so the strike ratio is 1.

As an example, we have printed a sample of Beef Index data (showing only month end prices for simplicity) below and shown the calculation.

The option is sold on 2004-04-30. The strike is 86.07. The settlement price is 84.65 for T+3.

The payoff for European call and put options is given as respectively:

$$P_{call} = max(S_t - K, 0) \tag{6.36}$$

$$P_{put} = max(K - S_t, 0) \tag{6.37}$$

So, for the call option the payoff will be $max(84.65 - 86.07, 0) = 0$. For the put option it will be $max(86.07 - 84.65, 0) = 1.42$.

In the plots below we show the option PnLs (premium and payoff) for the portfolios as defined above for both the synthetic and real data (Figs. 6.2, 6.3, and 6.4).

In the following table we calculate the total lifetime PnL for each option structure, which helps to explain our results. Even when we use the hedging effectiveness ratio to control for the PnL of the unhedged underlying option, we still find that the structure of the underlying option payoffs influences the hedging results (Table 6.5).

We note that in general, puts tends to perform better than calls, across all indices. We also note that the Beef Trim has the best performance in both directions. Other than this, it is hard to discern any other trends.

Table 6.3 Portfolio of option types

Option type	Maturity	Moneyness
Call	T+1, T+3, T+6, T+9	At the money
Put	T+1, T+3, T+6, T+9	At the money

Table 6.4 Sample of beef index data

Dates	Beef index
2004-04-30	86.07
2004-05-31	85.05
2004-06-30	87.96
2004-07-31	84.65
2004-08-31	84.21

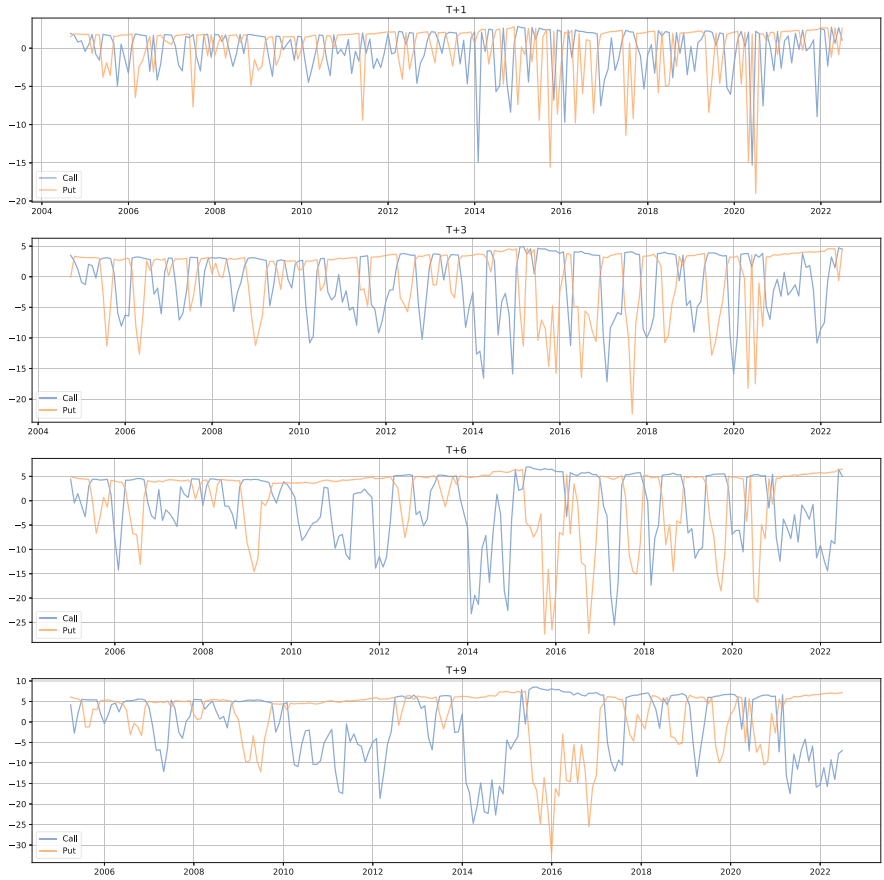


Fig. 6.2 Option PnLs beef index T+1,3,6,9

6.4.2.2 Portfolio Assumptions

We make the following assumptions regarding the practical setting for the study. We tested based on continuous selling of the portfolio of options every month, at the same volume. Additionally, we assume a transaction cost, TC, for the futures, which is calculated at a time t that is given by:

$$TC = 0.00015 * F_t * volume_t \tag{6.38}$$

where F_t is the future price at time t, and the volume at time t is typically given by the hedging ratio.

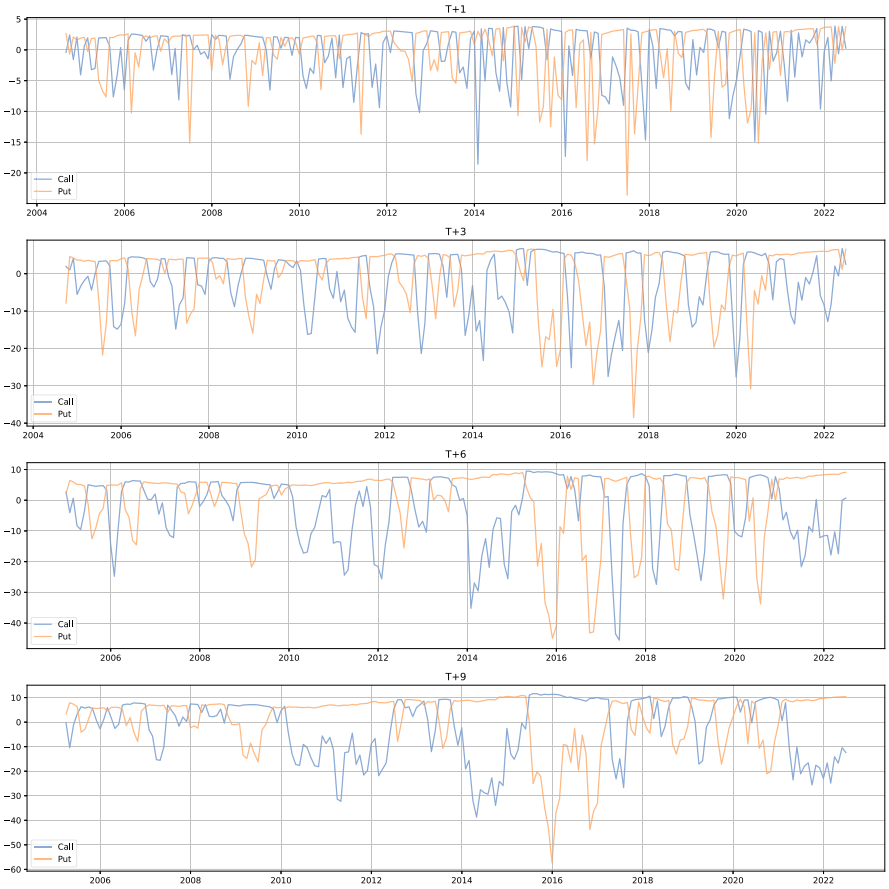


Fig. 6.3 Option PnLs cattle index T+1,3,6,9

6.5 Real Data: Results

6.5.1 Static Hedging

6.5.1.1 Hedging Effectiveness

We begin by presenting the hedging effectiveness results (Tables 6.6, 6.7, 6.8, 6.9, 6.10 and 6.11).

From these tables, we can note a number of trends for both standard deviation and value-at-risk. There is a clear trend in the performance of each of the methods across all of the indices: the delta method performs best, then min variance, then the VaR and C-VaR methods have similar performance, where which method is better depends on the particular option structure. Overall, these results are exactly as we

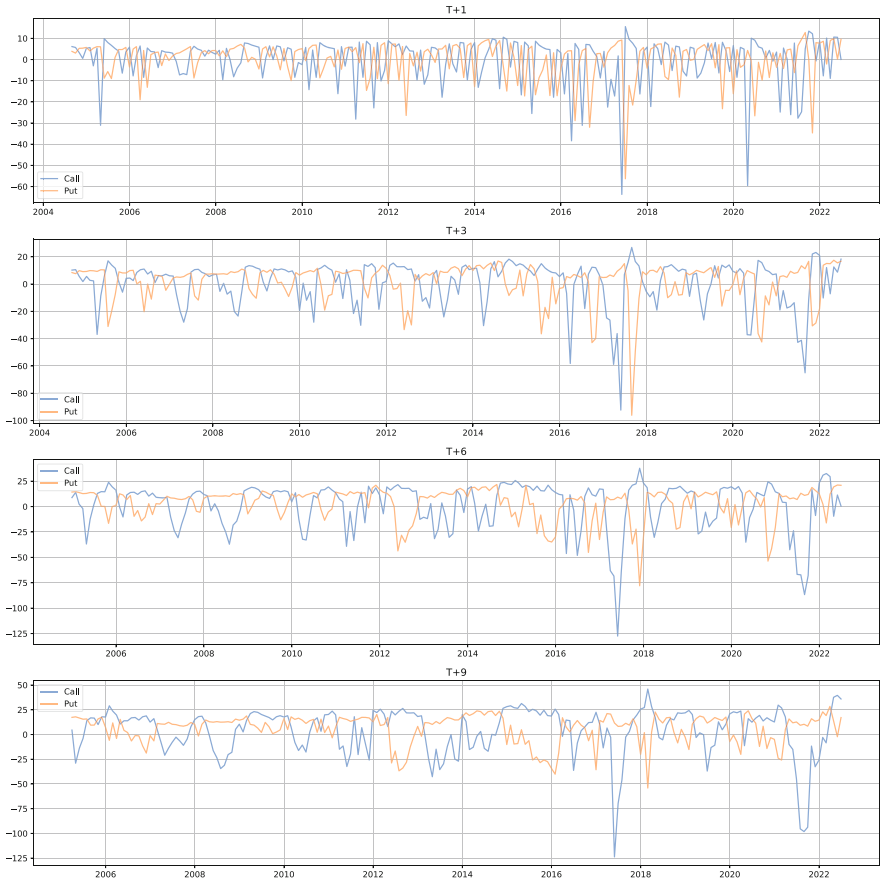


Fig. 6.4 Option PnLs beef trim index T+1,3,6,9

Table 6.5 Total PnL per option type

Index and option type	T+1	T+3	T+6	T+9
Beef index call	−20.9646	−125.450519	−166.665194	−96.890002
Beef index put	32.4546	32.297815	151.738139	354.323331
Cattle index call	−87.823801	−343.583559	−452.67942	−371.693973
Cattle index put	1.266199	−92.153559	50.69058	324.766027
Beef trim index call	−57.337704	223.368987	351.913833	867.158395
Beef trim index put	3.762296	403.348987	702.233833	1320.768395

would expect. The delta method hedges the option with the underlying asset and is provided as a benchmark of the performance with no basis risk, therefore we would expect this to have the best performance. The min variance uses the same delta but includes extra terms to adjust for correlation, volatility and price level differences between the index and the futures, so we would expect similar performance to the

Table 6.6 Hedging effectiveness for beef index risk measure: VaR-95%

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.3526	0.1745	0.0848	−0.1567
Call 3	0.4572	0.2406	0.1643	0.1888
Call 6	0.5464	0.3171	0.2395	0.3048
Call 9	0.6023	0.3504	0.0956	−0.0464
Put 1	0.5730	0.2868	0.3068	0.2363
Put 3	0.5004	0.2435	0.2363	0.2647
Put 6	0.5110	0.2641	−0.0792	0.2732
Put 9	0.5400	0.3785	0.1328	0.2236

Table 6.7 Hedging effectiveness for cattle index risk measure: VaR-95%

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.3938	−0.0173	−0.0156	−0.4344
Call 3	0.5707	0.1730	0.2038	0.1388
Call 6	0.4707	0.2615	0.1762	0.1913
Call 9	0.5522	0.1810	0.2230	0.0358
Put 1	0.5302	0.1285	0.1115	−0.0189
Put 3	0.5528	0.1530	0.2088	0.2411
Put 6	0.5660	0.2164	0.2627	0.2844
Put 9	0.5426	0.2150	0.2759	0.1639

Table 6.8 Hedging effectiveness for beef trim index risk measure: VaR-95%

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.5791	0.0552	−0.2276	−0.1735
Call 3	0.5144	0.1252	−0.1874	−0.3696
Call 6	0.5624	−0.0369	−0.2376	−0.2319
Call 9	0.6883	0.0369	−0.0338	−0.1303
Put 1	0.3934	−0.0888	−0.3572	−0.2598
Put 3	0.5202	0.0506	−0.0324	0.1792
Put 6	0.4892	0.0668	−0.0621	−0.3948
Put 9	0.6160	−0.0056	−0.3850	−0.3818

delta in terms of the trends that we observe, but a slightly worse overall performance, due to the basis risk which is introduced by using the futures as opposed to the index itself. In general, the VaR and C-VaR methods did not perform as well as the min variance method. Generally, (excluding the delta method for comparison since this is not cross-hedging) across SICs, Beef Index performs best, followed by Cattle Index, then Beef Trim Index, inline with what we would expect from the correlation.

The VaR and C-VaR methods rely on empirically determining the VaR and C-VaR, at a chosen confidence level, over a lookback window, the length of which is a hyperparameter. Using monthly data, there may be insufficient data points for an accurate empirical estimation of the distribution, so it is possible that using

Table 6.9 Hedging effectiveness for beef index risk measure: std

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.3825	0.0889	0.0276	−0.2955
Call 3	0.5215	0.2626	0.0892	−0.0073
Call 6	0.5023	0.3358	−0.1796	0.1798
Call 9	0.5373	0.3125	−0.2221	−0.2183
Put 1	0.4950	0.1746	0.0401	0.0442
Put 3	0.5423	0.2841	0.1950	0.2525
Put 6	0.5165	0.3030	−0.2628	0.1371
Put 9	0.4976	0.3167	−0.3406	0.0316

Table 6.10 Hedging effectiveness for cattle index risk measure: std

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.4528	0.0375	−0.0485	−0.3974
Call 3	0.5068	0.1535	0.0579	−0.1441
Call 6	0.4767	0.2054	−0.1131	−0.0702
Call 9	0.5190	0.2155	−0.1403	−0.1382
Put 1	0.5063	0.0893	−0.0555	−0.2360
Put 3	0.5154	0.1740	0.1056	0.0827
Put 6	0.5118	0.1807	−0.0299	0.0055
Put 9	0.5116	0.1978	−0.1387	0.0650

Table 6.11 Hedging effectiveness for beef trim index risk measure: std

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.5296	−0.0370	−0.1556	−0.2614
Call 3	0.4837	−0.0050	−0.1440	−0.3906
Call 6	0.5458	−0.0002	−0.3482	−0.1815
Call 9	0.5790	−0.0001	−0.2258	−0.3533
Put 1	0.4048	0.0254	−0.1771	−0.2597
Put 3	0.4028	0.0702	−0.1507	−0.1651
Put 6	0.3676	0.0462	−0.1812	−0.6069
Put 9	0.3965	0.0562	−0.3010	−0.4470

weekly frequency data, over the same period of time could yield better results, but an exploration of the impact of hedging frequency and hyperparameter choice was outside the scope of this study.

At the 54.5% log return correlation level, i.e. for the Cattle Index, we see that the min-variance, VaR and C-VaR methods have a positive impact in terms of risk reduction for most structures, but once the correlation is as low as 25.8% for the Beef Trim Index, we see that the min-variance method only has a small positive impact in terms of risk reduction and VaR and C-VaR increase the risk versus the unhedged portfolio.

It is also worth noting that across all of the hedging methods, there is no clear trend in hedging effectiveness performance vs maturity, which comes from the fact that hedging does not change the performance greatly across maturity relative to the unhedged portfolio. Hedging effectiveness whilst useful for measuring to see if certain objectives are met, is one-dimensional and for a full analysis a business may want to consider using hedging effectiveness under multiple risk measures as well as implementing other metrics such as risk-adjusted returns.

6.5.1.2 ECDF and KDE Plots

To further analyse our results and show that hedging effectiveness does not tell the whole story we present below the KDE and ECDF plots. This allows us to compare the distributions between the methods (Figs. 6.5, 6.6, 6.7, and 6.8).

These plots highlight that whilst there are clear differences in hedging effectiveness between the methods, the distributions have many similarities, and in particular, when we examine the CDFs, we see that no method is arguably significantly better than the unhedged portfolio. So, in contrast to our conclusion from the hedging effectiveness, these graphs suggest that more work may need to be done to determine an effective hedging strategy based on a range of measures (Figs. 6.9 and 6.10).

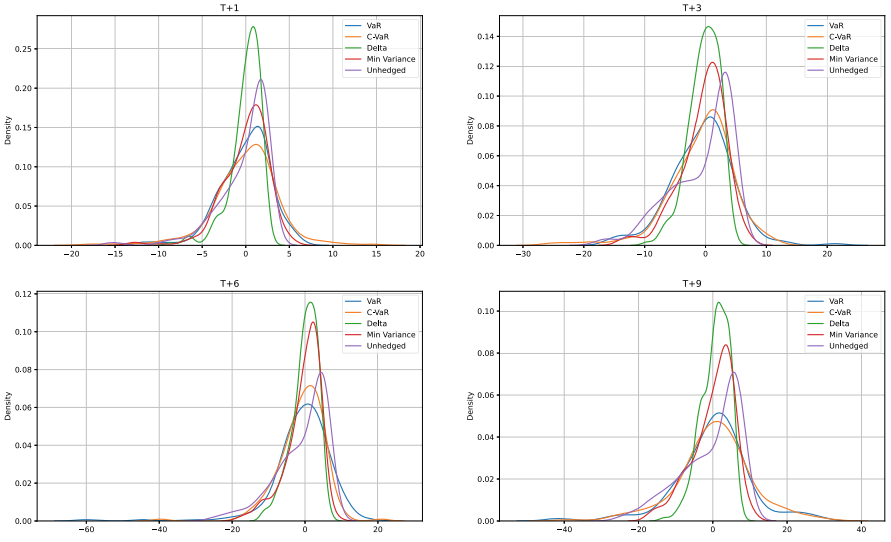


Fig. 6.5 Call option PnL KDE beef index

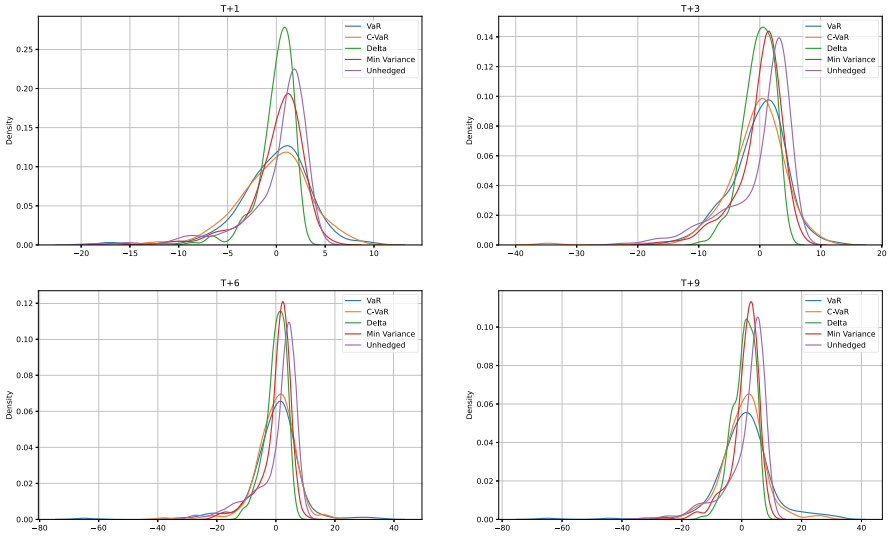


Fig. 6.6 Put option PnL KDE beef index

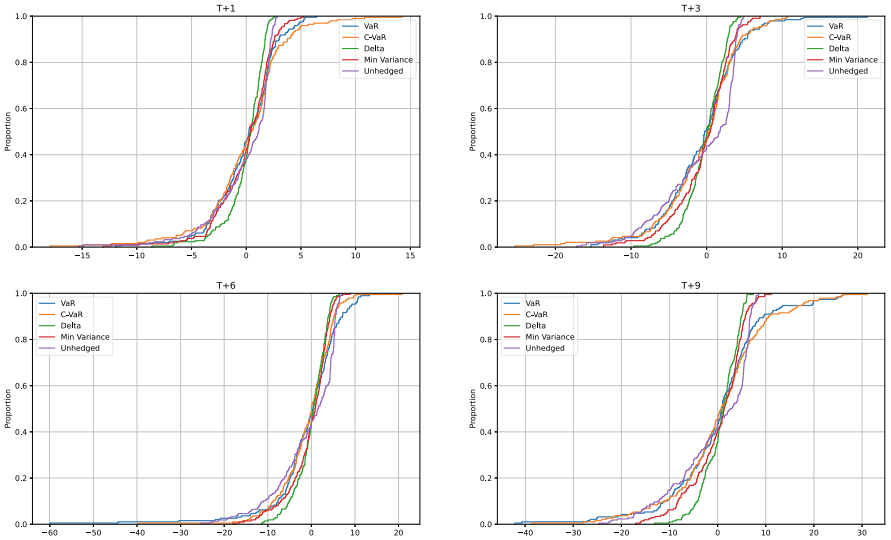


Fig. 6.7 Call option PnL empirical CDF beef index

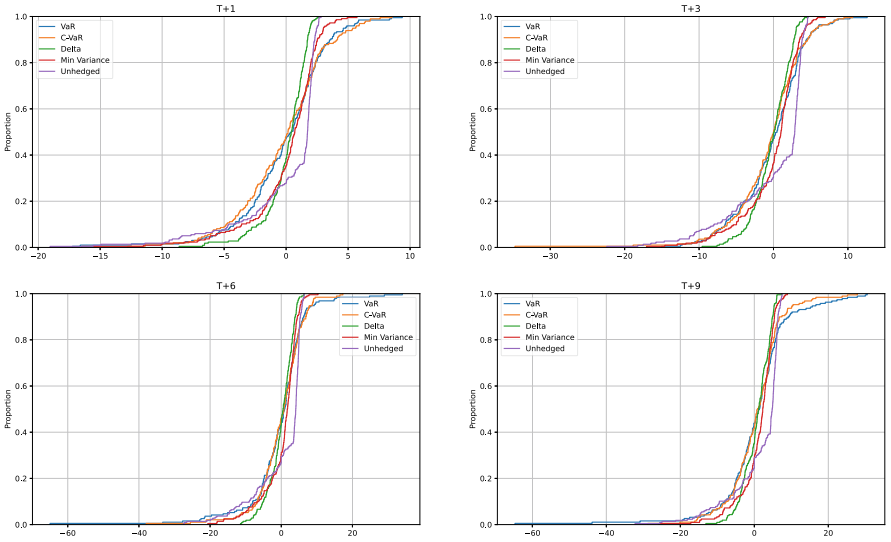


Fig. 6.8 Put option PnL empirical CDF beef index

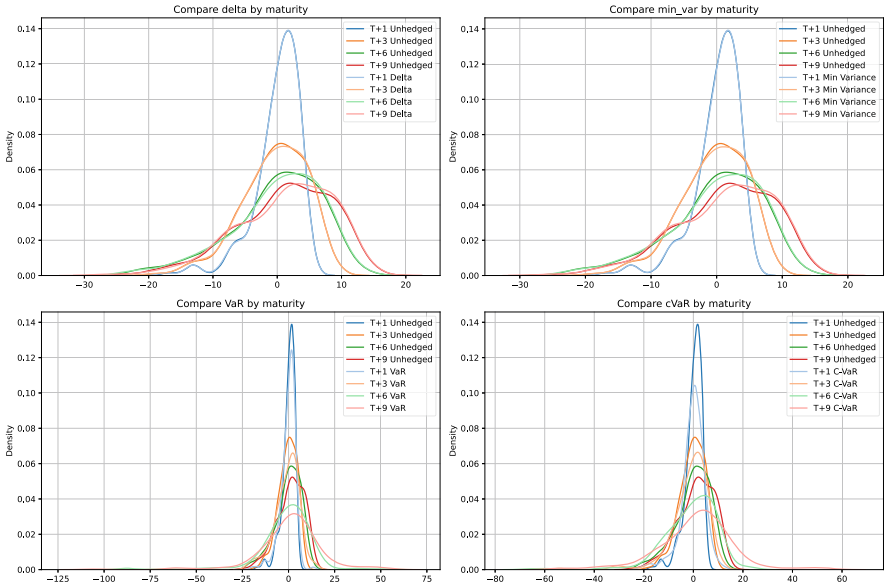


Fig. 6.9 Beef index call PnL by maturity

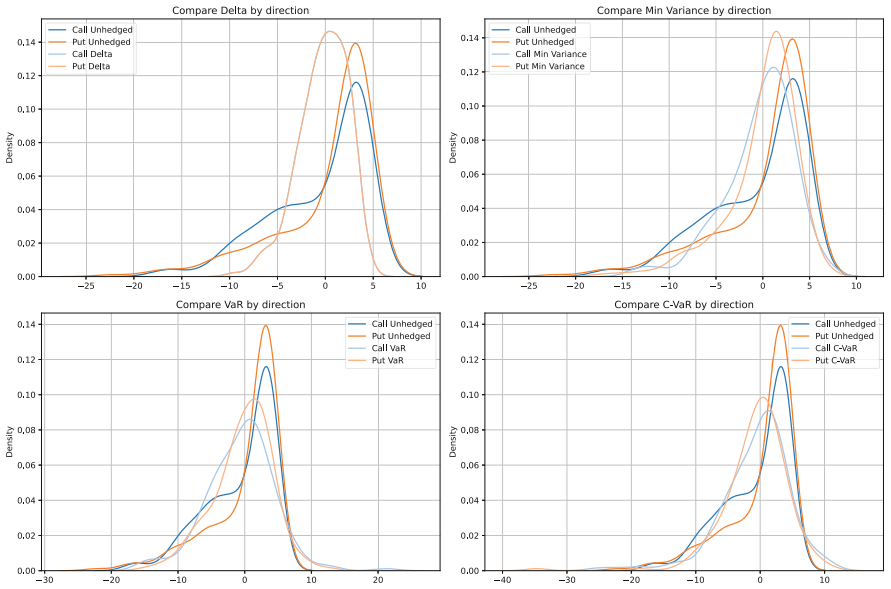


Fig. 6.10 Beef index PnL by direction for T+3

6.5.1.3 Option Parameters

In the graphs above we compare the distribution of the hedged option versus the distribution of the unhedged option for different maturities and option types for each of the different methods, to see how much hedging changes the distribution. We see that the hedged distribution is very close to the unhedged distribution in most cases, and we observe that the choice of option structure (i.e. maturity and direction) determines the PnL distribution to a much greater extent than the hedging method. This implies that as a business, even with the use of hedging, it is important to carefully consider which options to sell. In particular, we observe that longer maturity options have larger standard deviations.

6.5.1.4 Distribution Statistics

The distribution statistics for Beef Index are presented below, for the other indices please see the appendix (Tables 6.12, 6.13, 6.14, and 6.15).

For Beef Index, we see that for some structures, hedging can improve the mean PnL, however, it is important to remember that this is not the purpose of hedging and not something that we would theoretically expect (theoretically, hedging should decrease the variance for the delta and min-variance methods, and the downside risk for the VaR and C-VaR methods). We also notice that it is not consistently one hedging method or one portfolio structure where the mean PnL increases. We notice that on the call side, the mean PnL tends to be negative, so generally these options

Table 6.12 Mean for beef index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−0.2109	−0.1422	−0.1653	−0.1771	−0.1573
Call 3	−0.8467	−0.5860	−0.7068	−0.8417	−0.8101
Call 6	−0.9432	−0.4593	−0.6634	−1.0762	−1.1008
Call 9	−0.3777	0.2556	−0.0072	0.3926	0.2608
Put 1	−0.0409	−0.1416	−0.1031	−0.1641	−0.3338
Put 3	−0.3064	−0.5850	−0.4555	−0.5855	−0.6389
Put 6	0.0188	−0.4578	−0.2550	−0.7931	−0.7602
Put 9	0.8606	0.2574	0.5100	0.1601	0.1676

Table 6.13 Variance for beef index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	9.9243	3.7838	8.2382	9.3837	16.6556
Call 3	29.2269	6.6907	15.8911	24.2474	29.6565
Call 6	56.9035	14.0960	25.1000	79.1817	38.2790
Call 9	71.8468	15.3826	33.9543	107.3063	106.6336
Put 1	14.8385	3.7838	10.1095	13.6720	13.5544
Put 3	31.9415	6.6907	16.3692	20.6991	17.8467
Put 6	60.3011	14.0959	29.2963	96.1624	44.8980
Put 9	60.9454	15.3823	28.4532	109.5258	57.1545

Table 6.14 Skewness for beef index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−2.0687	−1.6811	−1.4224	−0.9722	−0.7528
Call 3	−1.0696	−0.7055	−0.9330	−0.1397	−1.1975
Call 6	−1.2129	−0.7592	−1.2110	−2.8252	−1.8626
Call 9	−0.9973	−0.5950	−0.8639	−0.6734	−0.3091
Put 1	−2.5875	−1.6811	−1.8912	−1.1733	−0.5877
Put 3	−1.7605	−0.7055	−1.3851	−0.4626	−0.2622
Put 6	−1.9495	−0.7592	−1.5268	−1.8864	−1.7593
Put 9	−2.0975	−0.5949	−1.6664	−1.7611	0.0959

(with or without hedging) are losing money. On the put side, there is a more clear trend that hedging, across all methods decreases the mean PnL. In general it seems that the VaR and C-VaR methods are performing worse, the same as we found for hedging effectiveness. For the Cattle Index, we see a similar result, although we note that on the put side, for hedging, the results are all negative showing that for this option direction on this index, hedging will always decrease the mean. Contrastingly, for the Beef Trim Index, we note that the values are positive across most of the portfolio and the mean value is higher with hedging when using the delta and min-variance methods on the call side but not on the put side. We note that for one or two structures the C-VaR method is performing very well, but for this index,

Table 6.15 Kurtosis for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	5.5737	3.5868	4.0113	1.9812	3.8910
Call 3	0.1126	0.5378	1.1850	0.4239	3.8831
Call 6	0.7112	0.1645	1.3641	15.0943	8.8220
Call 9	−0.0502	0.0811	0.3992	3.5329	1.7729
Put 1	6.3872	3.5868	5.4906	4.5041	0.8310
Put 3	1.9591	0.5379	2.4702	0.5640	0.3689
Put 6	2.1501	0.1645	2.3774	13.2697	6.8277
Put 9	2.8343	0.0810	4.1259	11.2378	2.7797

we should not consider high hedging mean values as indicative of success of the method as there is low correlation between the underlying asset and the futures, so it is most likely coincidence.

For variance, we have a very similar result to hedging effectiveness with standard deviation as this is testing the same aspect of the distribution. The delta and min-variance methods reliably decrease the variance—the delta method notably more than the min-variance as expected. Furthermore, we notice that as the maturity increases the variance typically increases, across all of the indexes and methods. We note that the VaR and C-VaR methods can decrease the variance, although this is not always the case and in some cases they can increase it significantly. Given that these methods also do not reliably increase the mean, we suggest that they are not as strong as the minimum variance method. Typically, the Beef Index has the lowest variance, and it is also important to note that the lower the correlation between the index and the futures, the less effective the min-variance method is at decreasing the variance.

For the skew, we note that hedging typically makes the skew significantly less negative, for the delta and min-variance methods for most of the structures. For the VaR and C-VaR methods in some cases they can improve the skew but generally do not. The Cattle Index has similar improvements to the Beef Index, for the min-variance method, and for Beef Trim, there is much smaller improvement as we would expect since it is less correlated to the futures. Similarly, for the kurtosis, the delta and min-variance methods are able to decrease the kurtosis for all of the indexes in most portfolio cases, with the Beef Index seeing the most improvement and the Beef Trim seeing the least. We also note that the VaR and C-VaR methods are sometimes able to decrease the kurtosis significantly but in other cases, will increase it significantly, so is not a consistently applicable result.

6.5.1.5 Static Hedging Summary

Overall, we conclude that cross-hedging can be useful to reduce the VaR-95% and standard deviation when the hedging instrument is sufficiently highly correlated with the asset the option is written on. In particular, we conclude that the min

variance method consistently performs the best compared to delta, across the entire testing portfolio; although the other methods could potentially be improved by hyperparameter optimisation. Across the indices, the log return correlation can be as low as 50% whilst still maintaining good hedging performance according to this hedging effectiveness metric.

We also conclude, however, that using simply VaR-95% as the hedging effectiveness measure is insufficient to fully capture the performance of the hedging technique. In particular, we highlight that from the CDFs and KDEs we see that no method is consistently better in terms of PnL distribution. From the characteristics of the distribution, we note that there is no significant consistent improvement in mean across the portfolio or methods, however, we see that for the delta and min-variance methods they can make the skew less negative and decrease the kurtosis.

Finally, we note that the PnL distribution is mostly determined by the underlying option type, i.e. maturity and direction, so from a business perspective it is important to consider carefully which structures to sell, even with the tool of hedging at hand.

6.5.2 Dynamic Hedging

As in the previous section, on static statistical hedging, in this section, we begin by presenting the results in terms of hedging effectiveness (Tables 6.16, 6.17, 6.18, 6.19, 6.20, and 6.21).

6.5.2.1 Hedging Effectiveness

We begin by noting that for the Beef Index, the min-variance method performs the best, and in some cases has a very high hedging effectiveness for both the VaR-95% and standard deviation risk measures, such as 72.42% for the Put T+9 structure (Table 6.24). The VaR and C-VaR methods generally tend to have negative hedging effectiveness on the put side (for both risk measures), for this index, but on the call side can somewhat improve the risk versus the unhedged option, but not consistently

Table 6.16 Hedging effectiveness for beef index risk measure: VaR-95%

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	−0.4034	−0.0922	−0.0362	−0.0072
Call 3	0.3218	0.4448	0.0307	−0.0420
Call 6	0.5943	0.5134	0.2449	0.1085
Call 9	0.5880	0.5265	0.0717	0.0717
Put 1	0.1077	0.0812	−0.1013	−0.1013
Put 3	0.3477	0.4565	−0.4102	−0.3580
Put 6	0.5575	0.6390	−0.1873	−0.1873
Put 9	0.6355	0.7242	0.0000	−0.0251

Table 6.17 Hedging effectiveness for cattle index risk measure: VaR-95%

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	−0.4045	0.0175	−0.0016	−0.0299
Call 3	0.2966	0.1516	0.1287	0.1287
Call 6	0.4179	0.0511	0.0647	0.0647
Call 9	0.5976	0.0043	0.1195	0.1195
Put 1	−0.1908	−0.0066	−0.1361	−0.1361
Put 3	0.0928	0.0431	−0.0848	−0.0421
Put 6	0.2570	0.0317	−0.1068	−0.1068
Put 9	0.1329	−0.0295	0.0000	0.0284

Table 6.18 Hedging effectiveness for beef trim index risk measure: VaR-95%

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.0641	−0.0101	0.0371	0.0371
Call 3	0.4661	0.1591	−0.1209	−0.0822
Call 6	0.5201	0.0310	−0.0763	−0.0739
Call 9	0.4095	−0.0574	0.0000	0.0000
Put 1	−0.3491	−0.0416	−0.4785	−0.4786
Put 3	0.4723	0.0466	−0.0520	−0.0520
Put 6	0.4397	0.0723	−0.1705	−0.1705
Put 9	0.2721	0.0115	−0.0590	−0.0979

Table 6.19 Hedging effectiveness for beef index risk measure: std

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	−0.0764	0.0199	−0.3346	−0.3274
Call 3	0.3909	0.3676	−0.0695	−0.0869
Call 6	0.4997	0.4373	0.0663	0.0487
Call 9	0.5651	0.4439	0.0648	0.0590
Put 1	0.1164	0.0520	−0.1246	−0.1235
Put 3	0.4025	0.3180	−0.3456	−0.3451
Put 6	0.5608	0.4912	−0.1806	−0.1820
Put 9	0.5801	0.4950	−0.0884	−0.0173

or as significantly as the delta or min-variance methods. For the min-variance the dynamic results are significantly better than in the static case for only the Beef Index. For the Cattle Index, the performance overall of the min-variance method is quite close to zero for both risk measures in most cases. It is worse than in the static case and has similar results to the Beef Trim Index. We note that also for the Cattle and Beef Trim Indices the delta method performs worse than the static delta method, particularly for the Cattle Index, which may partly explain why the min-variance method may be performing worse. For the VaR and C-VaR methods the results suggest that these methods do not apply very well, again, this may be due to the choice of hyperparameter, or due to limitations of the underlying method. We

Table 6.20 Hedging effectiveness for cattle index risk measure: std

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	−0.0228	0.0366	−0.2177	−0.2277
Call 3	0.3611	0.0801	−0.0417	−0.0441
Call 6	0.5058	0.0543	0.0311	0.0257
Call 9	0.5938	0.0559	0.0824	0.0828
Put 1	0.0195	0.0093	−0.1824	−0.1803
Put 3	0.2489	0.0384	−0.3011	−0.2914
Put 6	0.3320	0.0477	−0.2182	−0.2169
Put 9	0.2850	0.0413	−0.0429	−0.0584

Table 6.21 Hedging effectiveness for beef trim index risk measure: std

Hedging technique	Delta	Min variance	VaR	C-VaR
Call 1	0.0832	−0.0368	−0.0823	−0.0823
Call 3	0.4664	0.0421	−0.0498	−0.0478
Call 6	0.5525	0.0234	−0.0953	−0.0937
Call 9	0.5266	0.0023	−0.0668	−0.0670
Put 1	−0.1631	0.0033	−0.2295	−0.2299
Put 3	0.3797	0.0482	−0.2036	−0.2046
Put 6	0.3764	0.0406	−0.1905	−0.1913
Put 9	0.3170	0.0475	−0.1609	−0.2004

do find this somewhat surprising however, as we would expect a better estimate of the distribution and therefore the VaR and C-VaR using weekly data rather than monthly data. We note that for Beef Index the hedging effectiveness values are slightly higher for the risk measure VaR-95% compared with standard deviation, however, in order to draw conclusions about the distribution we should examine the plots and moments of the distribution. In some cases above we note that the T+1 option can perform badly, and suggest that this is due to there being very few payouts of the underlying option, making hedging unnecessary, as empirically there is very little risk that needs to be mitigated (Figs. 6.11, 6.12, 6.13, and 6.14).

6.5.2.2 ECDF and KDE Plots

From the distributions, we observe similarly to the static case that there is no one method that significantly outperforms the other methods. From the ECDFs, this is particularly true for the Call T+1,3,6 options. For the Call T+9 and Put T+6,9 we notice that the delta S-curve is much tighter than the unhedged and VaR/C-VaR methods, so at the lower end of the distribution they may be less likely to lose as much money, but at the upper end, they are less likely to make as much money. We also note that the unhedged, VaR and C-VaR are very closely related in terms of their ECDF shape. The best shape for a particular business depends heavily on the business's risk profile and key performance indicators (Figs. 6.15 and 6.16).

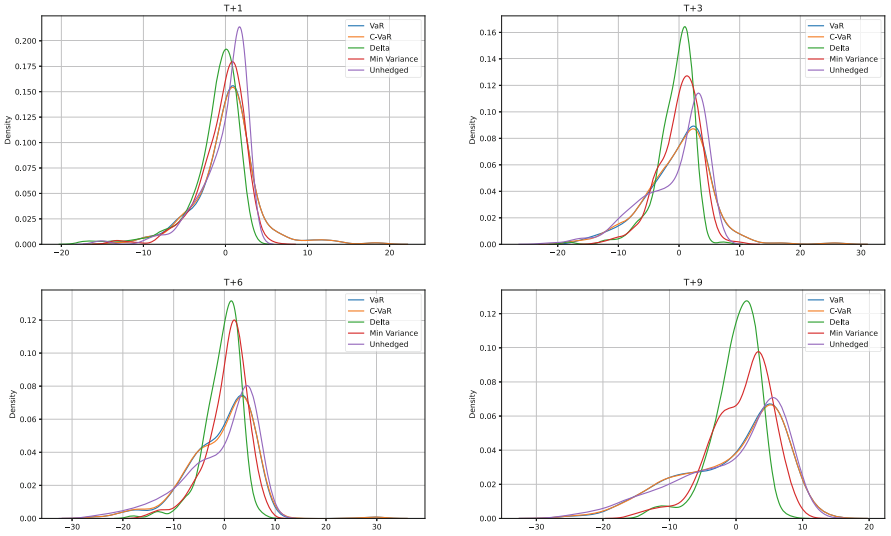


Fig. 6.11 Call option PnL KDE beef index

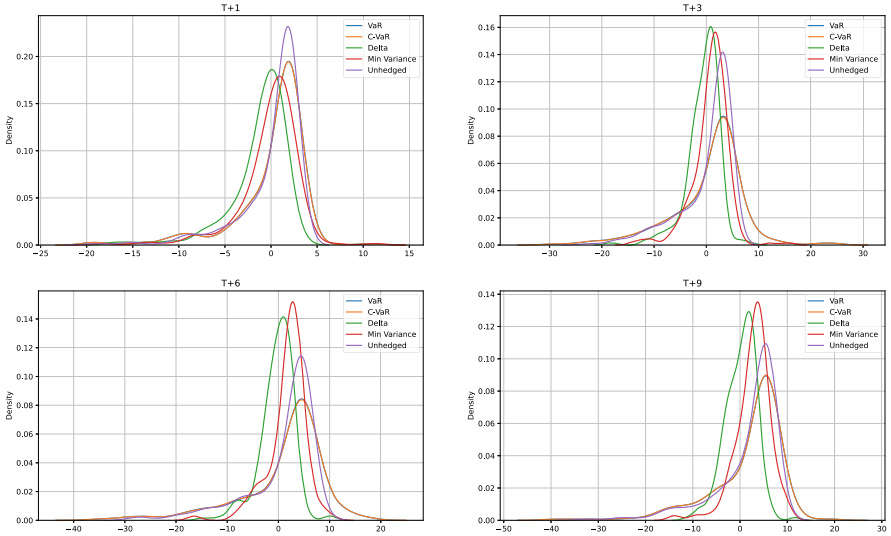


Fig. 6.12 Put option PnL KDE beef index

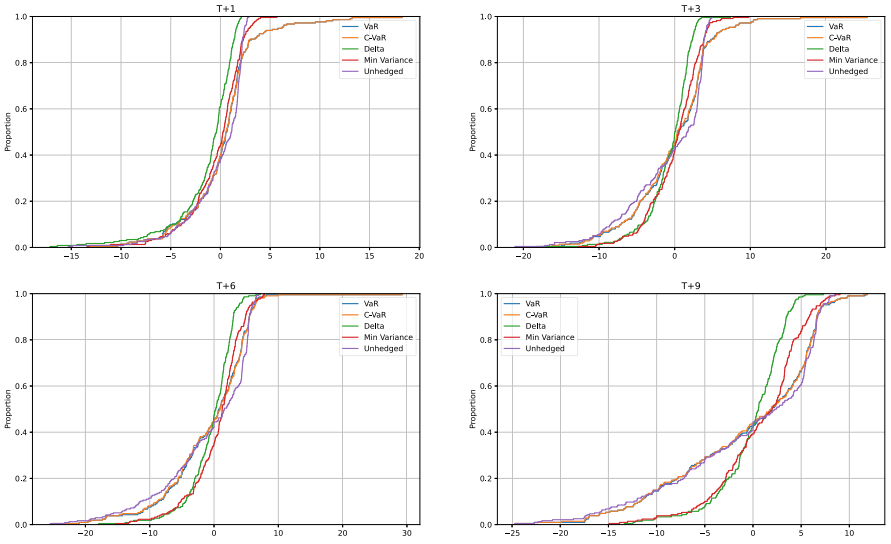


Fig. 6.13 Call option PnL empirical CDF beef index

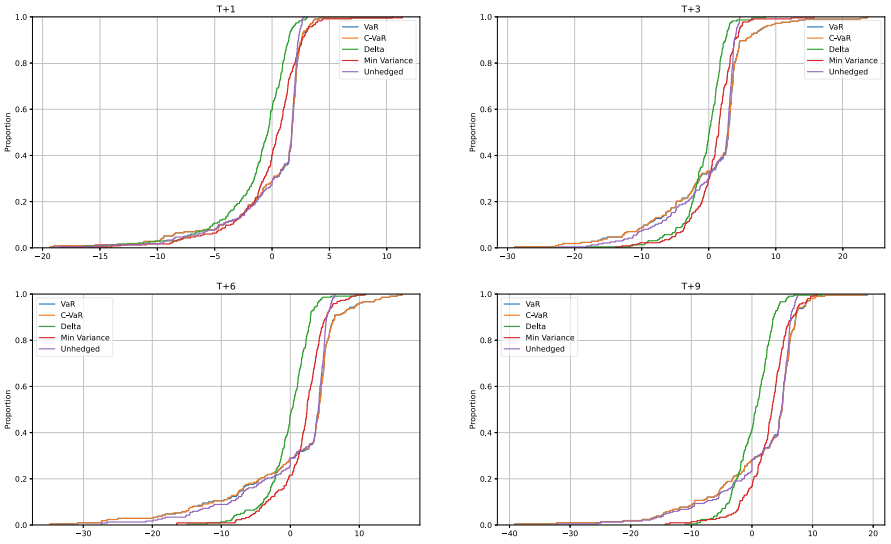


Fig. 6.14 Put option PnL empirical CDF beef index

6.5.2.3 Option Parameters

We observe that dynamically hedging can have a bigger impact on the unhedged PnL distribution, than in the static case, however, because dynamically hedging as

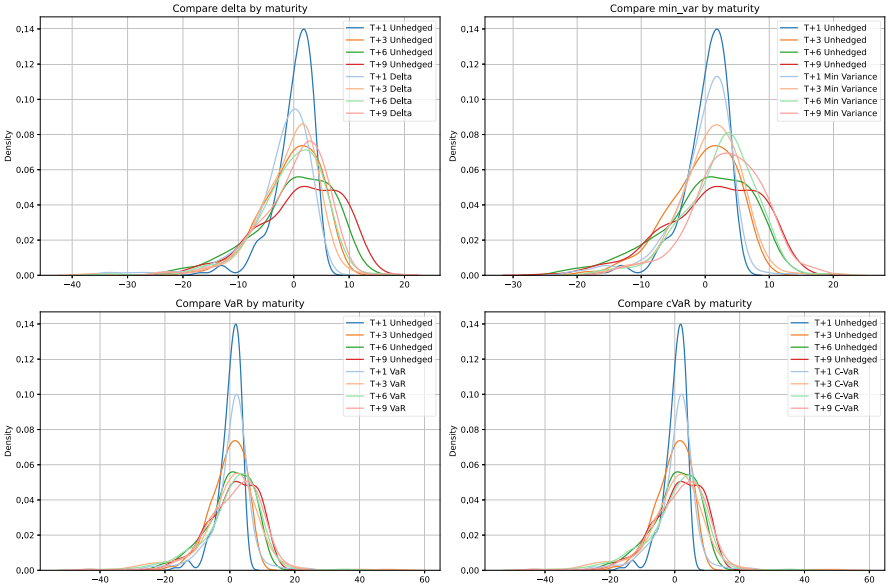


Fig. 6.15 Beef index call PnL by maturity

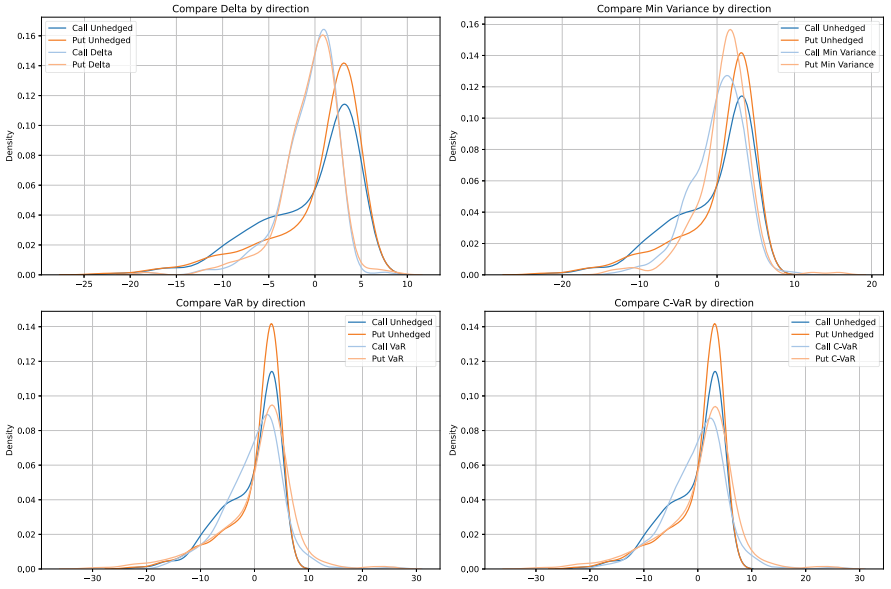


Fig. 6.16 Beef index PnL by direction for T+3

Table 6.22 Mean for beef index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−0.2109	−1.5944	−0.6160	0.0125	0.0229
Call 3	−0.8467	−1.0568	−0.1904	−0.2726	−0.3164
Call 6	−0.9432	−0.7550	0.1369	−0.8668	−0.8905
Call 9	−0.3777	−0.2997	0.3118	−0.4998	−0.6229
Put 1	−0.0409	−1.5176	−0.3387	−0.0109	−0.0257
Put 3	−0.3064	−0.9137	0.5410	−0.2057	−0.2167
Put 6	0.0188	−0.4788	1.3561	0.1408	0.1465
Put 9	0.8606	0.1096	2.3711	0.8453	0.7124

Table 6.23 Variance for beef index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	9.9243	11.4982	9.5335	17.6770	17.4870
Call 3	29.2269	10.8430	11.6877	33.4307	34.5264
Call 6	56.9035	14.2435	18.0182	49.6093	51.4909
Call 9	71.8468	13.5895	22.2201	62.8340	63.6220
Put 1	14.8385	11.5846	13.3342	18.7680	18.7293
Put 3	31.9415	11.4017	14.8573	57.8351	57.7877
Put 6	60.3011	11.6314	15.6077	84.0449	84.2494
Put 9	60.9454	10.7465	15.5447	72.1974	72.1399

Table 6.24 Skewness for beef index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−2.0206	−2.1602	−1.4222	0.5364	0.5635
Call 3	−0.9763	−1.5543	−0.4590	0.2107	0.1224
Call 6	−1.1598	−1.3885	−1.2136	−0.4853	−0.4757
Call 9	−0.9929	−1.2666	−0.7767	−0.8344	−0.8227
Put 1	−2.4211	−2.1187	−1.7445	−2.1850	−2.1826
Put 3	−1.5714	−1.4157	−0.7257	−0.8945	−0.8850
Put 6	−1.6836	−0.9210	−1.6324	−1.5501	−1.5414
Put 9	−1.7614	−0.4074	−1.4176	−1.7430	−1.7516

a method is less reliable across all structures in terms of improving the key metrics, this may not necessarily be advantageous (Tables 6.22, 6.23, 6.24, and 6.25).

6.5.2.4 Distribution Statistics

In contrast to static hedging, dynamic hedging almost always decreases the mean PnL in the case of delta. The min-variance method can significantly increase the mean PnL but not consistently across all structures. Similarly, for the VaR and C-VaR methods whilst we do see some improvement for some structures this is not

Table 6.25 Kurtosis for beef index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	5.5737	5.7770	3.1787	3.4977	3.5928
Call 3	0.1126	4.8820	0.6931	2.2960	2.3220
Call 6	0.7112	3.6672	1.7716	2.5063	2.2240
Call 9	−0.0502	2.2275	0.5414	−0.2136	−0.2214
Put 1	6.3872	5.7400	5.9066	5.6676	5.6724
Put 3	1.9591	4.9215	3.8318	2.4453	2.4262
Put 6	2.1501	3.3130	4.4283	2.4150	2.3836
Put 9	2.8343	1.2924	3.6728	3.9664	3.9671

consistent. In the case of the Cattle Index, generally delta or min-variance hedging will decrease mean PnL, but the VaR and C-VaR methods seem to improve it slightly, this is also the case for the Beef Trim Index.

The delta and min-variance methods in the dynamic case can decrease the variance significantly compared to the unhedged portfolio. The most notable improvements for the delta and min-variance methods are for the longer maturity options. The VaR and C-VaR methods seem to increase the variance, which aligns with the ECDF results which show much longer tails (on both the profit and loss side) for these methods. The performance for hedging is worse for the Beef Trim Index as we would expect from the lower correlation.

In some cases the min-variance method can make the skew more positive, but in particular we notice that the VaR and C-VaR methods always improve the skew for all of the indexes. The delta method, on the other hand, is less reliable. Similarly, for kurtosis, the VaR and C-VaR methods show significant improvement for the Cattle and Beef Trim Indexes, but none of the methods performs well for the Beef Index.

In terms of the distribution, the static method for delta and min-variance seems to be the best, but for VaR and C-VaR the dynamic method is better.

6.5.2.5 Dynamic Hedging Summary

Overall, we conclude that the delta and min-variance methods can reliably produce an improvement when considering the risk measures VaR-95% and standard deviation, and the VaR and C-VaR methods generally tend to increase the portfolio risk according to these measures. We also conclude that in the case of very high correlation, such as the Beef Index, both the static and dynamic min-variance methods perform well. In the case of a slightly lower correlation, the static method may be more reliable and we also show that for low correlation, none of the methods are able to hedge well.

From the distributions we note that the delta and min-variance methods have a tighter S-curve than the other methods, but there is no method that outperforms all the others, and that the underlying option PnL distribution is not significantly

impacted by hedging, although more so than in the static case. From the distribution statistics, we see that the delta and min-variance methods are the best for decreasing variance, but VaR and C-VaR can improve the mean, skewness and kurtosis for these indexes in this portfolio.

6.6 Conclusion and Next Steps

This paper examined different hedging techniques for options written on non-exchange-traded agricultural commodities using futures markets to hedge and evaluated performance based on risk reduction and distributional changes.

As introduced at the start of the paper, agriculture faces risks such as weather and supply chain disruption, necessitating effective risk management solutions. We outlined in the literature review that there are some available risk reduction solutions including: insurance (e.g. crop protection, revenue insurance, multi-peril cover), swaps and physical hedging. However, we also discussed that due to the basis risk that arises when applying these strategies they may have limited effectiveness. Consequently, options were proposed as an alternative risk transfer mechanism, however, the provider of these options, such as an insurance company, would still be exposed. So, in this paper we explored how cross-hedging options with the futures could offset some of the risk, despite the presence of basis risk between the commodity the option is written on and the futures, which prevents perfect hedging. Furthermore, as highlighted in the literature review, determining effective hedging strategies remains a challenge, particularly in the agricultural commodity space. We also emphasised the challenges within the beef market specifically, discussing some of the historical scenarios where the market has seen large moves, and hedging could have helped to mitigate the negative impact for certain participants.

Following this, we introduced basis risk and examined the Black-Scholes options pricing model as a basis for option pricing and hedging. We highlighted that its assumptions of constant volatility and lognormal asset prices do not perfectly hold for commodities and chose to calculate realised volatility given the difficulties of accurately determining implied volatility. We also introduced delta hedging, which offsets option exposure with positions in the underlying asset and this served as our benchmark hedging technique. Then we looked at minimum variance hedging which reduces risk by taking positions correlated with the option's value, adjustable for basis risk through parameters like correlation and volatility ratios between the hedging instrument and commodity. Finally, we introduced value-at-risk (VaR) and conditional VaR (C-VaR) to estimate tail risk given historical data. To evaluate the hedging performance we used hedging effectiveness and distributional analysis.

The results showed that when the correlation between the commodity and futures is sufficiently high, hedging can be effective at reducing risk as measured by VaR and standard deviation. Specifically, the minimum variance hedging method (in the static or dynamic case) consistently performed the best in reducing risk across the

testing portfolio. Hedging was still beneficial at correlation levels as low as 50%, but performance deteriorated significantly below that.

However, the analysis also highlighted the limitations of relying solely on risk reduction metrics in evaluating hedging performance. The empirical cumulative distribution functions and density plots revealed that no single method consistently outperformed the others across the entire profit and loss distribution. While delta and minimum variance hedging tended to produce tighter distributions, they also reduced upside potential.

Furthermore, the underlying option portfolio structure, in terms of maturity and direction, was found to be the primary driver of PnL distribution characteristics. This suggests carefully selecting which option contracts to sell is critical, even with the ability to hedge.

For further research, we suggest some potential avenues that were beyond the scope of this paper. Firstly, an inclusion of some factor-based modelling or an investigation of how the accuracy of weather forecasts or other similar factors could affect the hedging effectiveness. This could be together with an analysis of how businesses can integrate multiple risk management and modelling strategies, such as insurance, derivatives and other products. This could also involve exploring the optimal structure for a portfolio of options (based on maturity and direction) or other products, using mathematical models and empirical evidence to determine the best combination under the metrics we used (hedging effectiveness and distribution shape). Another area that we considered was the development of new dynamic hedging strategies based on machine learning algorithms, which could optimise hedging positions in real-time. An improvement that could be made on the ideas outlined in this paper, could be using parametric methods for volatility determination rather than realised volatility. Such models could include GARCH or even stochastic volatility models such as the ones referenced in the literature review.

Overall, in our paper, we determined that hedging can be a useful risk management tool in agricultural commodity markets, but effectiveness depends on the correlation between the commodity and hedging instrument. Performance should be evaluated across multiple distributional metrics beyond just risk reduction. And the specifics of the option portfolio must be optimized in conjunction with the hedging strategy. Overall though, static or dynamic minimum variance methods can consistently and reliably improve the performance from a risk reduction perspective. So, whether businesses are facing climate change, disease, geopolitical or other threats and uncertainties, we hope in this paper that we have outlined a compelling empirical case as to how hedging can aid with these difficulties and how it can go hand-in-hand with insurance and derivative products to create a more stable world.

Appendix

Distribution Characteristics Tables

In this section we include the distribution characteristics tables for static and dynamic hedging. The Beef Index tables are included in the main body under the Distribution Statistics for static and dynamic hedging (Tables 6.26, 6.27, 6.28, 6.29, 6.30, 6.31, 6.32, 6.33, 6.34, 6.35, 6.36, 6.37, 6.38, 6.39, 6.40, and 6.41).

Static Hedging

Table 6.26 Mean for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−0.5873	−0.4696	−0.5336	−0.4392	−0.2728
Call 3	−1.8492	−1.4243	−1.6922	−1.8785	−1.5804
Call 6	−2.1868	−1.4640	−1.8613	−1.7015	−1.5080
Call 9	−1.4603	−0.5476	−1.0215	−0.0381	−0.3270
Put 1	−0.2993	−0.4688	−0.3702	−0.3491	−0.3138
Put 3	−0.9656	−1.4228	−1.1329	−1.4275	−0.9472
Put 6	−0.7366	−1.4618	−1.0567	−1.1410	−1.4890
Put 9	0.3458	−0.5450	−0.0729	−0.9816	−0.7749

Table 6.27 Mean for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−0.8790	−0.7919	−0.8515	−0.9550	−0.8136
Call 3	0.8811	1.1038	1.0991	0.5999	1.3546
Call 6	1.7631	2.0055	2.5851	1.0351	2.4060
Call 9	5.3902	5.4735	6.9155	5.7888	7.3604
Put 1	−0.6962	−0.7900	−0.7501	−1.1241	−1.2427
Put 3	1.2840	1.1070	1.0833	0.4598	1.7703
Put 6	2.1694	2.0101	1.5702	−0.0099	1.8833
Put 9	5.5181	5.4791	4.5197	1.5514	1.4915

Table 6.28 Variance for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	22.3405	6.6895	20.6959	24.5609	43.6251
Call 3	79.6392	19.3699	57.0603	70.6811	104.2377
Call 6	142.0795	38.9137	89.7025	176.0261	162.7391
Call 9	166.8090	38.5972	102.6592	216.9109	216.0969
Put 1	27.4455	6.6895	22.7619	3 0.5784	41.9304
Put 3	82.4682	19.3699	56.2714	65.9703	69.3905
Put 6	163.2686	38.9133	109.5829	173.1818	161.4827
Put 9	161.8271	38.5968	104.1511	209.8416	141.4741

Table 6.29 Variance for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	149.0044	32.9681	160.2384	198.9867	237.0847
Call 3	292.2755	77.9039	295.2031	382.4818	565.1916
Call 6	557.6767	115.0349	557.9309	1013.7182	778.5046
Call 9	500.9028	88.7914	500.9619	752.7107	917.3208
Put 1	93.0474	32.9680	88.3725	128.9141	147.6476
Put 3	218.4219	77.9051	188.8343	289.1910	296.4834
Put 6	287.6363	115.0372	261.6935	401.3391	742.7581
Put 9	243.8102	88.7944	217.1969	412.6709	510.4654

Table 6.30 Skewness for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−1.4010	−1.0094	−1.2366	−0.6578	−0.4366
Call 3	−1.0205	−0.6052	−1.0780	−0.4603	−1.2504
Call 6	−1.2319	−0.8779	−1.3793	−1.3965	−1.0490
Call 9	−0.9291	−0.8561	−0.9467	−0.8613	−0.5335
Put 1	−1.9056	−1.0093	−1.5716	−1.0663	−0.6574
Put 3	−1.6436	−0.6051	−1.6345	−0.7862	−0.9962
Put 6	−1.8040	−0.8778	−1.8925	−0.8970	−1.2588
Put 9	−1.9731	−0.8560	−2.1552	−1.7681	−0.4370

Table 6.31 Skewness for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−2.2893	−1.5948	−2.4663	−1.2442	−0.6929
Call 3	−2.0774	−1.7253	−2.0364	−1.3142	−0.6354
Call 6	−1.9449	−1.1329	−1.6310	−1.7166	−1.1669
Call 9	−1.8936	−1.1393	−1.5844	−1.1711	−1.3723
Put 1	−2.1679	−1.5946	−1.8636	−0.9589	−0.3472
Put 3	−2.9721	−1.7250	−2.4915	−0.6194	−0.4029
Put 6	−1.8089	−1.1327	−1.7916	−0.8290	−0.7220
Put 9	−1.4466	−1.1390	−1.3927	−1.4289	−1.2631

Table 6.32 Kurtosis for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	1.7234	1.0167	1.7863	0.3614	1.9116
Call 3	−0.0241	0.2509	0.6050	0.1084	3.9259
Call 6	1.0725	0.2564	2.0758	6.3459	3.4160
Call 9	−0.2935	0.8715	0.0007	3.4842	2.3878
Put 1	3.3892	1.0166	2.7465	1.6180	2.1631
Put 3	2.3184	0.2509	2.5941	1.1036	1.7344
Put 6	2.4430	0.2564	3.0963	2.8164	4.2433
Put 9	4.0944	0.8714	5.7179	8.7876	1.7705

Table 6.33 Kurtosis for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	7.1503	3.9243	9.1571	2.4450	2.1047
Call 3	6.1841	4.8951	6.6880	3.2934	2.0029
Call 6	5.6227	2.0437	4.7585	5.4086	2.7921
Call 9	6.5858	2.5634	5.9925	3.8224	4.7108
Put 1	7.1741	3.9233	5.7334	2.6111	1.4392
Put 3	13.0155	4.8938	9.3500	2.5506	1.8857
Put 6	3.5724	2.0429	3.9982	0.7487	1.1668
Put 9	1.5814	2.5623	1.8069	2.7987	2.2816

Dynamic Hedging

Table 6.34 Mean for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−0.5873	−2.9629	−0.6850	−0.2663	−0.3318
Call 3	−1.8492	−2.5944	−1.8151	−1.7381	−1.7661
Call 6	−2.1868	−2.6152	−2.3326	−2.1077	−2.0187
Call 9	−1.4603	−2.0179	−2.3067	−1.1020	−1.1134
Put 1	−0.2993	−3.1472	−0.4445	−0.2417	−0.2611
Put 3	−0.9656	−3.1167	−1.0708	−0.6863	−0.6631
Put 6	−0.7366	−3.6875	−0.8872	−0.6458	−0.5530
Put 9	0.3458	−3.3263	0.2078	0.3295	1.2754

Table 6.35 Mean for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−0.8790	−5.4956	−0.9462	−0.5039	−0.5039
Call 3	0.8811	−1.1330	0.8400	1.1984	1.1898
Call 6	1.7631	−0.5406	1.6743	2.4324	2.4767
Call 9	5.3902	−0.4777	5.6861	5.6122	5.6175
Put 1	−0.6962	−5.4470	−0.9016	−0.5498	−0.5577
Put 3	1.2840	−0.9854	1.1724	1.5944	1.6306
Put 6	2.1694	−0.2394	1.6031	2.3945	2.3807
Put 9	5.5181	−0.0167	4.3429	5.2880	5.6543

Table 6.36 Variance for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	22.3405	23.3693	20.7348	33.1242	33.6754
Call 3	79.6392	32.5090	67.3926	86.4241	86.8111
Call 6	142.0795	34.6997	127.0819	133.3684	134.8644
Call 9	166.8090	27.5255	148.6778	140.4403	140.3273
Put 1	27.4455	26.3852	26.9391	38.3734	38.2328
Put 3	82.4682	46.5198	76.2583	139.6103	137.5404
Put 6	163.2686	72.8606	148.0608	242.3029	241.7913
Put 9	161.8271	82.7316	148.7378	176.0066	176.0281

Table 6.37 Variance for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	149.0044	125.2446	160.1679	174.5308	174.5316
Call 3	292.2755	83.2329	268.1574	322.1275	320.8781
Call 6	557.6767	111.6821	531.8373	669.0072	667.1213
Call 9	500.9028	112.2671	498.5604	570.1006	570.2705
Put 1	93.0474	125.8705	92.4385	140.6457	140.7535
Put 3	218.4219	84.0424	197.8860	316.4085	316.9538
Put 6	287.6363	111.8528	264.7452	407.6689	408.2326
Put 9	243.8102	113.7422	221.1892	328.5619	329.0632

Table 6.38 Skewness for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−1.4010	−1.4114	−1.5035	−0.2502	−0.2388
Call 3	−1.0205	−1.1451	−1.0303	−0.5620	−0.5644
Call 6	−1.2319	−1.3383	−1.3162	−1.1908	−1.1727
Call 9	−0.9291	−0.4144	−0.8679	−0.8036	−0.8016
Put 1	−1.9056	−1.5610	−1.9193	−1.8053	−1.8091
Put 3	−1.6436	−2.0571	−1.6128	−0.9675	−0.9593
Put 6	−1.8040	−2.1639	−1.7948	−1.2830	−1.3021
Put 9	−1.9731	−1.9983	−1.9253	−1.7191	−1.7185

Table 6.39 Skewness for beef trim index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	−2.2893	−2.5771	−2.2548	−2.1579	−2.1579
Call 3	−2.0774	−1.6360	−2.0031	−1.7067	−1.7125
Call 6	−1.9449	−1.1844	−1.8323	−1.0942	−1.1060
Call 9	−1.8936	−1.2436	−1.7725	−1.2831	−1.2852
Put 1	−2.1679	−2.5662	−1.9671	−1.3053	−1.3032
Put 3	−2.9721	−1.6331	−2.6457	−1.7105	−1.7111
Put 6	−1.8089	−1.2345	−1.6727	−1.1061	−1.1023
Put 9	−1.4466	−1.2949	−1.2792	−0.8522	−0.8490

Table 6.40 Kurtosis for cattle index

Hedging technique	Unhedged	Delta	Min-Var	VaR method	C-VaR method
Call 1	1.7234	2.8485	2.2133	1.9760	1.8524
Call 3	−0.0241	1.5112	−0.0345	0.2160	0.2121
Call 6	1.0725	2.6724	1.3142	1.7013	1.6613
Call 9	−0.2935	0.7579	−0.4026	−0.2197	−0.2211
Put 1	3.3892	2.4324	3.4664	3.3184	3.3390
Put 3	2.3184	5.2294	2.2252	1.5058	1.5531
Put 6	2.4430	5.4424	2.3923	1.7545	1.8077
Put 9	4.0944	4.7404	3.7732	3.1628	3.1602

Table 6.41 Kurtosis for beef trim index

Hedging method	Unhedged	Delta	min-var	VaR method	C-VaR method
Call 1	7.1503	9.6797	7.8914	9.1014	9.1013
Call 3	6.1841	3.7115	6.5482	4.4213	4.4572
Call 6	5.6227	2.3414	5.3321	3.6878	3.7293
Call 9	6.5858	2.2518	6.1277	5.1738	5.1692
Put 1	7.1741	9.6034	6.3596	2.9943	2.9829
Put 3	13.0155	3.7151	10.7479	6.0636	6.0472
Put 6	3.5724	2.4734	3.0350	1.8338	1.8262
Put 9	1.5814	2.3781	1.2323	0.4645	0.4541

Acknowledgments We would like to acknowledge the contributions of the individuals and organisations who have aided us in the research and writing of this paper. Firstly, we would like to express our appreciation for the contributions of Yujia (Bella) Li, Wanqing (Christine) Zhang and Chenwei Wang from University College London (UCL), who allowed us to work with them on their Masters' theses which formed the basis of some of the methods explored in this paper and helped with the initial research stages, particularly for the parametric and semi-parametric VaR and C-VaR methods. We would also like to thank their supervisor, Tomaso Aste, who worked with us in managing the students' projects.

We would also like to thank our colleagues, Barry Yu and Yuchen Pei, for their expert guidance throughout the evolution of the project. We have really appreciated their encouragement and technical support, which has enabled us to write this paper.

We would also like to thank Hirbod Assa for his continued support over the years and for inviting us to contribute this paper.

Finally, we would like to acknowledge the work of all the researchers whose works contributed to the background and methodology within the paper.

References

1. Irwin, S. H., & Sanders, D. R. (2012). Financialization and structural change in commodity futures markets. *Journal of Agricultural and Applied Economics, 44*(3), 371–396.

2. Animal and Plant Health Inspection Service, U.S. Department of Agriculture (2023). Bovine spongiform encephalopathy (bse). https://www.aphis.usda.gov/aphis/ourfocus/animalhealth/nvap/NVAP-Reference-Guide/Control-and-Eradication/Bovine-Spongiform-Encephalopathy#:~:text=As%20a%20result%2C%20U.S.%20beef,case%20was%20confirmed%20in%20U.S. Last Modified: May 30, 2023.
3. L.A. Times Archives. (2003). Cattle prices fall sharply again. https://www.latimes.com/archives/la-xpm-2003-dec-30-fi-wrap30-story.html. DEC. 30, 2003 12 AM PT.
4. Mead, D., Ransom, K., Reed, S. B., & Sager, S. (2020). The impact of the covid-19 pandemic on food price indexes and data collection. *Monthly Labor Review*.
5. Euromeat News. (2020). US meat prices have surged during the pandemic. https://www.euromeatnews.com/Article-US-Meat-prices-have-surged-during-the-pandemic/5487. Posted on May 25, 07:36.
6. IBISWorld. (2024). Beef cattle production in the US - market size, industry analysis, trends and forecasts (2024–2029). Accessed March 2024.
7. Assa, H., & Wang, M. S. (2021). Price index insurances in the agriculture markets. *North American Actuarial Journal, 25*(2), 286–311.
8. Goodwin, B. K. (2001). Problems with market insurance in agriculture. *American Journal of Agricultural Economics, 83*(3), 643–649.
9. Goodwin, B. K., Mishra, A. K., & Ortalo-Magné, F. N. (2003). What's wrong with our models of agricultural land values? *American Journal of Agricultural Economics, 85*(3), 744–752.
10. Ifft, J. (2022). What is price risk? price risk management for cow-calf producers: Part 1. Email: jifft@ksu.edu.
11. Schroeder, T. C., & Yang, X. (2001). Hedging wholesale beef cuts. In *2001 Annual Meeting, July 8–11, 2001, Logan, Utah* (Vol. 36091). Western Agricultural Economics Association.
12. Bieroth, C. W. (2010). Cross-hedging performance of wholesale beef in live cattle futures contracts revisited. Kansas State University, http://hdl.handle.net/2097/4943
13. Alcoforado, R. G., Egídio dos Reis, A. D., Bernardino, W., & Santos, J. A. C. (2023). Modelling risk for commodities in Brazil: An application for live cattle spot and futures prices. *Commodities, 2*(4), 398–416.
14. Ditsch, M. W., & Leuthold, R. M. (1996). Evaluating the hedging potential of the lean hog futures contract. ACE OFOR Report 14769, University of Illinois at Urbana-Champaign, Department of Agricultural and Consumer Economics.
15. Assa, H. (2015). A financial engineering approach to pricing agricultural insurances. *Agricultural Finance Review, 75*(1), 63–76.
16. Assa, H. (2016). Financial engineering in pricing agricultural derivatives based on demand and volatility. *Agricultural Finance Review, 76*(1), 42–53.
17. Ye, M., Wang, R., Tuo, G., & Wang, T. (2017). Crop price insurance in China: Pricing and hedging using futures market. *China Agricultural Economic Review, 9*(4), 567–587.
18. Johnson, L. L. (1960). The theory of hedging and speculation in commodity futures. *The Review of Economic Studies, 27*(3), 139–151.
19. Frechette, D. L. (2000). The demand for hedging and the value of hedging opportunities. *American Journal of Agricultural Economics, 82*(4), 897–907.
20. Lence, S. H. (1995). The economic value of minimum-variance hedges. *American Journal of Agricultural Economics, 77*(2), 353–364.
21. Lence, S. H. (1996). Relaxing the assumptions of minimum-variance hedging. *Journal of Agricultural and Resource Economics, 21*(1), 39–55.
22. Hirshleifer, D. (1988). Risk, futures pricing, and the organization of production in commodity markets. *Journal of Political Economy, 96*(6), 1206–1220.
23. Simaan, Y. (1993). What is the opportunity cost of mean-variance investment strategies? *Management Science, 39*(5), 578–587.
24. Lien, D., & Tse, Y. K. (2002). Some recent developments in futures hedging. *Journal of Economic Surveys, 16*(3), 357–396.
25. Hung, J.-C., Chiu, C.-L., & Lee, M.-C. (2006). Hedging with zero-value at risk hedge ratio. *Applied Financial Economics, 16*(3), 259–269.

26. Cao, Z., Harris, R. D. F., & Shen, J. (2010). Hedging and value at risk: A semi-parametric approach. *Journal of Futures Markets, 30*(8), 780–794.
27. Cong, J., Tan, K. S., & Weng, C. (2012). Var-based optimal partial hedging. *ASTIN Bulletin, 43*(3), 271–299.
28. Cong, J., Tan, K. S., & Weng, C. (2014). Conditional value-at-risk-based optimal partial hedging. *Journal of Risk, 16*(3), 49–83.
29. Ederington, L. H. (1979). The hedging performance of the new futures markets. *The Journal of Finance, 34*(1), 157–170.
30. Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81*(3), 637–654.
31. Mandelbrot, B. B. (1963). The variation of certain speculative prices. *The Journal of Business, 36*(4), 394–419.
32. Hagan, P. S., Kumar, D., Lesniewski, A. S., & Woodward, D. E. (2002). Managing smile risk. *Wilmott, 1*, 84–108. Archived from the original on 2022-04-30.
33. Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies, 6*(2), 327–343.
34. Chen, L. (1996). Stochastic mean and stochastic volatility — a three-factor model of the term structure of interest rates and its application to the pricing of interest rate derivatives. *Financial Markets, Institutions & Instruments, 5*, 1–88.
35. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327.
36. Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007.
37. Ederington, L. H. (1979). The hedging performance of the new futures markets. *The Journal of Finance, 34*(1), 157–170.
38. USDA. (2009). Index methodology. https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Prices_Received_and_Prices_Received_Indexes/#:~:text=The%20indexes%20measure%20the%20change,parity%20prices%20and%20parity%20ratios

Chapter 7
Crop Yield Insurance Analysis for Turkey: Spatiotemporal Dependence

Güven Şimşek and Kasirga Yildirak

Abstract Farming is among the most vulnerable segments of society due to the source of the income that is highly dependent on environmental risks. To maintain their production, farmers, who are critical components of agricultural production, need to protect themselves against production risks. For farmers to continue agriculture, it is crucial to provide insurance policies that at the very least protect their current income. Therefore, crop yield insurance has been discussed in this study. When a crop yield falls short of a predetermined threshold, crop yield insurance compensates for the resulting yield loss. This insurance product holds a prominent position among other agricultural insurances because yield insurance, which aims to keep agricultural production at a specific level, maintains sustainability in the ecosystem. Through the spatiotemporal modeling of crop yields and yield insurance, the impact of climate change, a major problem for agricultural insurance, has also been addressed. For the conditional crop yield distribution in this study, a hierarchical Bayesian technique is employed to characterize the spatiotemporal dependence. Wheat yield statistics from the years 2004 to 2022 were used for a total of 47 districts that are part of Ankara and Konya, which are at the top of the list in terms of wheat production volume. Premium rates have been obtained for the region, province, and chosen districts using the preferred model in accordance with model selection and performance criteria, and the results are presented. The R-INLA package program is used to perform all statistical analyses for this study.

G. Şimşek (✉) · K. Yildirak
Department of Actuarial Sciences, Faculty of Science, Hacettepe University, Ankara, Turkey
e-mail: guvensimsek@hacettepe.edu.tr

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_7

7.1 Introduction

Over the past few decades, crop insurance has played an important role in the agricultural market. As a result, selecting an accurate statistical model that affects crop yield distribution is crucial in obtaining a reasonable premium rate. Many statistical approaches have been considered to determine the distribution of agricultural yields [9, 12, 13, 37]. [17] and [36] use the beta distribution to determine crop yield distribution. [11] and [34] analyze the crop yield for the revenue insurance using the log-normal distribution.

Crop yields can vary depending on location, time of year, and crop type. As a result, traditional parametric and non-parametric approaches to crop yield modeling are not always feasible. In such cases, models with spatial, temporal, and spatiotemporal characteristics are more suitable for crop yield. [18] investigate the spatiotemporal effects of maize yield data from Brazil. They assess premium rates in the state of Paraná. [29] propose a linear mixed model with a spatio-temporal process for rice and cassava data from Thailand. [20] investigate the distribution of crop yield in Iowa corn and Oklahoma wheat. To estimate the distribution's parameters, a model is proposed using the Bayesian Kriging approach. [39] made use of a spatiotemporal model with water deficiency and water satisfaction index as explanatory variables to forecast the yield in state-run Turkish farms. Furthermore, the spatial effect across the related regions is investigated. [30] "The Integrated Nested Laplace Approximation" approach has been recently proposed by [26]. It has a very broad and extensible model class, including "linear mixed models" and spatiotemporal models [23, 28, 31]. [27] propose a model for determining crop yield insurance premium rates in the state of Paraná (Brazil). Under a Bayesian hierarchical framework, a dynamic spatiotemporal model is used. [30] calculated premium rates for districts using the INLA model and wheat yield data from Ankara and Konya districts from 2004 to 2022.

This study introduces the structure of district-based yield and discusses the significance of district dependency. The hierarchical Bayesian method, which models the conditional distribution of crop yield, reflects spatial or temporal dependencies among crop yields. We also discuss the approaches used in Bayesian modeling for analyzing model performance and selecting models.

In this study, we conduct a case study of 47 districts in Ankara and Konya, the cities with the highest wheat production in Turkey, from 2004 to 2022. We investigate the interdependence of specific subregions by taking into account spatial and temporal effects. We model spatial, temporal, and spatiotemporal effects to estimate wheat yield in Ankara and Konya. To handle space and time effects, we use a hierarchical Bayesian structure for the district-based crop yield data. We use this model to compute the premium rates associated with various coverage levels after selecting the best model.

Finally, we make our closing remarks and present our suggestions for future research.

The data and information used in the research are taken from a chapter in the doctoral dissertation titled "Impacts of Spatiotemporal Dependency and Asymmetric Information on The Analysis of Optimal Crop Yield Insurance." [30].

7.2 Bayesian Hierarchical Modeling

If the distribution of one parameter is conditional on another, we can define a model with a hierarchical structure. To clarify, the prior distribution is assigned to another prior parameter known as the hyperparameter. As a result, we can use hierarchical structure to investigate the spatial or temporal dependence between observations. The joint probability distribution can be used to express relationships between parameters [18]. The hierarchical Bayesian model has three levels [19]:

1. Data (Likelihood) level

$$Y|\nu_1, \nu_2 \sim \pi_1(Y|\nu_1, \nu_2) \tag{7.1}$$

2. Process (Parameter) level

$$\nu_1|\nu_2 \sim \pi_2(\nu_1|\nu_2) \tag{7.2}$$

3. Prior (Hyperparameter) level

$$\nu_2 \sim \pi_3(\nu_2) \tag{7.3}$$

The likelihood function is represented at the first level $\pi_1(Y|\nu_1, \nu_2)$ and the data Y that is conditionally independent of the given ν_1 and ν_2, where ν_1 and ν_2 represent parameter and hyperparameter, respectively. We could analyze spatial or temporal dependence among the observations in Y using the second level given in Eq. (7.2). The final equation represents the prior level with hyperparameter distribution π_3. The posterior distribution can be defined using Bayes' theorem using these levels:

$$\begin{aligned}\pi(\nu_1, \nu_2|y) &= \frac{\pi(Y, \nu_1, \nu_2)}{\pi(Y)} \\ &= \frac{\pi_1(Y|\nu_1, \nu_2)\pi(\nu_1, \nu_2)}{\int_{R_{\nu_1}} \int_{R_{\nu_2}} \pi_1(Y|\nu_1, \nu_2)\pi(\nu_1, \nu_2)d\nu_1 d\nu_2} \propto \pi_1(Y|\nu_1, \nu_2)\pi(\nu_1, \nu_2)\end{aligned} \tag{7.4}$$

where R_{ν_i} are the domains of ν_i. Here, the term $\pi(\nu_1, \nu_2)$ is the prior distribution for the data (likelihood) level. Also, $\pi(\nu_1, \nu_2)$ can be written in two parts as follows:

$$\pi(\nu_1, \nu_2) = \pi_2(\nu_1|\nu_2)\pi_3(\nu_2). \tag{7.5}$$

As a result, if we rearrange Eq. (7.5), we get:

$$\pi(\nu_1, \nu_2|Y) \propto \pi_1(Y|\nu_1, \nu_2)\pi_2(\nu_1|\nu_2)\pi_3(\nu_2). \tag{7.6}$$

As shown in Eq. (7.6), the model's joint posterior is proportional to *likelihood, parameter* and *hyperparameter*.

We apply a similar approach to our data set to see how a hierarchical Bayesian model can be used to analyze spatial data. We simply assume that the response variable Y_s is crop yield in a specific subregion, as shown below:

$$Y_s = \beta X_s^T + \gamma_s + \epsilon_s;\ s = 1, \ldots, n \tag{7.7}$$

where X_s denotes the vector of predictor variables and β indicates the slope of X_s. Here, γ_s and ϵ_s are the random effect and random error of Y_s, respectively. We could investigate the areal effect for data observations using γ_c. In this regard, we consider the following [1] proposed properties:

$$Y_s|\beta, \gamma_s, \sigma_\epsilon^2 \sim N(\beta X_s^T + \gamma_s, \sigma_\epsilon^2)$$

$$\gamma_s|\sigma_\gamma^2 \sim N(0, \sigma_\gamma^2)$$

$$\tau_\gamma \sim \text{IGamma}(a_1, b_1)$$

$$\tau_\epsilon \sim \text{IGamma}(a_2, b_2)$$

$$\beta \sim N(0, C)$$

where IGamma represents the Inverse Gamma distribution with parameters a and b. The precision parameter ($\tau = 1/\text{variance}$) and the variance-covariance matrix are denoted by τ and C, respectively. We define the joint posterior density using the properties mentioned above as follows:

$$\pi(\beta, \sigma_\gamma^2, \sigma_\epsilon^2|Y_s) = \pi_1(Y_s|\beta, \gamma_s, \sigma_\epsilon^2)\pi_2(\gamma_s|\sigma_\gamma^2)\pi_3(\sigma_\gamma^2)\pi_3(\sigma_\epsilon^2)\pi_3(\beta). \tag{7.8}$$

The random effect for Y_s is used here, which is the randomness arising directly from the observations $Y_s^{(i)}$, where i denotes the i-th district with $i = 1, \ldots, n$. We have not yet considered spatial properties in this case. To achieve this goal, we employ a common method proposed by [3], which is an intrinsic conditional autoregressive structure (ICAR) to investigate the spatial random effect. According to this method, the parameter γ mentioned in Eq. (7.8) is defined as follows:

$$\gamma_i|\gamma_{i\neq j} \sim N\left(\bar{\gamma}_i, \frac{1}{\tau_\gamma N_i}\right), \tag{7.9}$$

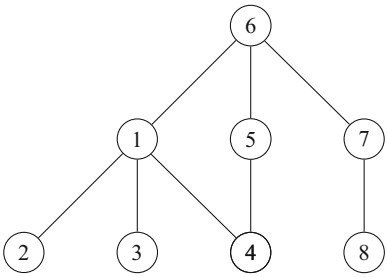


Fig. 7.1 Subregional neighborhood structure

where $\bar{\gamma}_i = N_i^{-1} \sum_{j \in N_i} \gamma_j$ and N_i displays the areas that are adjacent to district i. The mean of "spatial random effects" in the set of neighbors is the expectation of γ_i under the condition γ_j, whereas the conditional precision parameter τ_γ controls the spatial dependence between the observations. As an example of the term $\bar{\gamma}_i$, consider a region with eight subregions (Fig. 7.1).

Area 1, for example, has boundaries with $2, 3, 4$, and 6. As a result, the conditional expectation of γ_1 is as follows:

$$\gamma_1 | \gamma_{1 \neq j} \sim N\left(\frac{\gamma_2 + \gamma_3 + \gamma_4 + \gamma_6}{4}, \frac{1}{4\tau_\gamma}\right). \tag{7.10}$$

Because of the complexity of the models or the high dimension of the data, the MCMC methods for Bayesian modeling may take a long time. Rue et al. [26] recently introduced the INLA approach as an alternative to MCMC. INLA is a well-organized analytical approach applicable to a wide range of models, including spatial, temporal, spatiotemporal, generalized linear mixed models, and stochastic volatility models. The main advantage of INLA is that it reduces computational time compared to MCMC methods. Furthermore, one useful feature of INLA is that it allows us to approximate the posterior distribution of the parameter. INLA is discussed in greater detail in the following subsection.

7.3 Integrated Nested Laplace Approximation (INLA)

The INLA method was developed primarily for the latent Gaussian model (LGM). LGM is a hierarchical structure described in Sect. 7.2.

$$\begin{aligned} Y | \nu, \psi &\sim \pi(y | \nu, \psi) = \prod_{i=1}^{n} \pi(y_i | \nu_i, \psi), \\ \nu | \psi &\sim [\pi(\nu | \psi) = N(0, Q^{-1}(\psi))], \\ \psi &\sim \pi(\psi). \end{aligned} \tag{7.11}$$

The second part of Eq. (7.11) represents the latent process. The precision matrix Q^{-1} is equal to the inverse of the covariance matrix Σ. The parameter vector ν is described by a Gaussian Markov Random Field (GMRF) as shown below [21].

$$\pi(\nu|\psi) = \frac{|Q(\psi)|^{1/2}}{(2\pi)^{n/2}} e^{\left(-\frac{1}{2}\nu^T Q(\psi)\nu\right)}. \tag{7.12}$$

The LGM can be found using additive regression models. These models are defined in the same way that generalized linear models (GLM) are. In contrast to the linear predictor in GLM, the additive predictor for the LGM includes nonlinear effects such as seasonal and spatially structured random effects. We assume in our study that the response or dependent variable $y_i; i = 1, \ldots, n$ belongs to the exponential family [15]. Let μ_i represent the mean of the i-th observation of y. Then it is defined by the additive predictor η_i with the function $g(.)$. Here $g(.)$ denotes the link function, i.e. $g(\mu_i) = \eta_i$. The predictor η_i in its most common form is

$$\eta_i = \alpha_0 + \sum_{k=1}^{N_\beta} \beta_k x_{ki} + \sum_{j=1}^{N_f} f_j(z_{ji}) + \epsilon_i \ . \tag{7.13}$$

The terms used in Eq. (7.13) are defined as follows:

- The intercept is represented by the scalar α_0.
- The linear term is represented by $\beta = (\beta_1, \ldots, \beta_{N_\beta})$, which measures the effect of the vector of covariates x on the response variable.
- $f_j(.)$ is a function of the vector of covariates z that can be used to investigate nonlinear effects of z on the dependent variable, i.e. spatial or temporal random effects.
- The variables N_β and N_f represent the number of corresponding covariates, respectively.

Using Eq. (7.11), we can define the additive predictor η_i as a hierarchical structure. The term $\pi(y|\nu, \psi) = \prod_{i=1}^{n} \pi_1(y_i|\nu_i, \psi)$ denotes that each observation in data y is linked to the i-th element of latent field ν_i [24]. The random vector (latent field) is $\nu = \left(\alpha_0, \beta_{k=1}^{N_\beta}, f_{j=1}^{N_f}(.), \eta\right)$ and encompasses all parameters that are not directly visible in the data. Lastly, $\psi = (H_1, \ldots, H_{n_p})$ indicates the n_p-dimensional vector of hyperparameters H_i. Using these definitions, the joint posterior of ν and ψ can be defined as follows:

$$\begin{aligned} \pi(\nu, \psi) &\propto \pi(y|\nu, \psi)\pi(\nu|\psi)\pi(\psi) \\ &\propto \left(\prod_{i=1}^{n} \pi(y_i|\nu_i, \psi)\right)\pi(\nu|\psi)\pi(\psi) \end{aligned} \tag{7.14}$$

where the density function $\pi(\nu|\psi)$ is defined in Eq. (7.12). Thus, if we replace Eq. (7.12) into Eq. (7.14), the following equation is obtained:

$$\pi(\nu, \psi) \propto \pi(\psi)|Q(\psi)|^{1/2} exp\left(-\frac{1}{2}\nu^T Q(\psi)\nu + \sum_{i=1}^{n} log(\pi(y|\nu, \psi))\right) \tag{7.15}$$

The INLA method's goal is to approximate the marginal posterior distribution of each parameter vector ν and ψ. These marginals are given separately for ν and ψ as follows:

$$\pi(\nu_i|y) = \int \pi(\nu_i, \psi|y)d\psi = \int \pi(\nu_i|\psi, y)\pi(\psi|y)d\psi, \text{ and} \tag{7.16}$$

$$\pi(\psi_h|y) = \int \pi(\psi|y)d\psi_{-h}. \tag{7.17}$$

When the h-th hyperparameter is omitted, the vector of the remaining hyperparameters is denoted by ψ_{-h}. Both equations have the term $\pi(\psi|y)$ in common, as seen in Eqs. (7.16) and (7.17). Thus, we can define $\pi(\psi|y)$ to find an approximation for the marginal posterior distributions for all hyperparameters:

$$\begin{aligned} \pi(\psi|y) &= \frac{\pi(\nu_i, \psi|y)}{\pi(\nu|\psi, y)} \propto \frac{\pi(y|\nu, \psi)\pi(\nu|\psi)\pi(\psi)}{\pi(\nu|\psi, y)} \\ &\approx \frac{\pi(y|\nu, \psi)\pi(\nu|\psi)\pi(\psi)}{\tilde{\pi}(\nu|\psi, y)}\Bigg|_{\nu=\nu^*(\psi)} \end{aligned} \tag{7.18}$$

where $\tilde{\pi}(\nu|\psi, y)$ is the Gaussian approximation of $\pi(\nu|\psi, y)$ and $\nu^*(\psi)$ denotes the mode of ν for a given ψ.

Because the parameter vector ν has more elements than ψ in general, the approximation for the posterior conditional distributions $\pi(\nu_i|\psi, y)$ given ψ and y can be more complex. Three approaches can be used to approximate $\pi(\nu_i|\psi, y)$ [5]. These methods are detailed below:

1. With the help of the Normal distribution, the marginals from $\tilde{\pi}(\nu|\psi, y)$ are used to approximate $\pi(\nu_i|\psi, y)$. The Cholesky decomposition is also used to obtain the precision matrix. According to the other two approaches, this approach approximates $\pi(\nu_i|\psi, y)$ relatively quickly. However, this approach is typically not very good at approximation [25].
2. The Laplace Gaussian approximation is an alternative approach to the approximation for $\pi(\nu_i|\psi, y)$. The vector of parameters can be rewritten as $\nu = (\nu_i, \nu_{-i})$, where ν_{-i} denotes the vector of the remaining parameters when the i-th parameter is omitted. The joint posteriors for the parameter ν can then be approximated using the Laplace approximation, as follows:

$$\begin{aligned}\pi(\nu_i|\psi, y) &= \frac{\pi(\nu_i, \nu_{-i}|y)}{\pi(\nu_{-i}|\nu_i, \psi, y)} \\ &\approx \frac{\pi(y|\nu, \psi)\pi(\nu|\psi)\pi(\psi)}{\tilde{\pi}(\nu_{-i}|\nu_i, \psi, y)}\Bigg|_{\nu_{-i}=\nu^*_{-i}(\nu_i,\psi)}\end{aligned} \tag{7.19}$$

where $\tilde{\pi}(\nu_{-i}|\nu_i, \psi, y)$ represents Laplace Gaussian approximation of $\pi(\nu_{-i}|\nu_i, \psi, y)$ and $\nu^*_{-i}(\nu_i, \psi)$ is the mode of ν_{-i}. The Laplace Gaussian approximation works well, but it takes a long time to compute.

3. The final method, known as simplified Laplace approximation, is based on Taylor's series of Eq. (7.19). In comparison to the other two methods mentioned above, it is more rational and computationally efficient.

In Eqs. (7.16) and (7.17), we can write the approximated marginal posteriors of $\pi(\nu_i|y)$ and $\pi(\psi_h|y)$ as follows:

$$\tilde{\pi}(\nu_i|y) = \int \tilde{\pi}(\nu_i|\psi, y)\tilde{\pi}(\psi_l|y)d\psi, \tag{7.20}$$

$$\tilde{\pi}(\psi_h|y) = \int \tilde{\pi}(\psi|y)d\psi_{-h}. \tag{7.21}$$

Numerical integration can be used to obtain the solution of the given integral in Eq. (7.20):

$$\tilde{\pi}(\nu_i|y) = \sum_{l=1}^{L} \tilde{\pi}(\nu_i|\psi^{(l)}, y)\tilde{\pi}(\psi^{(l)}|y)\Delta_l \tag{7.22}$$

where Δ_l denotes the set of weights and $\psi^{(l)}$ represents some integration points. More detailed information about these approximations and INLA can be found in [26] and [4].

7.3.1 Adequacy of Models Based on Predictive Distribution

Assessing the model's adequacy is critical for making sound decisions following the modeling process. The most widely used Bayesian modeling approaches are based on predictive distribution. The data is divided into two categories: training and validation. The training data is used to fit the model, while the validation data is used to assess the prediction's accuracy. For testing model adequacy, the posterior predictive check method and/or leave-one-out cross-validation can be used.

- **Posterior predictive analysis**

  The posterior predictive distribution and p-value are two important quantities for the model posterior predictive checks [10]. These quantities' mathematical representations are provided by

$$\pi(y_i^{rep}|y) = \int_\theta \pi(y_i^{rep}|\theta_i)\pi(\theta_i|y)d\theta_i, \tag{7.23}$$

$$P(Y_i^{rep} \leq y_i|y). \tag{7.24}$$

  The first equation represents the posterior predictive distribution, where $\pi(y_i^{rep}|y)$ represents the density of a replicated observation Y_i^{rep}. The posterior predictive p-value is represented by $P(Y_i^{rep} \leq y_i|y)$ [33]. Using these values, we can infer whether the model is appropriate for the data or not. If $\pi(y_i^{rep}|y)$ has a large number of small values, it indicates that the related observation is an outlier. Thus, that model is not suitable for the data set. As a result, that model is unsuitable for the data set. Furthermore, if the values of $P(Y_i^{rep} \leq y_i|y)$ are close to zero or one, the model appears to be invalid for the data.

- **Leave-one-out-cross-validation**

  The conditional predictive ordinate (CPO) [22] and the probability integral transform (PIT) [7] are measures of goodness of fit, or prediction performance. The primary goal of these measures is to assign numerical scores to models based on their predictive distribution. The first criterion is defined as

$$\text{CPO}_i = \pi(y_i|y_{-i}) \tag{7.25}$$

  where y_{-i} denotes the vector of the remaining observations in the data set after omitting the i-th observation. High CPO values indicate that the model's predictive performance for y_i is good, whereas very low CPO values indicate that the i-th observation may be an outlier. Using the CPO values, we can also calculate the logarithmic score to select the best model. This score is determined as follows:

$$\text{Lscore} = -\frac{\sum_i^n \log(CPO_i)}{n}. \tag{7.26}$$

  In this case, a low log-score value indicates that the interested model is reasonable. In the case where each observation in the data set is independent of the others, the logarithmic score and the Akaike information criterion (AIC) are asymptotically equivalent [35].

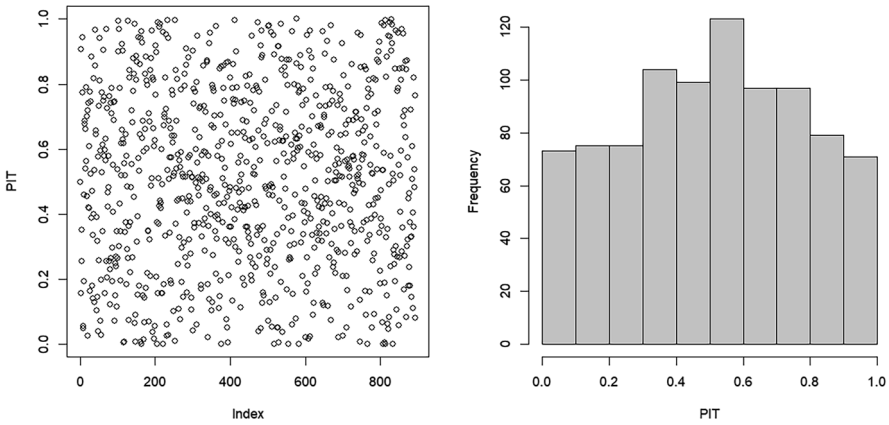


Fig. 7.2 Graph of the PIT values (left) and the histogram of PIT (right)

The second metric is computed as follows:

$$\text{PIT}_i = P\left(Y_i^{rep} \leq y_i | y_{-i}\right) \tag{7.27}$$

where PIT_i is the calibration of the i-th observation to the rest of the data, i.e. the i-th observation is removed. If PIT values are at the extremes, i.e. very small or very high, the observations associated with these values may be outliers. Furthermore, the PIT values must have a standard uniform distribution; otherwise, the model fit is unsatisfactory. We can also use the PIT histogram to assess the model's suitability for the data.

The PIT values and histogram of PIT values are shown in the figure below.

The histogram of these values based on the model, as seen in Fig. 7.2, represents that the distribution of PIT values is close to uniform.

7.3.2 Model Preference

To compare the fit of different models, the Deviance Information Criterion (DIC) proposed by [32] is used. The DIC model is defined by two terms: goodness of fit and penalty.

$$DIC = \bar{D} + p_D \tag{7.28}$$

where $\bar{D}$ is the posterior expectation of the Bayesian deviance $D(\nu)$ and is used to evaluate model fit.

$$D(\nu) = -2\log(\pi(y|\nu)) + 2\log(\pi(y)) \tag{7.29}$$

where the likelihood function is denoted by $\pi(y|\nu)$, and the term $2\log(\pi(y))$ can be omitted when comparing many models for the same data. The value p_D in Eq. (7.28) is referred to as the number of effective parameters and is used to assess the model's complexity. It is equal to the expectation of deviance minus the expectation of deviance.

$$p_D = \bar{D} - D(\bar{\nu}). \tag{7.30}$$

In this case, $\bar{\nu}$ denotes the expected value of ν, which can be represented by $E_{\nu|y}(\nu)$. Lastly, the model which has the smaller DIC value should be chosen over other models with larger DIC values.

In Sect. 7.4, we present the results of a hierarchical Bayesian model and premium calculations.

7.4 An Application of Spatiotemporal Models to the calculation of Crop Yield Insurance Premium

Obtaining premium rates based on statistical models becomes more difficult for crop yield modeling due to factors such as moral hazard and the dependence of catastrophic losses in terms of spatial and temporal effects. To model crop yield, we take spatial, temporal, and spatiotemporal effects into account. To capture spatial and temporal effects, we use a hierarchical Bayesian structure for district-based crop yield data.

Using the R-INLA package [4, 5] provided by R software, we obtain estimations for the considered models and calculate premium rates. We use ArcGIS software tools to display some of the graphs.

In this study, we use TUIK (Turkish Statistical Institute) crop yield data for wheat in the cities of Ankara and Konya in Turkey's Central Anatolia region from 2004 to 2022. Both cities play a significant role in wheat production. Ankara and Konya's total wheat production in 2022 is 2,135,587 tons, accounting for 13.35% of Turkey's total wheat production of 16,000,000 tons.

We simulate crop yield for each of these cities' districts. These districts are useful for considering spatiotemporal dependency because they share common borders, implying that they are neighbors.

We use the district IDs in this study for convenience. The corresponding district IDs are listed in Table 7.1. The data set in the application part of our study, as shown in the table, consists of 47 districts in Konya and Ankara. The following figure is provided to handle the geographical information for these districts.

The neighborhood structures among districts can be seen visually in Fig. 7.3. For example, District 4 Altinekin's neighbors from north to south-east are District 10 Cihanbeyli, District 39 Sarayonu, District 40 Selcuklu, and District 31 Karatay.

Table 7.1 The ID of the districts

District	District ID	District	District ID	District	District ID
Akoren	1	Elmadag	17	Kızılcahamam	33
Aksehir	2	Emirgazi	18	Kulu	34
Altındag	3	Eregli	19	Mamak	35
Altınekin	4	Evren	20	Meram	36
Ayas	5	Golbasi	21	Nallihan	37
Bala	6	Gudul	22	Polatli	38
Beypazari	7	Guneysinir	23	Sarayonu	39
Beysehir	8	Halkapinar	24	Selcuklu	40
Bozkir	9	Haymana	25	Seydisehir	41
Cihanbeyli	10	Huyuk	26	Sincan	42
Cankaya	11	Ilgin	27	Sereflikochisar	43
Celtik	12	Kadinhani	28	Tuzlukcu	44
Cubuk	13	Kalecik	29	Yalihuyuk	45
Cumra	14	Karapinar	30	Yenimahalle	46
Derbent	15	Karatay	31	Yunak	47
Doganhisar	16	Kecioren	32		

Fig. 7.3 The map of the districts with ID

Figure 7.4 is provided to show the structure of the neighborhood among districts. It represents nodes and lines between nodes in order to visualize the IDs of adjacent districts.

The neighborhood relationships between districts are easy to find in this figure. For example, as shown in Fig. 7.3, District 4's neighbors from right to left are District 10, District 39, District 40, and District 31, respectively.

The adjacency matrix, shown in the figure below, is another representation of the neighborhood.

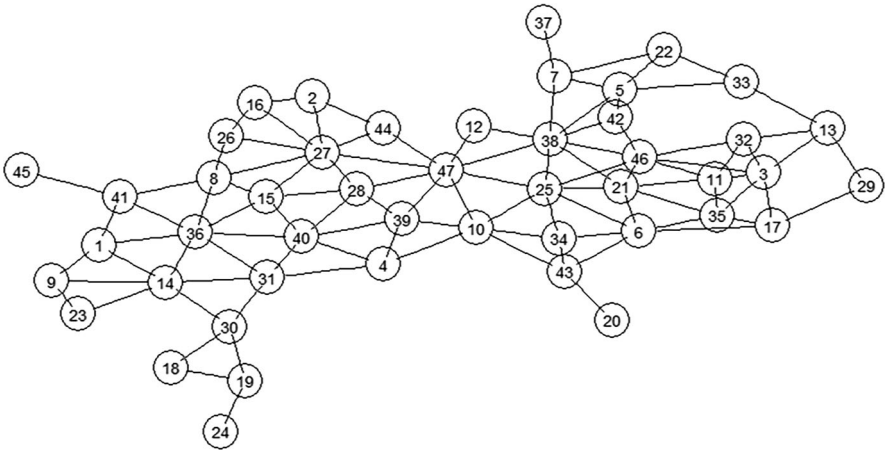


Fig. 7.4 The nodes of adjacent districts with their IDs

The adjacency matrix is a popular graphical tool in the literature. The numbers on the x and y axes of Fig. 7.5 represent the district IDs used in the case study.

7.4.1 Models Used in the Application

In this paper, we use various models to investigate the spatial, temporal, and spatiotemporal effects on wheat yield in crop yield insurance. The theoretical considerations and INLA results of these models for estimating district wheat yields are provided in the following subsections. Equation (7.7) gives the general theoretical modeling of the hierarchical Bayesian approach and Sect. 7.2 gives the related properties. According to [1], the conditional distribution of $Y_s|\beta, \gamma_s, \sigma_\epsilon^2$ is assumed to be normal with the parameters $\beta X_s^T + \gamma_s$ and σ_ϵ^2, respectively. Assume $Y_{s,t}$ are the total 893 yield observations for 47 districts over a 19-year period.

7.4.1.1 Basic Error Model (Model 1)

To begin, we examine the fundamental model, which includes an intercept term and an unstructured spatial random effect. Although this model is less informative than the other models used in this study, we use it as a baseline model to examine the effect of intercept and unstructured spatial effect on spatial models.

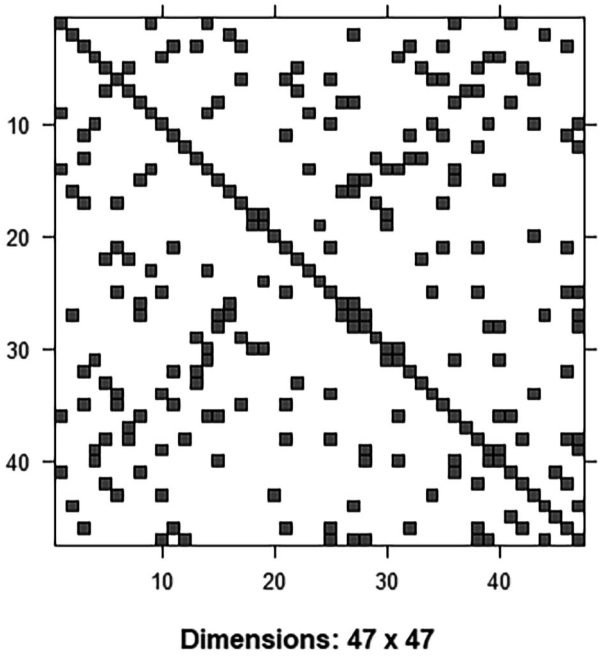


Fig. 7.5 The adjacency matrix for the districts with ID

For the years $t = 1, \ldots, 19$, let $Y_{s,t}^{(i)}$ denote the wheat yield in the district i with location s. The fundamental error model is given by:

$$\begin{aligned}
&\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \epsilon_{s,t}^{(i)}; \quad i = 1, \ldots, 47 \text{ and } t = 1, \ldots, 19 \\
&\alpha_0 \sim N(0, \sigma_{\alpha_0}^2) \\
&\epsilon_{s,t} \sim N(0, \sigma_{\epsilon_{s,t}}^2) \\
&\gamma_{s_1} \sim N(0, \sigma_{\gamma_{s_1}}^2) \\
&\log \tau_{\gamma_{s_1}} \sim \log \text{Gamma}(a_{s_1}, b_{s_1}) \\
&\log \tau_{\epsilon_{s,t}} \sim \log \text{Gamma}(a_{\epsilon_1}, b_{\epsilon_2})
\end{aligned} \tag{7.31}$$

where;

- The intercept term for the model, α_0 , represents the mean wheat yield for the districts.
- The spatially unstructured effect is denoted by $\gamma_{s_1}^{(i)}$, and we assume that it is independent and identically distributed (*iid*).

7.4.1.2 Spatial Model (Model 2)

We extend the basic error model (Model 1) to the spatial model (Model 2) in this model to investigate spatial dependence. To accomplish this, we add the spatially structured random effect to Eq. (7.31). The following is the second model [5]:

$$\begin{aligned}
&\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \epsilon_{s,t}^{(i)}, \\
&\gamma_{s_2} \sim N(0, \sigma_{\gamma_{s_2}}^2), \\
&\log \tau_{\gamma_{s_2}} \sim \log \text{Gamma}(a_{s_2}, b_{s_2})
\end{aligned} \tag{7.32}$$

where $\gamma_{s_2}^i$ denotes the spatial structured term. We define the properties of this term in Eq. (7.9).

7.4.1.3 Spatiotemporal Models (Model 3–6)

In order to investigate temporal effects on crop yield estimation, we revisit the spatial model in Eq. (7.32). In this section, we first present the parametric representation for spatial-temporal modeling proposed by Bernardinelli et al. [2]. The following is the definition of **Model 3**:

$$\begin{aligned}
&\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \left(\alpha + \delta_{s,t}^{(i)}\right) t + \epsilon_{s,t}^{(i)}, \\
&\delta_{s,t} \sim N(0, \sigma_{\delta_{st}}^2), \\
&\log \tau_{\delta_{s,t}} \sim \log \text{Gamma}(a_{\delta_{s,t}}, b_{\delta_{s,t}}).
\end{aligned} \tag{7.33}$$

Here, the term $\left(\alpha + \delta_{s,t}^{(i)}\right) t$ consists of two components where:

- The term α refers to a main linear trend that represents the overall trend effect.
- $\delta_{s,t}^{(i)}$ denotes the time trend associated with the district i and is used to define the interaction between space and time.

For the spatial-temporal modeling, we present a non-parametric formulation borrowed from Knorr-Held [14]. **Model 4** is provided by:

$$\begin{aligned}
&\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \phi_{t_1} + \phi_{t_2} + \epsilon_{s,t}^{(i)}, \\
&\phi_{t_1} \sim N(0, \sigma_{\phi_{t_1}}^2), \\
&\phi_{t_2} \sim N\left(0, \frac{1}{\tau}\right),
\end{aligned} \tag{7.34}$$

where the model's parameters have the same characteristics as the previous models except for the terms ϕ_{t_1} and ϕ_{t_2}. ϕ_{t_1} denotes temporally unstructured random effect and it is modeled by using a Gaussian exchangeable prior, i.e. $\phi_{t_1} \sim N(0, \sigma^2_{\phi_{t_1}})$. Furthermore, the term ϕ_{t_2} denotes the structured temporal effect, which is represented by the random walk model of order 1 (rw1) as follows:

$$\Delta x_i = x_i - x_{i-1}$$

For the temporally structured random effect ϕ_{t_2}, we use rw1. Another goal of this research is to investigate the spatial-temporal interactions that explain differences in yield amount based on the spatial and temporal interaction trends of different districts. In this sense, Eq. (7.34) can be extended by adding an unstructured interaction term $\delta_{s,t}$, resulting in **Model 5** as follows:

$$\eta^{(i)}_{s,t} = \log(Y^{(i)}_{s,t}) = \alpha_0 + \gamma^{(i)}_{s_1} + \gamma^{(i)}_{s_2} + \phi_{t_1} + \phi_{t_2} + \delta^{(i)}_{s,t} + \epsilon^{(i)}_{s,t}. \tag{7.35}$$

In this section, we define $\delta_{s,t}$ as the interaction term between the unstructured effects γ_{s_1} and ϕ_{t_1}. We define $\boldsymbol{\Theta_\delta}$ as a structure matrix for the unstructured effects γ_{s_1} and ϕ_{t_1}, and can be represented by the *Kronecker product* i.e, $\boldsymbol{\Theta_\delta} = \boldsymbol{\Theta_{\gamma_{s_1}}} \otimes \boldsymbol{\Theta_{\phi_{t_1}}} = \boldsymbol{I} \otimes \boldsymbol{I} = \boldsymbol{I}$ [6]. Because γ_{s_1} and ϕ_{t_1} represent unstructured spatial and temporal effects, the interaction term $\delta_{s,t}$ has no spatial or temporal effect. Using this result, we assume that the interaction term is normally distributed with a mean of 0 and a variance of $\dfrac{1}{\tau_{\delta_{s,t}}}$, and that it is *iid*.

Finally, we look at how the spatial and temporal structure affects the interaction term. As a result, the estimation of the parameters associated with the term $\delta_{s,t}$ is not obtained under the assumption that $\delta_{s,t}$ is *iid*, as given in Eq. (7.35). The goal of using **Model 6** as shown in Eq. (7.36) is to investigate the overall trend that is correlated with both spatial and temporal characteristics of the data based on their neighbors [38]. As a result, the interaction term $\delta_{s,t}$ is introduced into the model as a random effect. We use the Besag-York-Mollie (BYM) model for spatially structured random effects and the rw1 process for structured temporal random effects.

The following are the fixed and random effects for the structured spatiotemporal interaction model (**Model 6**):

$$\begin{aligned} &\eta^{(i)}_{s,t} = \log(Y^{(i)}_{s,t}) = \alpha_0 + \gamma^{(i)}_{s_1} + \gamma^{(i)}_{s_2} + \phi_{t_1} + \phi_{t_2} + \delta^{(i)}_{s,t} + \epsilon^{(i)}_{s,t}, \\ &\delta_{s,t} \text{ is the random effect.} \end{aligned} \tag{7.36}$$

Following a thorough examination of each model, we compare the models under consideration in the following section.

7.4.2 Model Selection

We examine six models for the district-based distribution of wheat yield. The basic error model is the first model that takes into account the spatially unstructured random effect for districts, i.e. it is assumed to be *iid*. Then we phase in different models that include spatial, temporal, and spatiotemporal effects. In this section, we present the comparison criteria for the models considered in Sect. 7.4.1. These criteria could also be used for model selection and model diagnostics.

In our case study, we use two criteria, DIC and Lscore, to select a reasonable model to estimate wheat yield. We select the best model based on the lowest DIC and Lscore values. The following table shows the DIC values, which are the sum of $\bar{D}$ in the first column and p_D in the second column, as well as the Lscore values. The model formulations are also shown in this table.

As shown in Table 7.2, **Model 6** has the lowest DIC and Lscore values (the values highlighted in bold in Table 7.2), indicating that it is the best model among all considered models. **Model 6** is a structured spatiotemporal interaction model that includes spatial, temporal, and spatiotemporal effects. In this work, the precisions τ for the random effects are assigned using R-INLA priors Gamma(1,0.00005).

The following graph compares the observed and fitted values obtained by using the chosen model (**Model 6**) over the last four years (2019–2022) (Fig. 7.6).

Here, studies were conducted using different a priori distributions to examine the connections between computation time efficiency and parameter estimations for the INLA and MCMC approaches. The NIMBLE package available in R has been used for MCMC algorithms [8]. Table 7.3 presents comparisons for **Model 6**, which was determined to be the best model by applying the INLA technique.

Table 7.2 $\bar{D}$, p_D , DIC and Lscore results for model selection

Models	$\bar{D}$	p_D	DIC	Lscore
Model 1	$\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \epsilon_{s,t}^{(i)}$			
	417.48	39.42	456.90	0.255
Model 2	$\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \epsilon_{s,t}^{(i)}$			
	417.64	39.50	457.14	0.257
Model 3	$\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \left(\alpha + \delta_{s,t}^{(i)}\right) t + \epsilon_{s,t}^{(i)}$			
	283.92	49.02	332.94	0.188
Model 4	$\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \phi_{t_1} + \phi_{t_2} + \epsilon_{s,t}^{(i)}$			
	−37.20	58.93	21.73	0.014
Model 5	$\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \phi_{t_1} + \phi_{t_2} + \delta_{s,t}^{(i)} + \epsilon_{s,t}^{(i)}$			
	−40.19	62.37	22.18	0.015
Model 6	$\eta_{s,t}^{(i)} = \log(Y_{s,t}^{(i)}) = \alpha_0 + \gamma_{s_1}^{(i)} + \gamma_{s_2}^{(i)} + \phi_{t_1} + \phi_{t_2} + \delta_{s,t}^{(i)} + \epsilon_{s,t}^{(i)}$			
	−569.02	237.34	**−293.46**	**−0.124**

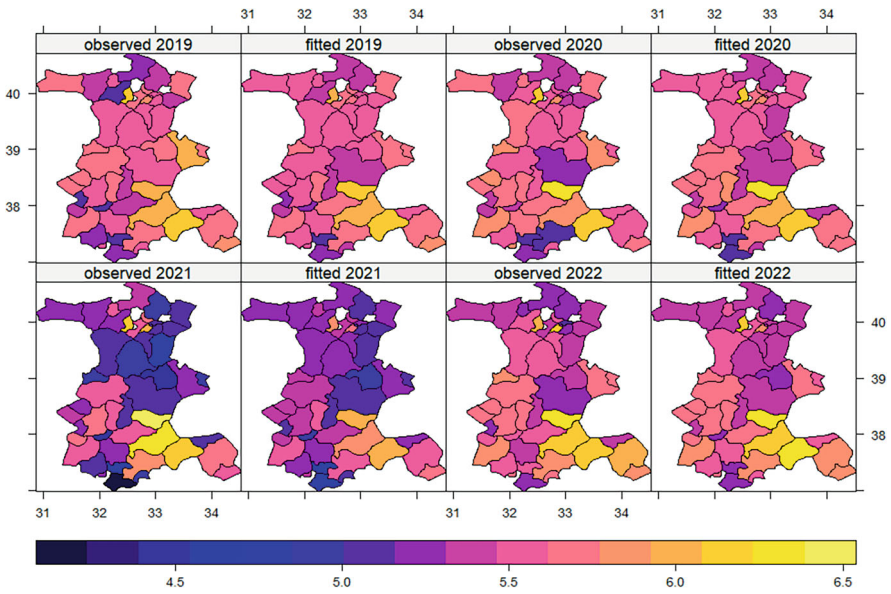


Fig. 7.6 The observed and the fitted values of wheat yield according to space and time

Table 7.3 Comparison of model running times for INLA and MCMC

		Gamma(1,0.00005)		Gamma(1,0.5)		Gamma(1,1)	
		Mean	Sd	Mean	Sd	Mean	Sd
α_0	INLA	5.45	0.01	5.45	0.15	5.45	0.20
	MCMC	5.44	0.03	5.43	0.08	5.46	0.11
τ_ϵ	INLA	32.39	2.59	34.12	2.69	35.36	2.78
	MCMC	32.65	2.53	32.68	2.49	31.81	2.33
$\tau_{\phi_{t_1}}$	INLA	32.04	10.82	10.69	4.17	5.69	2.21
	MCMC	2063.17	8934.81	10.01	3.84	5.75	2.22
$\tau_{\phi_{t_2}}$	INLA	11, 730.98	15, 591.13	9.80	4.06	5.14	2.11
	MCMC	5834.38	12, 545.88	8.93	3.67	5.05	2.02
$\tau_{\gamma_{s_1}}$	INLA	22, 195.76	24, 462.18	2.29	2.47	1.18	1.26
	MCMC	6783.04	17, 061.04	21.52	5.33	12.65	2.98
$\tau_{\gamma_{s_2}}$	INLA	22, 389.39	24, 713.49	2.26	2.46	1.16	1.24
	MCMC	457.25	735.51	13.01	4.15	7.69	2.32
τ_δ	INLA	14.29	2.78	11.64	1.91	10.20	1.52
	MCMC	39.19	7.65	29.23	4.49	24.75	3.25

To explore how alternative prior distributions affect the parameter estimates for the two approaches, here we employ not only the Gamma(1,0.00005) prior distribution for random effects, but also the Gamma(1,0.5) and Gamma(1,1) prior distributions. Regarding the INLA and MCMC approaches, the constant term α_0 and the τ_ϵ precision parameter for Gaussian observations have very similar values within

Table 7.4 Comparison of running times, mean absolute error (MAE) and root mean squared error (RMSE) for INLA and MCMC approaches with the prior distribution Gamma(1,0.00005)

Approach	User	System	Elapsed	MAE	RMSE
INLA	0.16	0.12	5.26	0.1077	0.1462
MCMC	25.70	0.44	64.74	0.1067	0.1459

three prior distributions; however, the two approaches have different values for the other random effects under different prior distributions. In comparison to other spatial and temporal random effects under three distinct prior distributions, INLA and MCMC have more accurate values when we examine the mean value for the precise parameter τ_δ, which displays spatial-temporal interaction. Furthermore, for Gamma(1,0.5) and Gamma (1,1) prior distribution options, it is noted that the estimate values of the precision parameters for random effects close for INLA and MCMC techniques as we move away from the non-informative prior distribution (Gamma(1,0000.5)).

Based on information provided by R's system.time command, Table 7.4 compares the running times for INLA and MCMC. The R session's CPU time is indicated in column "user," the operating system's CPU time is indicated in column "system," and the wall clock time required to run the process is indicated in column "elapsed" [16]. Table 7.4 illustrates how much more computationally burdensome and time-consuming the MCMC approach was in determining its realizations compared to the INLA method. MAE and RMSE metrics were employed to assess the predictive performance of both the INLA and MCMC methodologies. The close proximity of these values suggests that the predictive capabilities of both models are comparable. Hence, considering the similarity in predictive power and the quicker generation of solutions by the INLA model, it implies that utilizing the INLA approach is more efficient for our specific research.

7.4.3 Crop Yield Insurance Premium Calculation

In this section, we will calculate the premium rates for crop yield insurance. Goodwin and Ker [9] propose using the equation below to calculate the premium rate π_r.

$$\pi_r = \frac{P(y < c\bar{y})\,[c\bar{y} - E(y|y < c\bar{y})]}{c\bar{y}} \tag{7.37}$$

In this equation, $P(y < c\bar{y})$ represents the probability that the realized yield is less than the prespecified threshold (yield loss level) $c\bar{y}$, where $c; 0 < c < 1$ is the insurer's coverage rate and $\bar{y}$ is the expected value of the yield, i.e. $\bar{y} = \int_{-\infty}^{\infty} yf(y)dy$. $E(y|y < c\bar{y})$ denotes the conditional expectation of realized yield if it is less than the yield threshold. The insured's premium is represented by the

numerator in Eq. (7.37). Divide the premium by the yield threshold to get the premium rate. The following are the mathematical formulas for $P(y < c\bar{y})$ and $E(y|y < c\bar{y})$.

$$P(y < c\bar{y}) = \int_0^{c\bar{y}} f(y)dy$$

$$E(y|y < c\bar{y}) = \frac{\int_0^{c\bar{y}} yf(y)dy}{\int_0^{c\bar{y}} f(y)dy}$$

We solve the above integrals using numerical methods. Assume the crop yield is normally distributed. The probability distribution function is then

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

where μ denotes the mean whereas σ is the standard deviation.

Using the selected model ((**Model 6**)), we compute the average premium rates associated with the specified coverage levels $c = 0.70, 0.80, 0.90$. The premium rates for the top two districts in terms of wheat production volume in Ankara and Konya are shown in the table below. The outcomes are presented separately for each city's top two districts.

The first two districts in Table 7.5, Polatlı and Haymana, are districts of Ankara, and they are ordered from highest to lowest in terms of wheat production in 2022. Cihanbeyli and Karatay, the third and fourth districts in Konya, are ranked highest to lowest in terms of wheat production in 2022. We assume that the crop yield insurance amount is the yield in kilos per decare. As a result, the premium rate is expressed as a percentage of the insurance amount. The premium rate is higher

Table 7.5 Premium rates in the districts chosen

District	Level of coverage (%)	Premium rate (%)
Polatlı	70	0.16
	80	0.73
	90	2.40
Haymana	70	0.33
	80	1.14
	90	3.07
Cihanbeyli	70	0.77
	80	1.97
	90	4.24
Karatay	70	5.22
	80	7.53
	90	10.39

for higher coverage levels. It is not possible to conclude that premium rates are solely related to wheat production volume. Because premium rates are calculated using spatial, temporal, and spatiotemporal characteristics, the relationship between premium rates and wheat production volume is not a single indicator of wheat yield risk. If an extreme weather event occurs in a given year, premium rates may change dramatically from 1 year to the next, according to our chosen model.

Table 7.5 shows that premium rates in Konya districts are higher than premium rates in Ankara districts. According to our chosen model, the wheat yield risk in Konya is higher than the risk in Ankara. Furthermore, among the four districts, Karatay has the highest premium rates based on three different coverage limits. The reason for this is that Karatay has consistently produced the highest wheat yield. Karatay's high yield suggests that wheat yield is more likely to remain below the threshold value in the event of damage.

7.5 Conclusion

By introducing the INLA model in this study, we ensure that complex models for agricultural insurance are analysed quickly in terms of computation time. We demonstrated the usability of the INLA model in this study by not using explanatory variables in the models and instead examining the effect of spatial and temporal effects on wheat yield. Other meteorological data affecting agricultural production, for example, can be added as explanatory variables for districts, and models with higher explanatory power can be established.

We propose using a hierarchical Bayesian method to account for not only spatial and temporal effects, but also the effect of spatiotemporal dependency among geographical subregions. According to the model results, the structured spatiotemporal interaction model is the best model. In addition to the modeling results, we provide a methodology for calculating the premium, which is a useful tool for assessing the risk of yield loss. The premium calculation results are obtained in relation to the specified coverage levels by using the selected model for the selected districts.

While the average wheat yield data for 47 districts was examined in this study, the study proceeded on the average yield because no data was available to directly observe the effects of rainfed and irrigated farming practices for the districts. If data on rainfed and irrigated agricultural areas for the districts, as well as farmer-based data, can be obtained in the future, models for this study can be developed by incorporating these effects as explanatory variables.

We can apply the INLA model to estimate crop yield based on farmers' data if we can acquire farmer-based data on crop yield and covariates like demographic, socioeconomic, and meteorological variables. Additionally, we can extend this methodology to other scenarios as well.

Finally, we hope to estimate the farmer's yield by improving a dependent aggregate claims model in the case of claim frequency and claim severity dependency.

Assuming that yield is a function of the farmer's surroundings and effort, we hope to obtain a risk score for the farmer and, as a result, a moral hazard map in a future study. As a result of the farmer-based premium calculation, we can obtain yield insurance more efficiently.

Author Contributions All authors contributed equally to the writing of the manuscript.

Conflicts of Interest No potential conflict of interest was reported by the author(s).

References

1. Awondo, S., Datta, G., Ramirez, O., & Fonsah, E. (2012). Estimation of crop yield distribution and insurance premium using shrinkage estimator: A hierarchical bayes and small area estimation approach. Technical Report.
2. Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine, 14*(21–22), 2433–2443.
3. Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics, 43*(1), 1–20.
4. Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.
5. Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with r-inla. *Spatial and Spatio-Temporal Epidemiology, 4*, 33–49.
6. Clayton, W. D. (1996). Generalized linear mixed models. In S. R. Gilks, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.
7. Dawid, A. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General), 147*(2), 278–290.
8. de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics, 26*(2), 403–413.
9. Goodwin, B., & Ker, A. (1998). Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. *American Journal of Agricultural Economics, 80*(1), 139–153.
10. Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: a comparison of mcmc and inla. In *Statistical modelling and regression structures* (pp. 91–110). Springer.
11. Jung, A., & Ramezani, C. (1999). Valuing risk management tools as complex derivatives: An application to revenue insurance. *Journal of Financial Engineering, 8*, 99–120.
12. Ker, A., & Coble, K. (2003) Modeling conditional yield densities. *American Journal of Agricultural Economics, 85*(2), 291–304.
13. Ker, A., & Goodwin, B. (2000). Nonparametric estimation of crop insurance rates revisited. *American Journal of Agricultural Economics, 82*(2), 463–478.
14. Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine, 19*(17–18), 2555–2567.
15. Martins, T., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with inla: New features. *Computational Statistics & Data Analysis, 67*, 68–83.
16. Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., & DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. *Spatial and Spatio-Temporal Epidemiology, 31*, 100301.

17. Nelson, C. H., & Preckel, P. V. (1989). The conditional beta distribution as a stochastic production function. *American Journal of Agricultural Economics, 71*(2), 370–378.
18. Ozaki, V., Ghosh, S., Goodwin, B., & Shirota, R. (2008). Spatio-temporal modeling of agricultural yield data with an application to pricing crop insurance contracts. *American Journal of Agricultural Economics, 90*(4), 951–961.
19. Park, E., Brorsen, B., & Harri, A. (2016). Using bayesian spatial smoothing and extreme value theory to develop area-yield crop insurance rating. Technical Report.
20. Park, E., Brorsen, B. W., & Harri, A. (2019). Using bayesian kriging for spatial smoothing in crop insurance rating. *American Journal of Agricultural Economics, 101*(1), 330–351.
21. Pereira, M., & Desassis, N. (2019). Efficient simulation of gaussian markov random fields by chebyshev polynomial approximation. *Spatial Statistics, 31*, 100359.
22. Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological), 52*(1), 175–184.
23. Riebler, A., Held, L., Rue, H., et al. (2012). Estimation and extrapolation of time trends in registry data-borrowing strength from related populations. *The Annals of Applied Statistics, 6*(1), 304–333.
24. Rodrigues, Â. (2017). *Spatio-temporal modelling of tornados with R-INLA, at the county-level in Texas and Ocklahona*. Ph.D. Thesis.
25. Rue, H., & Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of Statistical Planning and Inference, 137*(10), 3177–3192.
26. Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series b (Statistical Methodology), 71*(2), 319–392.
27. Ruiz-Cárdenas, R., & Krainski, E. T. (2010). Evaluating spatio-temporal models for crop yield forecasting using inla: implications to pricing area yield crop insurance contracts. http://www.ime.unicamp.br/sinape/sites/default/files/Ruiz-Krainski_paper_SINAPE2010.pdf
28. Ruiz-Cárdenas, R., Krainski, E. T., & Rue, H. (2012). Direct fitting of dynamic models using integrated nested laplace approximations-inla. *Computational Statistics & Data Analysis, 56*(6), 1808–1828.
29. Saengseedam, P., & Kantanantha, N. (2017). Spatio-temporal model for crop yield forecasting. *Journal of Applied Statistics, 44*(3), 427–440.
30. Şimşek, G. (2020). Impacts of spatıotemporal dependency and asymmetrıc informatıon on the analysıs of optımal crop yıeld insurance.
31. Sørbye, S., Illian, J., Simpson, D., Burslem, D., & Rue, H. (2019). Careful prior specification avoids incautious inference for log-gaussian cox point processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 68*(3), 543–564.
32. Spiegelhalter, D., Best, N., Carlin, B., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (Statistical Methodology), 64*(4), 583–639.
33. Stern, H., & Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine, 19*(17–18), 2377–2397.
34. Stokes, J. R. (2000). A derivative security approach to setting crop revenue coverage insurance premiums. *Journal of Agricultural and Resource Economics*, 159–176.
35. Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 44–47.
36. Tirupattur, V., Hauser, R. J., & Chaherli, N. M. (1996). Crop yield and price distributional effects on revenue hedging. OFOR Working Paper Series, no. 96–05.
37. Turvey, C., & Zhao, J. (1999). Parametric and non-parametric crop yield distributions and their effects on all-risk crop insurance premiums. Technical Report.

38. Urdangarin, A., Goicoa, T., & Ugarte, M. D. (2022). Space-time interactions in bayesian disease mapping with recent tools: Making things easier for practitioners. *Statistical Methods in Medical Research, 31*(6), 1085–1103.
39. Yildirak, K., Kalaylıoglu, Z., & Mermer, A. (2015). Bayesian estimation of crop yield function: drought based wheat prediction model for tigem farms. *Environmental and Ecological Statistics, 22*, 693–704.

Chapter 8
Model and Forecast Combination for Predictive Yield Distributions in Crop Insurance

Yong Liu, Austin Ford Ramsey, and Ziqin Zhou

Abstract Multiple-peril crop insurance policies require statistical modeling of probability distributions of crop yields. Unfortunately, no single parametric distribution is likely to capture the true data generating process. Likewise, non-parametric approaches converge to the true distribution at a slow rate; yield histories are often of modest size. Recognizing these shortcomings, model and forecast combination are now being applied in crop insurance settings. Model and forecast combination avoid the dangers inherent in selecting a single model for the predictive yield distribution. The component models for the combination can be selected ad-hoc or based on the idea of distributional similarity. Crop yields are spatially correlated, so the model for one insured unit may be related to another, and can then be used in the pool of potential models for the combination.

We briefly review the literature on model and forecast combination and its application in agricultural insurance settings. We then turn toward an empirical application involving crop yield insurance for major row crops in the U.S. Southeast. A variety of individual models and combinations are estimated at the county level. Insurance premiums and premium rates are calculated from the estimated distributions. Implications of model and forecast combination for insurance rates, premiums, and government subsidies are discussed. We conclude by suggesting future research in this area.

Y. Liu · Z. Zhou
Texas A&M, College Station, TX, USA
e-mail: yong.liu@ag.tamu.edu; ziqin.zhou@tamu.edu

A. Ford Ramsey (✉)
University of Georgia, Athens, GA, USA
e-mail: aframsey@uga.edu

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_8

8.1 Introduction

Multiple-peril crop insurance programs have proliferated worldwide. Such programs are often subsidized by national or local governments with the aim of stabilizing crop revenues and farm household incomes. In the United States, the federal crop insurance program is now the most expensive federal program providing subsidies to production agriculture [13]. The size of the U.S. program continues to grow; total liability has increased from $117 billion in 2012 to more than $173 billion in 2022 as more crops and agricultural producers are brought under federal crop insurance program policies. As liability and subsidies increase, the program has become the subject of discussions around opportunities to reduce program costs [10].

The continued expansion of subsidized crop insurance in the U.S. and the world has placed increased focus on the actuarial details underlying insurance policies. Aside from determining the indemnities that are paid out to agricultural producers, the actuarial models also affect government expenditures on these programs. Government outlays in the U.S. program are primarily for premium subsidies, whereby the government pays a portion of the premium, and for private insurers that operate the program. The U.S. program is a public-private partnership in that the United States Department of Agriculture determines premiums for the policies, but policies are sold and administered by private firms. In 2022, subsidies to agricultural producers (in the form of premium subsidies) amounted to just over $11 billion, while payments to private firms to operate the program (program delivery costs) were in the area of $2 billion. Clearly, the actuarial details of the programs have the potential to result in major differences in premium rates and expenditures on the part of taxpayers.

A key criteria in the U.S. federal crop insurance program is that policies should—at least attempt to—be actuarially fair. The premiums paid on crop insurance policies should match the risk exposure faced by farmers, at least to the extent that this is feasible. In other words, premiums should be equivalent to expected losses. Premiums that are actuarially fair avoid complications stemming from moral hazard and adverse selection. Moral hazard involves agricultural producers increasing their risk exposure because they do not bear the full cost of the increased risk. Adverse selection implies that the demand for agricultural insurance is dependent on a producer's risk. In any event, the basic problem resulting in moral hazard and adverse selection is information asymmetry.

Adverse selection and moral hazard are key problems for agricultural insurers. Santeramo [39] argues that moral hazard and adverse selection are the main factors that drive low participation in some crop insurance programs. An early study of the U.S. crop insurance program by [23] provides evidence of adverse selection. They found that farmers mainly participated in the program in order to receive subsidies or due to the possibility of adverse selection. In the contemporary program, the extremely high share of planted acres under federal crop insurance policies limits the negative impacts and potential death spiral from adverse selection. While most

studies focus on adverse selection, more recent work has identified advantageous selection in some crop insurance settings. He et al. [19] found evidence of advantageous selection for two types of crop insurance in the Philippines.

Actuarially fair prices may not completely mitigate the issue of moral hazard. A number of studies have examined crop insurance markets and found significant changes in producer behavior that are indicative of moral hazard [46]. Adequate design of crop insurance and reinsurance policies is one approach for controlling morally hazardous behavior. The design of contracts where both parties' exposure to risk increases with overall risk can mitigate moral hazard as shown in the reinsurance context in [2]. Assa et al. [3] uses a similar contract design to price frost insurance while effectively removing the risk of moral hazard. Controlling both adverse selection and moral hazard are necessary for maintaining viable crop insurance programs.

Changes in the system of rating for crop insurance policies in the U.S. have resulted in a healthier program since the mid-1990s. While average loss ratios exceeded one in the late 1980s, the program-wide loss ratio averaged just under 0.88 over the period 2011–2020. Some of the changes in loss ratios can be attributed to the increasing number of agricultural producers who place their crops under a federal crop insurance program policy. For major row crops, more than 90% of all planted acres are enrolled in a subsidized crop insurance policy. Most of the policies sold are based on losses to farm revenue or farm yields and are multiple-peril crop insurance policies: these policies protect against multiple causes of loss to crop production.

8.2 Actuarial Models for Crop Insurance Policies

Most multiple-peril crop insurance policies are priced based on data on historical crop yields, prices, and farm revenue. Historical data used to rate the policies may be collected—and policies may be sold—at the farm level or the county level. The principal policies sold through the U.S. federal crop insurance program are revenue policies and yield policies. Yield policies pay out when crop yields are low and revenue policies pay out when crop revenue is low; revenue policies account for the additional risk of declining prices over the growing season. The most popular policy is termed Revenue Protection (RP) and is a revenue policy based on farm-level crop yields. In 2022, RP policies comprised just over 70% of all liability in the federal crop insurance program.

Revenue insurance policies require a probability distribution for revenue. As revenue is the product of quantity and price, this probability distribution is constructed as a joint probability distribution over crop yields and crop prices. Yield insurance policies only require a probability distribution over crop yields as the payout price is known at the time the policy is sold. The key difference from a rating perspective is that prices are stochastic for revenue insurance policies. The loss from a revenue policy in the U.S. crop insurance program is

$$\text{Loss} = \max(0, Y_P P_P \lambda - Y_H P_H), \tag{8.1}$$

where Y_P and P_P are planting-time yield and price, respectively, which are both known at the time the policy is sold. Y_H and P_H are harvest-time—or realized—yield and price. λ is a coverage level between 0 and 1. In practice, most federal crop insurance policies allow for selection of a coverage level between 0.55 and 0.85. Only two variables in Eq. (8.1) are stochastic and the policy can be rated by specifying their joint distribution as

$$p(Y_H, P_H) = f(Y_H, P_H). \tag{8.2}$$

In practice, the joint distribution is constructed by combining marginal distributions for the two variables. This can be achieved using the distribution-free approach of [22] or copulas following the work of [42].

For yield insurance policies, the loss is given by

$$\text{Loss} = \max(0, Y_P P_P \lambda - Y_H P_P), \tag{8.3}$$

with the price being deterministic. The only information necessary to price the policy concerns the univariate probability distribution

$$p(Y_H) = f(Y_H). \tag{8.4}$$

Although revenue insurance is more popular compared to yield insurance, we focus on yield insurance as this allows for a clear picture of modeling issues related to the yield distribution. A number of authors have considered the construction of joint distributions for revenue insurance including [14, 36], and [38]. Many of the concerns addressed in this chapter apply equally to revenue insurance because the distribution of revenue is ultimately a joint distribution of yield and price.

An actuarially fair price for an insurance policy should result in an average loss ratio near one. The loss ratio is defined as

$$\text{LR} = \frac{\text{Loss}}{\text{Premium}} \tag{8.5}$$

and if the prices are actuarially fair then the expected loss equals the premium. This implies that the expected loss ratio is one. Loss ratios that deviate, on average, from one typically indicate departures from the actuarially fair price. Although loss ratios in the U.S. crop insurance program overall have averaged near one in the past decade, there has been heterogeneity across crops and locations.

Loss ratios averaging over one would indicate that the policies are likely to be under-priced, while loss ratios averaging under one would indicate that policies are over-priced. There are several important implications of mis-pricing for subsidized crop insurance programs. High average loss ratios would indicate large payouts in excess of premiums collected. In such situations, the total amount of premium

subsidies paid by the government could be lower, but payments of indemnities or other government backstops of insurance could be exceptionally large. If loss ratios are lower on average, then participation in the program may decline leading to more pronounced problems of adverse selection.

The motivating question is how to construct $f(Y_H)$ in such a way that it accurately represents the risks faced by the producer and comes as close as possible to generating premium rates that are actuarially fair. Ideally, we would use a model that exactly matches the data generating process. No such model is available from statistical or economic theory, which means that this is a question of model specification. Specifying the appropriate model for $f(\cdot)$ is not so simple because the data used to estimate the distribution are typically of moderate size. Individual farms may only have yield records going back several decades, while many county yield records only go back to the middle of the twentieth century.

This general lack of yield data complicates modeling of the yield distribution. Many early crop insurance and agricultural production studies utilized parametric distributions. Nelson [30] compared the normal distribution and beta distribution in calculating crop insurance premium rates. Gallagher [9] utilized a gamma distribution. Sherrick et al. [40] found the Weibull distribution to provide a good fit to corn yields. Claassen and Just [6] applied the inverse lognormal distribution to short series of farm-level yields. Park et al. [32] fit generalized Pareto distributions: a type of extreme value distribution. A disadvantage of parametric distributions is that there is no guarantee that the chosen parametric distribution encompasses the true data generating process. While parametric models converge quickly to the data generating process, this is only true if the correct parametric model has been specified.

In contrast to parametric models, non-parametric approaches make less stringent assumptions. Although non-parametric approaches converge to the true model at a slower rate, they are able to approximate any distribution to an arbitrary degree given a large enough sample size. Non-parametric estimation of yield distributions in the crop insurance context was proposed by [15] and refined in [25]. Non-parametric density estimation also serves as the basis for spatio-temporal model combination implemented in [28]. Some semi-parametric approaches were discussed by [24]. Parametric and non-parametric models for crop yields were compared by [31] in an application to Brazilian corn, soybeans, and wheat. A challenge for most non-parametric methods is the moderate size of yield data and the fact that yields are non-negative. Standard kernel density estimation approaches usually do not account for cases where the random variable has a bounded support.

Regardless of whether the underlying model is parametric or non-parametric, most methods of density estimation require data that are independent and identically distributed. These criteria pose a problem for crop insurance programs as yield distributions change over time. Arata et al. [1] indicates that worldwide, nearly half country-crop combinations show slowdowns in yield growth, while a quarter of the combinations considered show an increase in yield variability. Most studies first de-trend the observed yields, or make other adjustments, and then estimate the yield distributions on a set of normalized yields [18, 41]. Other approaches to

structural change in yield distributions include mixtures with different trends in the components as in [43], time-varying distributions as in [48], and the use of quantile regression as in [35].

As shown above, modeling of the yield distribution is of primary importance for multiple-peril crop insurance policies. The increasing demand for revenue insurance in the United States has also placed importance on modeling of the distribution of crop revenue. Many of the same modeling concerns that arise in the yield distribution characterize the distribution of revenue. Typically, prices in the federal crop insurance program are assumed to align with assumptions of the Black-Scholes model, although various alternative distributions were considered by [16]. The implications of the Black-Scholes model are used in deriving the variance of expected prices. An examination of the use of Black-Scholes for constructing this measure in crop revenue insurance was conducted by [17]. They found that the Black-Scholes model was preferred to several alternatives. While [5] found the volatility factor used in the federal crop insurance program to be an unbiased estimator of realized price volatility, they suggested improvements that could be made to better measure price risk over the insurance period. Regardless of how the distribution of prices is modeled, marginal distributions for prices and yields can be combined with copulas to obtain the joint distribution of revenue as in [38]. However, because prices are observed at the same frequency as yields, models for revenue insurance still require strong assumptions given relatively small samples [21].

In addition to modeling distributions of yields from a single county or unit of interest, some authors have studied how yield distributions could be spatially smoothed or related through a spatial model. Park et al. [32] used Bayesian kriging to smooth crop yield densities across counties. This approach was expanded on by [33] who, instead of smoothing across physical space, smoothed densities across climate space. They found that smoothing in climate space proved superior to smoothing in physical space, particularly in locations with missing data or varying climate. An approach for interpolating missing yield data through the use of Bayesian hierarchical models was developed by [34]. A frequentist approach allowing a flexible spatial structure through the use of nonparametric spatially-varying coefficients was applied by [45]. The idea of borrowing information from other units, or smoothing yield distributions across space, relates to spatial correlation in the weather conditions that cause yield losses.

8.3 Model and Forecast Combination for Predictive Yield Distributions

There is an inherent danger in choosing any single model for an analysis. Parametric models are likely to be mis-specified. If this is the case, they will not converge to the true data generating process and result in biased estimates. This danger is widely

recognized in the time series literature in the context of forecasting. Small samples exacerbate the problem. An early review of forecast combination techniques was conducted by [7], while a modern discussion is available in [44]. A review of model averaging in economics can be found in [29]. Model and forecast combination are directly related to the actuarial modeling of crop yield distributions because the yield distribution used in crop insurance policies is a forecast.

A forecast is a prediction about the future and is typically (although not always) derived from a statistical model. Forecast combination is distinguished from model averaging in that the focus of model averaging is on dealing with model uncertainty and avoiding model selection. Forecast combination is the combination of forecasts arising from different models. The concept is more general because one of the component forecasts in a combination could be obtained from a model average. Forecasts could also consist of expert opinions for which no statistical model can be obtained. In spite of this subtlety in their focus, both model and forecast combination aim to produce more reliable and accurate forecasts.

Recently, a number of authors have turned to model or forecast combination as a means of obtaining more accurate yield distributions, and therefore more accurate insurance rates. We focus on two approaches for model or forecast combination in this study. The first is a form of Bayesian model averaging, specifically Bayesian averaging of frequentist estimates, applied in the crop insurance setting by [26]. The second is a linear pooling procedure implemented in the crop insurance context by [37]. These represent two different approaches to the issue of model specification and choice in an insurance rating setting. Bayesian averaging of frequentist estimates is formally an approach for model combination, while linear pooling combines forecasts from different models.

Concepts of model and forecast combination can be used to pool information from different units by combining their models or forecasts. The aim is to bring information from other units to bear in estimating the predictive crop yield distribution for the unit in question. The difference with standard applications of model or forecast combination is that the alternative models or forecasts in the set of candidates are estimated on other units. Data from other units are known to be related to the unit in question; in this case, yield distributions in counties that are geographically proximate are likely to be similar due to correlation in weather, soil features, and other aspects of crop production. Therefore, the models estimated on data from other counties may serve as a reasonable set of candidate models for model or forecast combination. This observation motivates the analysis to follow. Individual models and forecasts are generated for each county in question using that county's data. These models or forecasts are then considered as the set of component forecasts for every county and combined with the two combination approaches discussed below. The result is a combined model or forecast that is unique for every county under consideration, even though the set of component models or forecasts is the same.

8.3.1 Bayesian Averaging of Frequentist Estimates

The first combination scheme considered is an analogue to Bayesian model averaging. Consider the case where there are J models and we are interested in estimating a parameter or other functional θ. Without loss of generality, θ could also be a vector of parameters. In the context of density estimation for agricultural insurance, θ encompasses the parameters of the yield distribution. For instance, in the case of a normal distribution, $\theta = (\mu, \sigma^2)$. The posterior distribution of θ given data X is given by

$$p(\theta|X) = \sum_{i=1}^{J} p(\theta|M_i, X)p(M_i|X), \tag{8.6}$$

where M_i denotes model i. The posterior distribution of θ is the sum of the posterior distribution of θ under each model multiplied by the posterior probability of the model itself. The posterior distributions of θ under each model are readily obtained by estimating the individual models. The posterior model probabilities are given by

$$p(M_i|X) = \frac{p(X|M_i)p(M_i)}{\sum_{i=1}^{J} p(X|M_i)p(M_i)}, \tag{8.7}$$

and the computation of this probability involves evaluation of potentially high dimensional integrals. The preceding discussion is nothing more than Bayesian model averaging and is discussed in some detail in [20].

A key feature of the approach detailed in [26] is that the underlying models are mixtures of normal distributions. This simplifies, considerably, the estimation of $p(X|M_i)$ as

$$p(X|M_i) = \int p(X|\tau_i, M_i)p(\tau_i|M_i)d\tau_i, \tag{8.8}$$

where τ_i are the parameters associated with model i. In the case where the underlying models are mixtures of normals, τ_i would include the means and variances of the individual normal mixture components and the mixing weights. Ker et al. [26] rely on a result shown in [8] where $p(X|M_i)$ can be approximated by the Bayesian Information Criterion associated with each model. In particular,

$$p(X|M_i) = \exp(-1/2BIC_i), \tag{8.9}$$

where BIC_i is the Bayesian Information Criterion for model i. Assuming all candidate models have equal prior model probability, the weights on the individual models in Eq. (8.7) are then given by

$$p(M_i|X) = \frac{\exp(-1/2BIC_i)}{\sum_{i=1}^{J}\exp(-1/2BIC_i)}. \tag{8.10}$$

The resulting model averaged distribution is given by the sum of the individual distributions each weighted by $p(M_i|X)$ as shown in Eq. (8.6). This procedure is referred to here are Bayesian averaging of frequentist estimates because the underlying models are frequentist and do not incorporate prior probabilities in estimation of their parameters. The underlying models in [26] are based on a mixture of two normal distributions with trends in means as applied in [43].

Bayesian model averaging is based on an in-sample measure of model fit: Bayesian Information Criteria in this case. Another potential issue with Bayesian model averaging is that it assumes that the true model is a member of the set of models under consideration. Yao et al. [47] state that Bayesian model averaging is inappropriate if the data generating process is not one of the models under consideration. Under such a situation, Bayesian model averaging asymptotically converges to whichever model in the set of component models is closest in Kullback–Leibler divergence. They suggest an alternative to Bayesian model averaging which they refer to as stacking of Bayesian predictive distributions.

8.3.2 Linear Pooling

Bayesian model averaging combines models and is based on a measure of in-sample fit. As noted above, [26] essentially weight models based on their Bayesian Information Criterion. One can also consider out-of-sample fit as a criteria for combining models or forecasts. Such is the case for linear pooling. The idea behind linear pooling was developed in [11] and later applied in macroeconomic forecasting by [12]. In linear pooling, the log predictive score function is used to construct the combination of forecasts. The log predictive score function is, as we show later, one possible criteria for evaluating forecast accuracy. Again, suppose that we have J possible models and that the forecasts from these models can be combined into a single forecast. We are interested in the predictive density of crop yields; i.e. we are combining density forecasts, not point forecasts.

The combined predictive density can be represented by the following

$$\sum_{i=1}^{J} w_i f(x_t|X_{t-1}, M_i), \tag{8.11}$$

where $\sum_{i=1}^{n} w_i = 1$ and $w_i \geq 0$ for all i. Thus the density is a linear combination and finite mixture of the individual predictive distributions with weights w_i. The predictive density is for the variable of interest x at time t and is based on all information available which is given in X_{t-1}. According to [11], the term linear prediction pool is coined for this mixture by [4]. They also suggest that the finite

mixture of Eq. (8.11) can be evaluated with a log score function as

$$\sum_{t=1}^{T} \log \sum_{i=1}^{J} w_i f(x_t | X_{t-1}, M_i). \tag{8.12}$$

The log score shown in Eq. (8.12) is the sum of log scores for the individual models. The log predictive score function for the single model in county i is given by

$$LS(X_T, M_i) = \sum_{t=1}^{T} \log f(x_t | X_{t-1}, M_i), \tag{8.13}$$

and is an evaluation of the predictive accuracy of the model at T points in time. Taking the log predictive score function as a measure of forecast ability, the natural step is to choose weights w_i to maximize the log score given in Eq. (8.12). Doing so maximizes the forecast ability (given the log score as a measure of this ability) of the mixture of component forecasts.

8.4 Combining Densities for Multiple-Peril Yield Insurance

We illustrate the application of the model combination techniques by applying them to forecast yield densities for three crops: upland cotton, peanuts, and flue-cured tobacco. Our choice of these crops is motivated by the observation that most published research in crop insurance is geared toward corn and soybeans. However, sizeable amounts of the other major row crops (cotton, peanuts, and tobacco) are under federal crop insurance policies. These crops have also differed from corn and soybeans in terms of average loss ratios. Flue-cured tobacco, in particular, had an average loss ratio well above one over the last decade.

Yield data were obtained from the National Agricultural Statistics Service for the period from 1955–2020 where available. In the case of tobacco, the data only run through 2000. We selected major producing states in the Southeast with enough counties with complete yield information to facilitate the analysis. Based on these criteria, we modeled cotton yield distributions in Arkansas, Georgia, Louisiana, Mississippi, and Tennessee; peanuts in Alabama, Georgia, and North Carolina, and flue-cured tobacco in North Carolina.

Before fitting the county-level individual model, we first detrended the observed yields for each county using a local linear non-parametric regression and addressed potential heteroscedasticity in the residuals using the method introduced by [18]. Denote the residuals from the local linear non-parametric regression as $\hat{\epsilon}_t$. Then the following equation can be estimated to determine the degree of heteroskedasticity

$$\ln(\hat{\epsilon}_t) = \alpha + \gamma \ln \hat{Y}_t + \eta_t \tag{8.14}$$

where γ is the form of heteroskedasticity and takes a value of zero when constant variance and two when the variance is proportional. Yields are then adjusted one step ahead such that

$$Y_t^* = Y_{T+1} + \hat{\epsilon}_T \frac{Y_{T+1}}{Y_t}^{\hat{\gamma}/2} \tag{8.15}$$

The resulting yields Y^*, referred to as adjusted yields hereafter, are assumed to follow a normal distribution and are treated as identically and independently distributed, aligning with common practice in the literature. These adjusted yields are utilized for estimating the corresponding parameters of the yield distributions. The underlying normal distributions are later combined with Bayesian model averaging and linear pooling.

The underlying yields are treated as independent and identically distributed only within each county. This assumption is used to justify the estimation of the component models in each county. In other words, we are assuming that after detrending and adjusting for heteroskedasticity, the data in each county are independently drawn from the same distribution. This assumption is typically required to obtain consistent estimated of the density using standard methods whether parametric or non-parametric. However, even if the yield observations are independent and identically distributed, we have no guarantee that the assumption of the normal distribution for the conditional yield distribution is correct.

Table 8.1 shows summary statistics of the yields for each crop. Variation in the raw yields is large, particularly in cotton and peanut where there have been substantial changes to yields over the period. In contrast, there is less variation in tobacco yields as tobacco has not received as much attention in terms of research and development of yield improving technology. The tobacco program, a federal policy that operated until the early 2000s, was a marketing quota. Farmers wishing to market tobacco were required to hold quota and in many cases, quota could only be sold or rented out within a certain geographic area. This program reduced incentives for growers to undertake the implementation of technologies that would

Table 8.1 Summary statistics of crop yields

		Raw yields			Adjusted yields		
Crop	Counties	Mean	Median	S.D.	Mean	Median	S.D.
Cotton	49	671.4	639.5	236.4	1017	1024	234.8
Peanut	22	2732.2	2761.0	968.3	4140	4247	962.8
Tobacco	23	1855.7	1822.5	322.6	2392	2409	354.8

Note: cotton and peanut data cover the period from 1955 to 2020, while tobacco data span from 1955 to 2000. Adjusted yields are calculated by detrending the raw yields, then adjusted for the heteroscedasticity to the most recent year for each crop. "S.D." is an abbreviation for "standard deviation"

lead to major yield enhancements [27]. The adjusted yields in Table 8.1 refer to the detrended yields used to form the probability distributions for the insurance policies. As the yields are projected at the end of the time series, the mean and median adjusted yields are substantially larger than those statistics for the raw yields. There is still substantial variation in the adjusted yields for all three crops.

To assess the performance of our models, we employ the most recent 30 years of data as a test set. Specifically, for cotton and peanut data, we initiate the individual model estimation using data from 1955 to 1990. Subsequently, we forecast the one-year ahead yield density in 1991, in line with the typical requirements of crop insurance programs for next-year yield forecasts. This one-year ahead estimation and forecasting process continues through the test set, concluding with density estimation in 2020 using data spanning from 1955 to 2019. The tobacco data undergo the same procedure, with the only distinction being the data span from 1955 to 2000. In each year of the test set, we construct Bayesian model averaging and linear pooling combinations using all county models within each crop category. Bayesian model averaging weights are calculated using Eq. (8.10), while the weights of linear pooling are determined by maximizing Eq. (8.12). Note that the maximization required for the linear pooling density is essentially a convex optimization problem and can be implemented using widely available statistical software.

Having combined the individual county models according to Bayesian model averaging and linear pooling, the log scores can be evaluated in each year. Figure 8.1 shows the mean predictive log-scores (across all counties) in each year. The average log scores accruing to the models based on individual yield densities and Bayesian model averaging are surprisingly similar. We found that Bayesian model averaging tends to place a large weight on individual models in the density combination. This may not been surprising as Bayesian model averaging converges asymptotically to the best-fitting model in terms of Kullback-Leibler divergence. Clearly, the best-fitting model in-sample will be the individual model estimated on own data. In contrast, the model constructed by linear pooling occasionally performs better than either the individual or Bayesian model averaging approaches. While differences are minimal in most years, there are occasional large differences. The overall superiority of linear pooling in this case may not be surprising given that it optimizes the weights to achieve the maximum log score.

Figures 8.2 and 8.3 show the weights that are assigned to different counties by the Bayesian model averaging and linear pooling. Figure 8.2 shows the weight for three selected counties in 2021 (for cotton and peanuts) and 2001 (for tobacco). The county of interest is highlighted in black. Both BMA and linear pooling tend to place a high amount of weight on the county of interest (i.e. the own county component model) for these counties. Figure 8.3 shows the average weight on the own county component model across all years. BMA tends to place larger weight on the own county model; not surprising given that BMA is averaging across models using an in-sample measure of fit. Linear pooling, on the other hand, tends to place less weight on the own county model. The lack of weight for the own county could

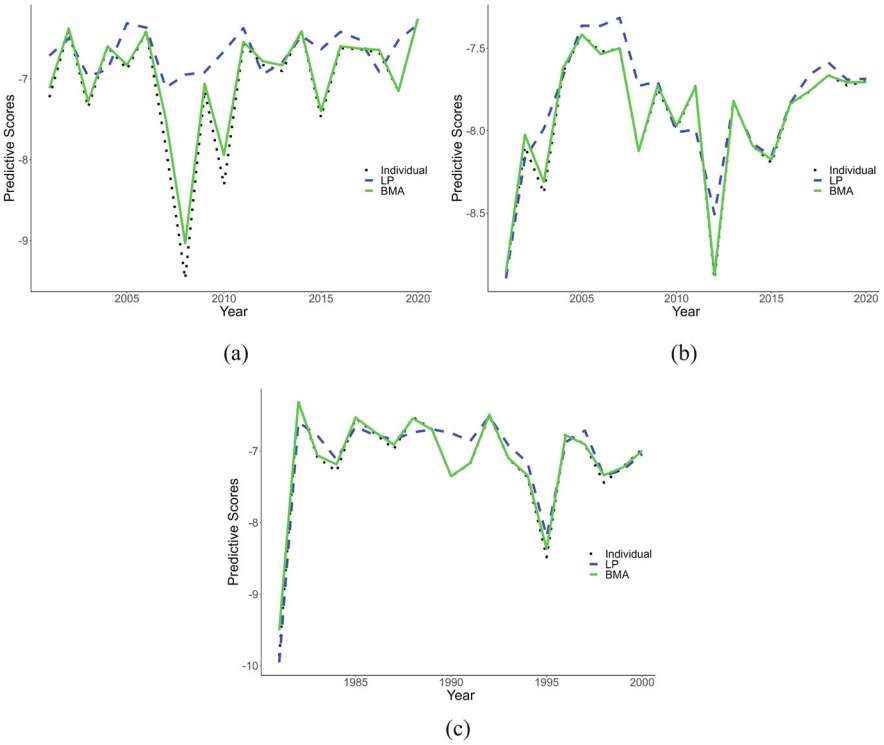


Fig. 8.1 Yearly Mean Predictive Log-scores by Crops *Note:* The predictive log scores are computed as the annual averages of all counties within each crop category. (**a**) Cotton: 2001–2020. (**b**) Peanut: 2001–2020. (**c**) Tobacco: 1981–2000

result from the out of sample measure of fit along with the somewhat restrictive Gaussian model used for the underlying yield distributions.

The practical impact of pooling models or forecasts is best examined by comparing the premium rates for yield insurance that would prevail under the different actuarial approaches. For simplicity, we consider the rates that would be generated for a yield insurance policy with a coverage level of 0.85. Figure 8.4 shows the difference in the premium rates between Bayesian model averaging, linear pooling, and a rate derived in a similar way to the calculation of rates in the federal crop insurance program. This rate, which we denote as the Risk Management Agency (RMA) rate, is simply calculated using the mean loss implied by the adjusted yields. As is evident from the maps, the differences are relatively minor for Bayesian model averaging. Differences tend to be larger for linear pooling. Linear pooling may better capture extreme losses in the left tail of the distributions, thus resulting in the much larger rates observed in cotton and tobacco.

Figure 8.5 shows premium rates for 2021 and 2001 (the final predicted year in the samples). We find that there are only minor differences between the rates produced

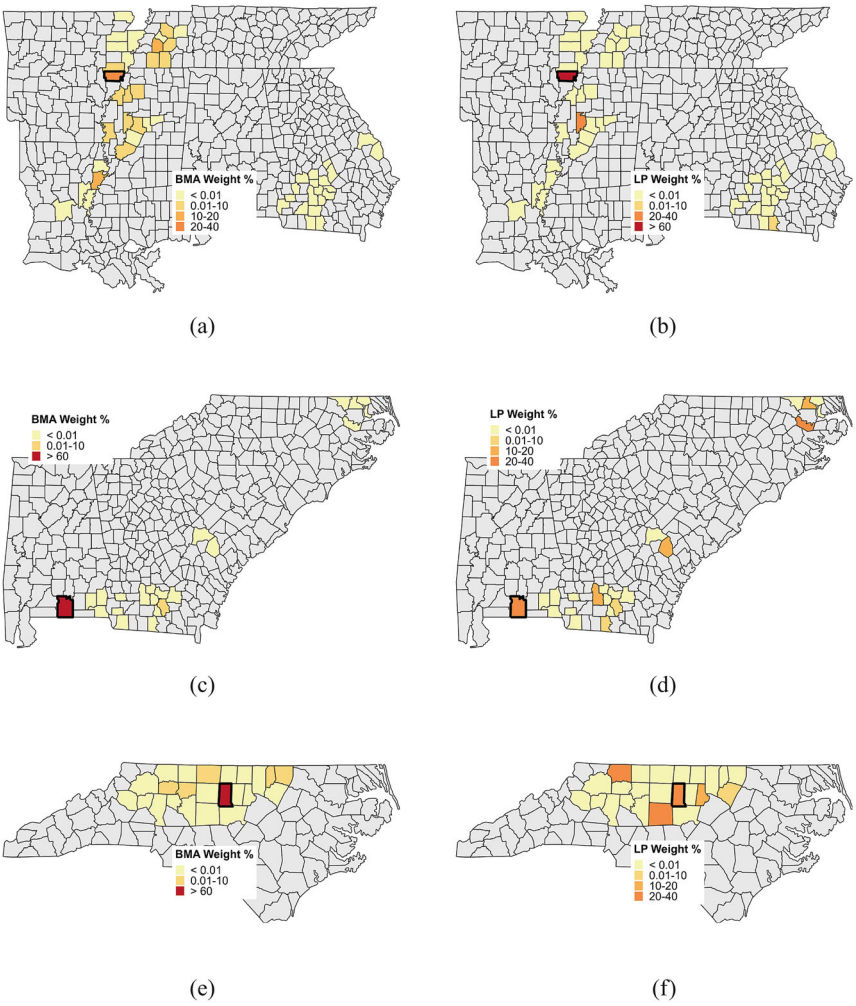


Fig. 8.2 Maps of Weights for Selected Counties *Note:* The maps illustrate the geographic distribution of weights for the forecast combination using Bayesian Model Averaging (BMA) and Linear Pooling (LP) across three representative counties, with the selected county outlined in black. (**a**) Cotton: 2021 BMA, Lee Cty, AR. (**b**) Cotton: 2021 LP, Lee Cty, AR. (**c**) Peanut: 2021 BMA, Covington Cty, AL. (**d**) Peanut: 2021 LP, Covington Cty, AL. (**e**) Tobacco: 2001 BMA, Alamance Cty, NC. (**f**) Tobacco: 2001 LP, Alamance Cty, NC

by the individual models and the RMA approach based on the empirical distribution. This also seems to be the case with Bayesian model averaging. The major difference comes from linear pooling. In general, rates are smaller for peanuts and cotton, although there are some outliers for cotton. The rates tend to be much higher for tobacco. These observations accord with some stylized facts about production of these commodities. Cotton and peanuts tend to have less yield risk compared

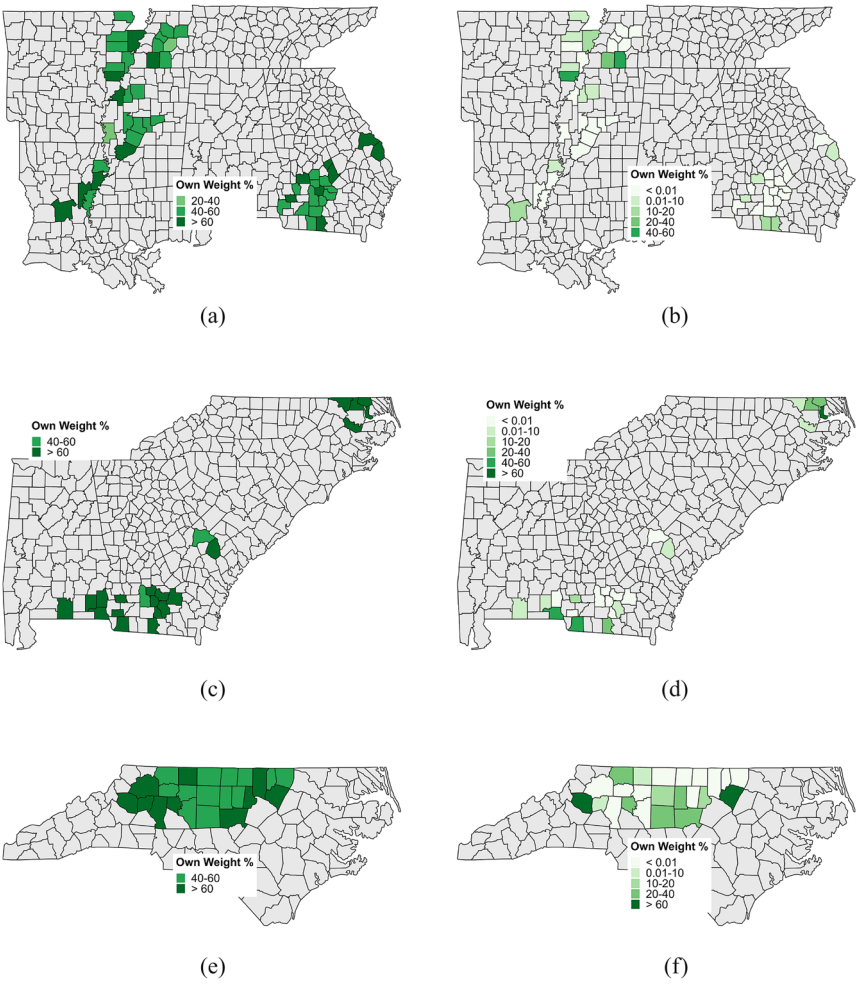


Fig. 8.3 Maps of Average Weight on Own County *Note:* The maps display the temporal average of weights from each county's own model by crop, derived from the forecast combination using Bayesian Model Averaging (BMA) and Linear Pooling (LP). (**a**) Cotton: BMA. (**b**) Cotton: LP. (**c**) Peanut: BMA. (**d**) Peanut: LP. (**e**) Tobacco: BMA. (**f**) Tobacco: LP

to other commodities, with price risk being a more major concern. On the other hand, tobacco has tended to have especially high loss ratios in the federal crop insurance program, possibly suggesting that some aspects of yield risk are not properly accounted for in the simple models estimated here.

Overall, we find that model and forecast combination can result in premium rates that differ substantially from individual parametric distributions. These differences are most pronounced for forecast combination techniques that rely on out-of-sample predictive accuracy as a metric for combining individual models. Model and

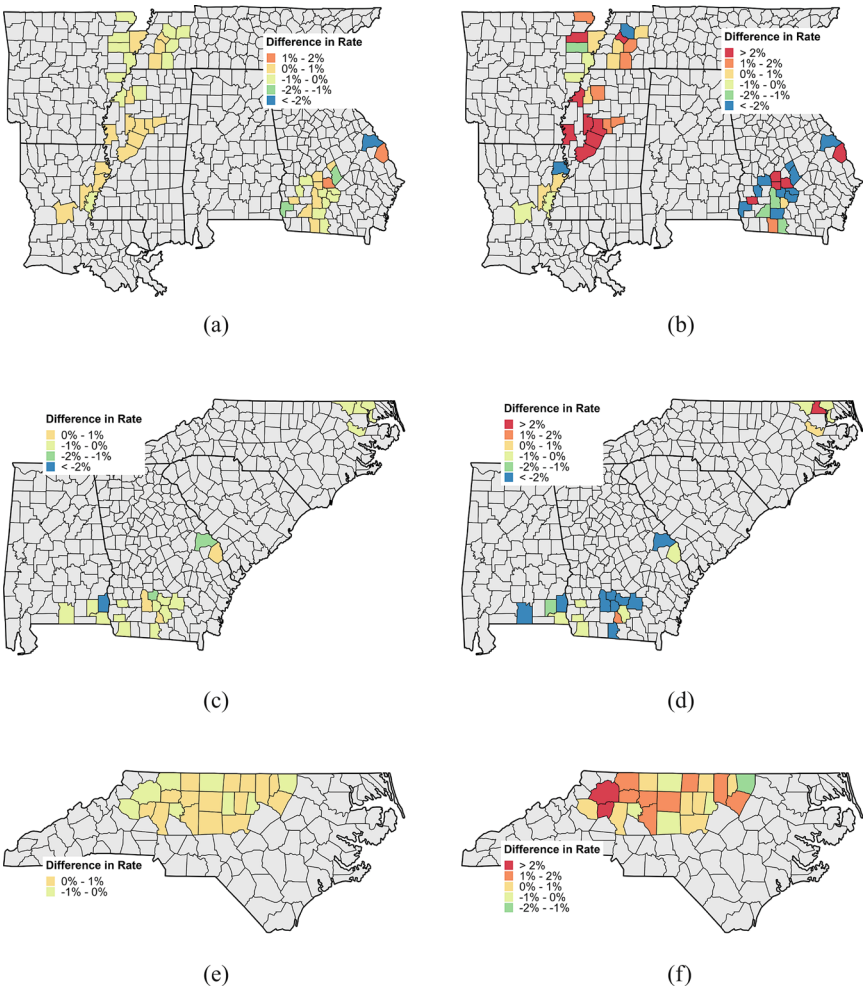


Fig. 8.4 Maps of Predicted Premium Rate Differences *Note:* The maps depict the percentage differences in premium rates between the proposed methods and the currently employed empirical rates by the RMA, specifically at an 85% coverage level. The RMA's empirical rates are calculated as the mean loss based on the adjusted yields. (**a**) BMA and RMA: Cotton 2021. (**b**) LP and RMA: Cotton 2021. (**c**) BMA and RMA: Peanut 2021. (**d**) LP and RMA: Peanut 2021. (**e**) BMA and RMA: Tobacco 2001. (**f**) LP and RMA: Tobacco 2001

forecast combination provide increased flexibility for estimating yield distributions, thus allowing the analyst to dispense with the strong assumptions that typically accompany parametric models of yield distributions.

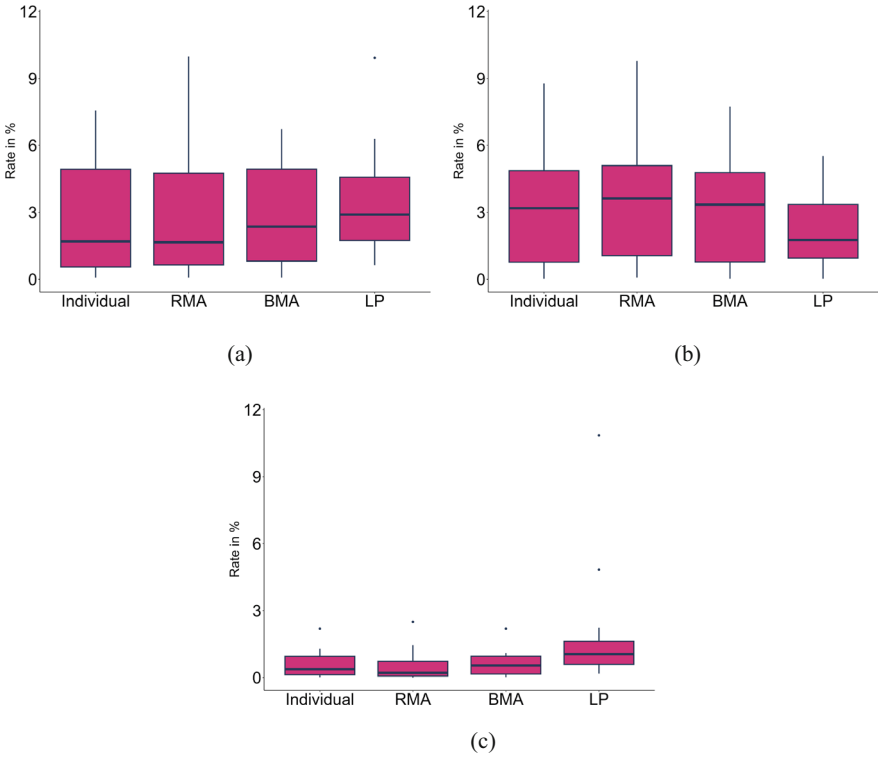


Fig. 8.5 Box-plots of Estimated Premium Rates by Method. *Note:* The estimated premium rates are of all counties in each crop at the coverage level of 85%. (**a**) Cotton Premium Rates 2021. (**b**) Peanut Premium Rates 2021. (**c**) Tobacco Premium Rates 2001

8.5 Conclusion

This chapter evaluates two methods for model or forecast combination in generating predictive crop yield distributions for crop insurance. The predictive yield distribution is a key statistical construct underlying the pricing of most multiple-peril crop insurance policies. Because subsidized multiple-peril crop insurance is becoming more popular worldwide, estimation and modeling of yield distributions has taken on new importance. Small changes in the way that the distributions are constructed can have sizeable impacts on government outlays and indemnities paid to agricultural producers.

There is little theoretical direction for models of the yield distribution. Any parametric model is likely to be mis-specified, although the effects of this mis-specification are not obvious and usually require empirical investigation. Limited data is available which also hinders the use of non-parametric approaches. To avoid the dangers of selecting a single model for evaluation of the yield distribution,

several authors have applied model or forecast combination in the crop insurance context.

We apply Bayesian model averaging and linear pooling to combine yield density forecasts across counties. The models in the pools of component forecasts are models estimated on nearby counties. The exercise considers rating insurance policies for upland cotton, peanuts, and flue-cured tobacco. Rates differed by county for all three crops. Linear pooling generated rates that were much higher in the case of tobacco. Because of the expanded pool of models, both Bayesian model averaging and linear pooling have the potential to better capture non-normal aspects of the yield distribution. This may result in improved modeling of the left tail of the distribution which is typically of most interest in crop insurance applications.

Several issues remain for further research. There has not been an evaluation of the forecast combination paradox in this context. This paradox has to do with the observation that simple forecast combination methods (such as averaging) can, and often do, outperform more complicated approaches. Model and forecast combination appear to have mostly been applied to the same underlying statistical model estimated on different geographical units. There is no reason why these methods could not be applied to different statistical models for the same unit, or for different units. The reasoning for combining models from nearby units is that the units are typically subject to similar weather conditions. However, as discussed in [11], forecasts that do not have superior fit overall, sometimes receive significant weight in a forecast combination. In any event, there are several avenues for improved crop insurance rating through model and forecast combination. Computational advances now allow for the routine application of such procedures, bringing with them the potential for better actuarial methods.

References

1. Arata, L., Fabrizi, E., & Sckokai, P. (2020). A worldwide analysis of trend in crop yields and yield variability: Evidence from fao data. *Economic Modelling, 90*, 190–208.
2. Assa, H. (2015). On optimal reinsurance policy with distortion risk measures and premiums. *Insurance: Mathematics and Economics, 61*, 70–75.
3. Assa, H., Wang, M., & Pantelous, A. A. (2018). Modeling frost losses: Application to pricing frost insurance. *North American Actuarial Journal*, 22(1), 137–159.
4. Bacharach, J. (1974). *Bayesian dialogues*. Christ Church College, Oxford University. Unpublished manuscript.
5. Bulut, H., & Fortenbery, T. R. (2022). Reevaluating the use of volatility factor in crop insurance premium rating. *Journal of the Agricultural and Applied Economics Association, 1*(2), 124–135.
6. Claassen, R., & Just, R. E. (2011). Heterogeneity and distributional form of farm-level yields. *American Journal of Agricultural Economics, 93*(1), 144–160.
7. Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559–583.
8. Dasgupta, A., & Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association, 93*(441), 294–302.
9. Gallagher, P. (1987). Us soybean yields: estimation and forecasting with nonsymmetric disturbances. *American Journal of Agricultural Economics, 69*(4), 796–803.

10. GAO (2023). *Crop insurance: Update on opportunities to reduce program costs*. United States Government Accountability Office. https://www.gao.gov/products/gao-24-106086
11. Geweke, J., & Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics, 164*(1), 130–141.
12. Geweke, J., & Amisano, G. (2012). Prediction with misspecified models. *American Economic Review, 102*(3), 482–486.
13. Glauber, J. W. (2013). The growth of the federal crop insurance program, 1990–2011. *American Journal of Agricultural Economics, 95*(2), 482–488.
14. Goodwin, B. K., & Hungerford, A. (2015). Copula-based models of systemic risk in us agriculture: Implications for crop insurance and reinsurance contracts. *American Journal of Agricultural Economics, 97*(3), 879–896.
15. Goodwin, B. K., & Ker, A. P. (1998). Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. *American Journal of Agricultural Economics, 80*(1), 139–153.
16. Goodwin, B. K., Roberts, M. C., & Coble, K. H. (2000). Measurement of price risk in revenue insurance: implications of distributional assumptions. *Journal of Agricultural and Resource Economics*, 195–214.
17. Goodwin, B. K., Harri, A., Rejesus, R. M., & Coble, K. H. (2018). Measuring price risk in rating revenue coverage: Bs or no bs? *American Journal of Agricultural Economics, 100*(2), 456–478.
18. Harri, A., Coble, K. H., Ker, A. P., & Goodwin, B. J. (2011). Relaxing heteroscedasticity assumptions in area-yield crop insurance rating. *American Journal of Agricultural Economics, 93*(3), 707–717.
19. He, J., Rejesus, R., Zheng, X., & Yorobe, Jr. J. (2018). Advantageous selection in crop insurance: Theory and evidence. *Journal of Agricultural Economics, 69*(3), 646–668.
20. Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical Science, 14*(4), 382–417.
21. Hungerford, A., & Goodwin, B. (2014). Big assumptions for small samples in crop insurance. *Agricultural Finance Review, 74*(4), 477–491.
22. Iman, R. L., & Conover, W.-J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation, 11*(3), 311–334.
23. Just, R. E., Calvin, L., & Quiggin, J. (1999). Adverse selection in crop insurance: Actuarial and asymmetric information incentives. *American Journal of Agricultural Economics, 81*(4), 834–849.
24. Ker, A. P., & Coble, K. (2003). Modeling conditional yield densities. *American Journal of Agricultural Economics, 85*(2), 291–304.
25. Ker, A. P., & Goodwin, B. K. (2000). Nonparametric estimation of crop insurance rates revisited. *American Journal of Agricultural Economics, 82*(2), 463–478.
26. Ker, A. P., Tolhurst, T. N., & Liu, Y. (2016). Bayesian estimation of possibly similar yield densities: implications for rating crop insurance contracts. *American Journal of Agricultural Economics, 98*(2), 360–382.
27. Kirwan, B. E., Uchida, S., & White, T. K. (2012). Aggregate and farm-level productivity growth in tobacco: Before and after the quota buyout. *American Journal of Agricultural Economics, 94*(4), 838–853.
28. Liu, Y., & Ker, A. P. (2020). Rating crop insurance contracts with nonparametric bayesian model averaging. *Journal of Agricultural and Resource Economics, 45*(2), 244–264.
29. Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys, 29*(1), 46–75.
30. Nelson, C. H. (1990). The influence of distributional assumptions on the calculation of crop insurance premia. *Applied Economic Perspectives and Policy, 12*(1), 71–78.
31. Ozaki, V. A., Goodwin, B. K., & Shirota, R. (2008). Parametric and nonparametric statistical modelling of crop yield: Implications for pricing crop insurance contracts. *Applied Economics, 40*(9), 1151–1164.

32. Park, E., Brorsen, B. W., & Harri, A. (2019). Using bayesian kriging for spatial smoothing in crop insurance rating. *American Journal of Agricultural Economics, 101*(1), 330–351.
33. Park, E., Brorsen, B. W., & Harri, A. (2020). Spatially smoothed crop yield density estimation: physical distance versus climate similarity. *Journal of Agricultural and Resource Economics, 45*(3), 533–548.
34. Park, E., Harri, A., & Coble, K. H. (2022). Estimating crop yield densities for counties with missing data. *Journal of Agricultural and Resource Economics, 47*(3), 634–655.
35. Ramsey, A. F. (2020). Probability distributions of crop yields: a bayesian spatial quantile regression approach. *American Journal of Agricultural Economics, 102*(1), 220–239.
36. Ramsey, A. F., Ghosh, S. K., & Goodwin, B. K. (2020). Rating exotic price coverage in crop revenue insurance. *Agricultural Finance Review, 80*(5), 609–631.
37. Ramsey, A. F., & Liu, Y. (2023). Linear pooling of potentially related density forecasts in crop insurance. *Journal of Risk and Insurance, 90*(3), 769–788.
38. Ramsey, A. F., Goodwin, B. K., & Ghosh, S. K. (2019). How high the hedge. *Journal of Agricultural and Resource Economics, 44*(2), 227–245.
39. Santeramo, F. G. (2018). Imperfect information and participation in insurance markets: evidence from Italy. *Agricultural Finance Review, 78*(2), 183–194.
40. Sherrick, B. J., Zanini, F. C., Schnitkey, G. D., & Irwin, S. H. (2004). Crop insurance valuation under alternative yield distributions. *American Journal of Agricultural Economics, 86*(2), 406–419.
41. Skees, J. R., Black, J. R., & Barnett, B. J. (1997). Designing and rating an area yield crop insurance contract. *American Journal of Agricultural Economics, 79*(2), 430–438.
42. Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika, 9*(6), 449–460.
43. Tolhurst, T. N., & Ker, A. P. (2015). On technological change in crop yields. *American Journal of Agricultural Economics, 97*(1), 137–158.
44. Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2022). Forecast combinations: An over 50-year review. *International Journal of Forecasting, 39*(4), 1518–1547.
45. Wen, K. (2023). A semiparametric spatio-temporal model of crop yield trend and its implication to insurance rating. *Agricultural Economics, 54*(5), 662–673.
46. Wu, S., Goodwin, B. K., & Coble, K. (2020). Moral hazard and subsidized crop insurance. *Agricultural Economics, 51*(1), 131–142.
47. Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis, 13*(3), 917–1007.
48. Zhu, Y., Goodwin, B. K., & Ghosh, S. K. (2011). Modeling yield risk under technological change: Dynamic yield distributions and the us crop insurance program. *Journal of Agricultural and Resource Economics*, 192–210.

Chapter 9
A Recursive Method on Estimating ARFIMA in Agricultural Time Series

Simon Wang and Nina Ni

Abstract In this paper, we apply a recursive method to financial data to determine their corresponding Hurst exponent and the optimal Autoregressive Fractionally Integrated Moving Average (ARFIMA) models. We begin by introducing the long-range dependence phenomenon and methods to address it in time series modeling. Then, a recursive algorithm, where the Hurst exponent is estimated by applying an autoregressive filter to the data repeatedly until it converges, is empirically tested with simulated data for stability and convergence. Finally, we apply this convergence approach to real commodity data sets. We identify the optimal ARFIMA models for each commodity studied and estimate the Hurst exponent as well as their corresponding ARFIMA parameters. Our results provide a stable method for estimating the Hurst index and fitting stationary long-memory processes to ARFIMA models.

9.1 Introduction

Modeling commodity prices and their derivatives has been problematic for researchers and practitioners, one major obstacle is the existence of long-range dependence (LRD) within the data, which limits the standard short memory model approach. It is well known that many macroeconomic time series data are followed by long memory process, this phenomenon in economics was first discovered by [1]. He summarized that the economic and financial historical data typically exhibit some distinct low-frequency non-periodic cyclical patterns.

Mathematically speaking, a stationary process Y_t has the long memory property, if for its auto-correlation function $\rho(k) = Cov(Y_t, Y_{t+k})/Var(Y_t)$ satisfies the

S. Wang (✉)
Stable Group Ltd., London, UK
e-mail: simon@stableprice.com

N. Ni
WisdomTree in Europe, London, UK

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_9

following:

$$\lim_{n\to\infty} \sum_{k=-\infty}^{n} |\rho(k)| = \infty.$$

In other words, a long memory series has an auto-correlation function that decays hyperbolically, slower than the geometrical tail "short memory" e.g ARMA. Thus, it may be predictable at long horizons. The economic application of LRD has been extended from macroeconomics to finance. For example, in [2] the author provides details of estimating long memory models to price. [3] finds the long-memory process that provides a good explanation of the behavior of inflation. Simultaneously, another paper [4], shows that the properties of long memory models and their response to shocks are quite different from high-order autoregressive models.

On the other hand, [5] introduces the concept of a long memory to measure the persistence of stationary processes. In [5] , the Autoregressive Fractionally Integrated Moving Average (ARFIMA) process is referred to model a time series with long-range dependence. For a stationary time series, an ARFIMA refers to an Autoregressive moving average (ARMA) model where the innovations are fractional white noise, which can be rewritten in lag operator notation as:

$$\Phi(B)(1-B)^d(Y_t-\mu) = \Theta(B)\varepsilon_t.$$

where d is the fractional integration parameter that allowed to take non-integer values, B is the lag operator ($B^n Y_t = Y_{t-n}$, $n = 1, 2, 3...$), $\Phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ specifies the AR lag polynomial, and $\Theta(B) = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)$ specifies the MA lag polynomial. The properties of Y_t depend on the value of the fractional integration parameter d. In an ARFIMA process, $0 < |d| < 0.5$, thus the process Y_t can be considered as a long memory process.

Regarding the estimation methods for ARFIMA process, there exists three main approaches: the maximum likelihood methods (see [6–8]), the semi-parametric methods (see [9, 10]) and the heuristic methods (see [11–13]). The maximum likelihood estimator provides a consistent approach to estimate all the parameters of interest simultaneously, but it usually generates unstable results and high computational costs due to the ill-behaved likelihood function. The last two approaches also called two-step estimation methods, in other words, to fit an ARFIMA(p,d,q) model we estimate the long memory parameter d first, then the second step of the procedure is to fit ARMA model to the data. More specifically, the Geweke & Porter-Hudak (or GPH) method [9] is based on the behavior of the spectral density of ARFIMA process near frequency zero, but it shows a bias in the presence of strong autocorrelation in the ARMA process. A modified version of the GPH estimator by using a smoothed periodogram is introduced to reduce the bias by [10]. However, many researchers such as [14] and [15] suggest the estimators using the heuristic

methods is more robust without studying the asymptotic properties, and this is the reason why in this paper we focus on the heuristic approach on detecting the long-range dependence.

Fractal analyses became more popular in the finance community recently, one particular application is the use of Hurst exponent in financial or economical time series data. Hurst exponent or Hurst index provides a measure for long term memory and fractality of a time series. Due to its robustness and few assumptions, it finds a wide applications in time series analysis. In [16], Peters uses the Hurst process and R/S Analysis in testing and researching Capital Markets. He introduced the Fractal Markets Hypothesis (FMH), which avoids the classical assumptions that returns are log-normal and uncorrelated for financial mathematical models. For a stationary process, the values of the Hurst exponent, denotes as $H \in (0, 1)$, divides time series into three categories: a completely random series with $H = 0.5$ and an anti-persistent series or a persistent (trend) series with $H < 0.5$ and $H > 0.5$ separately. The strength of anti-persistent increases as H approaches 0 and the strength of trend increases as H approaches 1.

It is proved in [9], the following relation exists between the Hurst exponent H and the fractional parameter d for the same time series.

Theorem 9.1 *Let Y_t be an ARFIMA process with parameter d $(-0.5 < d < 0.5)$, if and only if, it is also a stationary process with Hurst exponent $H = d + 0.5$.*

Proof See[9]. □

It is known in empirical study that the estimation of ARIFMA model is not stable due to the ill-behaved likelihood function, see [17–23] for details. Here, from Theorem 9.1, we employ a convergence algorithm to fit the data. The algorithm estimates an initial Hurst exponent first, then several loops follow-by to de-factionalize the data until the most stable Hurst index is found, the finally Hurst index can be transferred to our fractional parameter to fit the selected ARFIMA model. In the rest of the paper, we will test a recursive approach on estimating the Hurst exponent and further the ARFIMA parameters for the corresponding time series. Then, we will apply the algorithm to a set of commodity data, the matching Hurst index and optimal AFRIMA models will be reported.

9.2 A Recursive Approach on Estimating Hurst Exponent

Due to the nature of our study here, when we choose the methods for estimating the Hurst exponent, the stability and efficiency are considered primarily. Thus, three methods are short-listed and used in our research, see [23–26] for details. These are: Rescaled Range Statistic Method (RS in short), Aggregated Variance Method (AV in short) and a recursive implementation of the AV method (OT in short).

9.2.1 Existing Methods Estimate the Hurst Exponent

Range Statistic Method (**RS** in short) or the rescaled range or R/S analysis is one of the most well-known technique for estimating Hurst parameter. It is able to distinguish a random series from a fractal series, irrespective of the distribution of the underlying series (Gaussian or non-Gaussian). Now, given a sequence of n observations $y_1, y_2, \cdots, y_n$, then R captures the maximum and minimum cumulative deviations of the observations y_t from its mean μ , and it is a function of time:

$$R(n) = \max_{1 \leq t \leq n} \left[\sum_{i=1}^{t} (y_i - \mu) \right] - \min_{1 \leq t \leq n} \left[\sum_{i=1}^{t} (y_i - \mu) \right].$$

$S(n)$ is the standard deviation of the original time series:

$$S(n) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2}.$$

Then the R/S ratio of R and the standard deviation S can be calculated and it is called the rescaled range statistic or R/S statistic. Meanwhile, the R/S ratio of the original time series can be estimated by fitting the power law

$$E(R/S) = Cn^H.$$

where C is a positive, finite constant independent of n. And for some value of n, the Hurst exponent H can be calculated by

$$H = \frac{\log(R/S)}{\log(n)}, \ 0 < H < 1.$$

Hence the estimate of H can be found by calculating the slope of the logarithm of R/S against $\log(n)$ using regression, the first order liner function is adopted in this paper. Based on our observation from the data itself, for RS method, we choose a constant time span of 2^{10} as the observation rolling windows size. This method is able to obtain an robust estimator even for data followed by a heavy-tailed probability density function (see [27]). On the other hand, the exact distribution of the R/S statistic is difficult to determine. So, it is biased and affected by the non-stationarity in the data. One possible way in [26] to reduce the bias is to ignore the points on the extreme left and right of the log-log plot, and use only the points in the central region of the plot, the former due to the influence of the short-term dependence structure while the latter because only a few observations are included. However, based on our empirical study, we aplly another possible approach in our paper: instead of using the slope of common logarithm (using base 10) of R/S

against $\log(n)$ we are going to adjust it to the binary logarithm (using base 2), that is $H = \log_2(R/S)/\log_2(n)$, which allows us to estimate the slope with all the available points.

Generally speaking, the **Aggregated Variance method** (**AV** in short) introduced in [28] is the analysis of the variances of aggregated time series processes. Given a sequence of n observations $y_1, y_2, \cdots, y_n$, one property of long-memory process is that the variance of the sample mean μ_n converges to zero slower than the rate n^{-1}, where n is the sample size. For a large sample size n they have following relationship (see [29]):

$$Var(\mu_n) \sim cn^{2H-2}.$$

where $c > 0$. This suggest the AV method for estimating H: dividing the series into n/m blocks of size m and compute the sample mean in each blocks

$$\mu_m(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} y_i\,, \quad k = 1, 2, \cdots, n/m.$$

and the sample variance

$$s^2(m) = (n/m - 1)^{-1} \sum_{k=1}^{n/m} [\mu_m(k) - \mu_n]^2.$$

Hence plotting $\log s^2(m)$ against $\log(m)$ should yield points scattered with slope equal to $2H - 2$, then the Hurst exponent H can be estimated by evaluating the slope through regression, again here a first order liner function is adopted. For the AV method, we choose a constant block size of 1000 for splitting the data.

A slight improvement from aggregated variance method is we can update the block size constantly inside the estimation process. We implement another method that will be explained later in full details (OT in short), where instead of using a fixed size block size for splitting the data, it is dynamically adjusted during the estimation process until the block size is smaller than 5.

9.2.2 Stability for Current Methods

In order to see which method works the best, we conduct an further experiment on the simulated data where we know the exact Hurst exponent in Table 9.2. For generating artificial ARFIMA process, we choose the fast algorithm processes in [30] where the calculation speed (number of arithmetic operations) is improved from order T^2 to $T\log(T)$, T is the length of the time series.

In the Table 9.2, we test the accuracy of all chosen methods on estimating the Hurst exponent for simulated stationary data at all Hurst range. Recall in an ARFIMA process, the process is stationary if $-0.5 < d < 0.5$, and according to Theorem 9.1, one can have $d = H - 0.5$. Thus, we simulate 10,000 data for each of the **AFRIMA(0,d,0), AFRIMA(0.5,d,0), AFRIMA(0,d,0.3), AFRIMA(0.5,d,0.2)** and **fractional Brownian Motion processes** with the fractional parameter d ranges from -0.5 to 0.5 separately. Then, we apply all three different methods 5000 times on each group of the 10,000 data from a totally of 55 groups and calculated the means and standard deviations from the Hurst exponent they have estimated. Finally, we calculate the average mean error and average standard deviation in each of the five models for comparison. Generally speaking, the method with the smallest average error and average standard deviation is the most accurate one for all range of Hurst exponent estimation, and we do find the AV and OT method perform better in terms of mean values. However, after a closer look at the real data sets we have, we realized that most of them may have a Hurst index around 0.5, and the RS method had better results in this area in terms of mean values and low variance in general. We decided to keep all three estimations in the following study and use the average of them as our initial Hurst exponent estimation.

In Figs. 9.1 and 9.2, we visualize part of the simulated estimation results for three different methods in box-plots. Again for ARFIMA(0,d,0) and ARFIMA(p,d,q) models, one can see all method generally work well when $d = 0$, i.e. $H = 0.5$. However, we also notice that the traditional heuristic approaches work better with ARFIMA(0,d,0) mode or the pure fractional noise. This can be explained by the short memory part of the time series can act like a disruptor when the measure for long memory is applied. From this observation, one can see if one can remove the

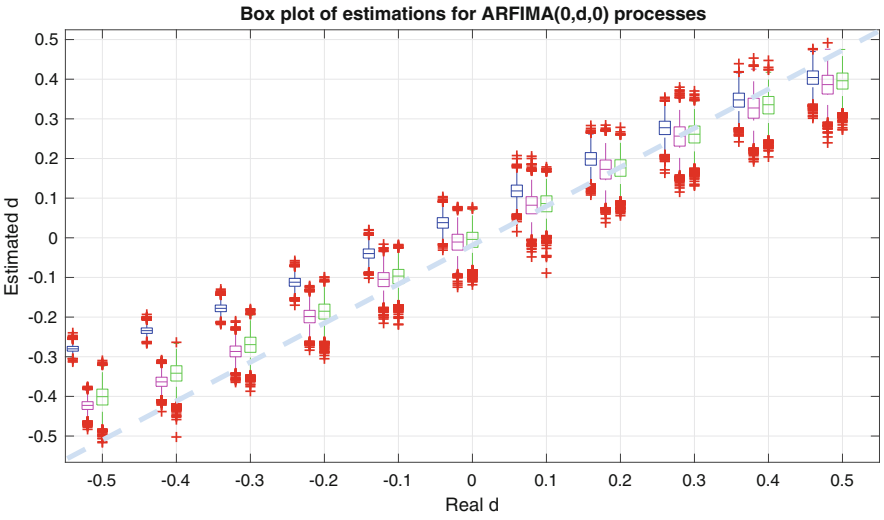


Fig. 9.1 ARFIMA(0,d,0) process box-plot (Blue: RS, Magenta: AV, Green: OT)

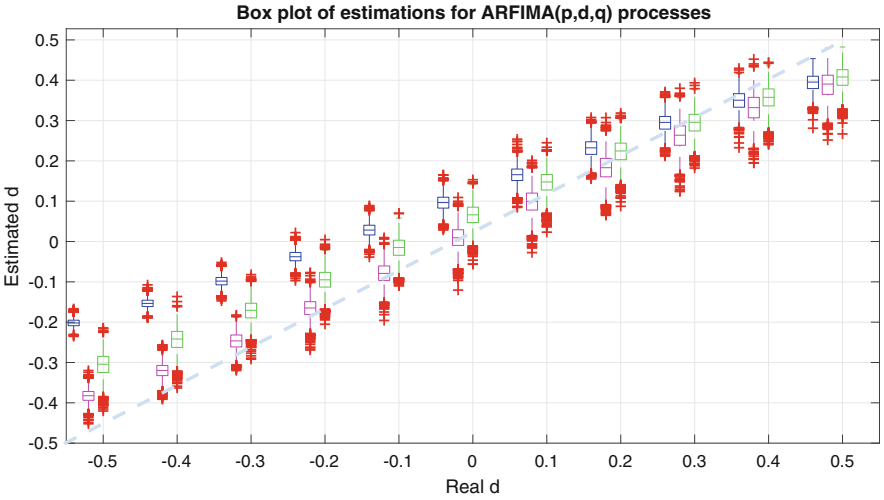


Fig. 9.2 ARFIMA(0.5,d,0.2) process box-plot (Blue: RS, Magenta: AV, Green: OT)

short memory part from a time series data, and only estimate the Hurst exponent for the long memory part, a more accurate fitting algorithm may be found for the measure for long memory.

9.2.3 A Recursive Approach

Adopting the idea from [23], where the authors used an algorithm on testing the stability of Hurst estimation, we adopt a similar convergence procedure in estimation the parameters for an ARFIMA model. The procedure goes as follows: we first estimate an initial Hurst exponent from the methods we mentioned above, then an infinite autoregressive filter, $Y_t = (1-B)^d X_t = \sum_{s=0}^{\infty} b(s) B^s X_t$, is applied on the data X_t with the Hurst exponent $d = H - 0.5$ we just estimated. With the newly generated data Y_t, we hire the traditional time series methodology (i.e. Box-Jenkins Method) to find an adequate ARIMA model for it. After estimate the polynomials of the ARIMA model, a new time series is generated by the filter: $\widehat{Y}_t = \dfrac{\widehat{\Phi}(B)}{\widehat{\Theta}(B)} X_t$ in order to find the "pure" ARFIMA(0,d,0) process. Finally, a recursive method is applied to estimate Hurst exponent again from $\widehat{Y}_t$ until the Hurst index converges. We provide the following list for the details of each step in our algorithm:

1. De-seasonality and inflation adjust the data if needed.
2. Check the stationarity and generate the initial time series data X_t.

(a) Augmented Dickey-Fuller (ADF) test.
(b) Take the increment, log difference or return ratio if needed.
(c) Adjust observed data and generate the initial stationary time series X_t.

3. Initial calculation of the Hurst parameter for X_t.

(a) Rescaled Range Statistic Method (R/S procedure).
(b) Aggregated Variance Method.
(c) Recursive implementation of the AV method.
(d) Use the average of (a), (b) and (c) as our initial estimation for Hurst.

4. Use $H = d + 0.5$ to find the initial value of d, the fractional parameter for the ARFIMA process.

(a) See details and proofs in [9].
(b) Note, H is from d, but one cannot use d to estimate H.

5. Calculate the underlying time series Y_t for the ARIMA model after removing the fractional fact d.

(a) Using the infinite autoregressive filter: $Y_t = (1 - B)^d X_t = \sum_{s=0}^{\infty} b(s) B^s X_t = \sum_{s=0}^{\infty} b(s) X_{t-s}$.
(b) where $b(s) = \prod_{k=1}^{s} \frac{k + d - 1}{k} = \frac{\Gamma(s - d)}{\Gamma(-d)\Gamma(s + 1)}$.
(c) X_t is the observed data after stationary adjustments. Choose a large s to stop the summation. (In our case we choose $s = 150$).

6. Hire traditional time series methodology to find the most adequate ARIMA model for the de-fractionalized data Y_t and estimate the parameters.

(a) One can choose the Box-Jenkins Method here.
(b) In our practice, we provide a set of time series models from $ARMA(0, 0)$ to $ARMA(6, 6)$ and estimate all models in parallel.
(c) The results polynomials for the $ARMA$ ($\Phi(B)Y_t = \Theta(B)\epsilon_t$) are saved as $\widehat{\Phi}(B)$ and $\widehat{\Theta}(B)$.

7. Using the ARMA filter, calculate $\widehat{Y_t} = \frac{\widehat{\Phi}(B)}{\widehat{\Theta}(B)} X_t$ to find the pure $ARFIMA(0, d, 0)$ process.
8. Re-estimate fractional parameter d from the "pure" process $\widehat{Y_t}$. And repeat steps 3–8.
9. Stop the algorithm until d converges.

9.2.4 Convergence and Stability

To test the convergence and stability of our algorithm, we conduct another two experiments on the simulated data in Tables 9.3 and 9.1 (Table 9.2). In Table 9.3,

we test the three different Hurst exponent methods together with their average on simulated data to monitor the convergence of our algorithm in controlled environments. We simulate 10,000 data for three types of ARFIMA models: with AR term, with MA term and with AR and MA terms. The "Steps" in the second column of the table indicate the current loop of the recurrent algorithm as we proposed above. One can find, for all simulated data with all methods, we observe the convergence in the values of d (thus the convergence of Hurst exponent), the convergences happen quickly and usually stable within the first 10 steps. The most right column provides the average error for each method at each step, which, again, indicated the AV and OT method work better in most of the cases. Note, if we take the Hurst exponent as the average of the three methods, we result some decent results in terms of overall error and convergence speed.

In Table 9.1, we provide the empirical results for our algorithm on estimating the parameters for ARFIMA model when apply on simulated data. These are fitting results from six individual random run of our algorithm. Six scenarios are provided here to simulated ARFIMA process both with positive and negative Hurst exponent and different ranges of the fractional parameter. The simulated time series all have 10,000 time steps, which is similar to the real data we have, and one can find a random selection of our fitting procedure can result some adequate performance with acceptable average errors in an overall prospective.

9.3 Examples on Fitting Commodity Data

9.3.1 Commodity Data Sets and Their Stationary

In this section, we fit commodity data into ARFIMA models and select the best model for each product using the algorithm above. The data we choose is the daily trading data from Bloomberg for the past 20 or 30 years. There are commodity index or future prices ranging from industry oil, metal, livestock (animal) to non-livestock (crops). Unlike the stock prices, the price for commodities and the relevant financial derivatives are less liquid and may not be completely hedgeable. This may be due to the properties of physical goods and their storage consideration. But, such properties provide a good opportunity to estimate the Hurst exponent, as a Hurst exponent that is not 0.5 indicates the underlying time series is not a completely random series which is unpredictable. Table 9.4 summaries the statistical property of the selected data together with the p-values from the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests and the augmented Dickey–Fuller test (ADF) tests for stationary. One can tell, most of the data are heavily skewed, with fat tail when compare to normal distribution whose skewness is 0 and kurtosis is 3.

On the other hand, from the p-values, it is not hard to see in general, all of the prices are not stationary. This may cause troubles for us, as our alogrithm assumption requires the stationarity of the data. Thus, instead of studying the price

data directly, we choose to work with the volatility of the price. The log differences (log return, defined as $Y_t = \log X_t - \log X_{t-1}$) and the first difference (increments, defined as $Y_t = X_t - X_{t-1}$) are typical choices for empirical study for financial time series as they fit the underlying assumption for most of the financial mathematics theories and easy to work with. Also, it is mentioned in [31] the absolute values of the log return is a good measure for data's long memory. In the following section, we will only work absolute values of log returns and the increments data. Table 9.5 shows the KPSS and ADF tests for the processed data (Diff: increments; Log: log difference), except for "CME FEEDER CATTLE INDEX" and perhaps "BRENT CRUDE OIL INDEX" in KPSS tests, all other commodities exhibit stationarity after processing.

Note, the disadvantage of taking the log differences or calculate the increments is that we change the data distributions. So the interpretations of the Hurst exponent will be different as well when compared to the Hurst exponent calculated from the original data. For instance, a Hurst exponent bigger than 0.5 for the original price data indicates long-rang dependence for the price itself, which means the future price of this commodity is largely depend on its previous price for a long time. This could be caused by the product's natural characteristics or trading behaviors from traders. However, if we find commodity increments' or log returns' Hurst exponent is bigger than 0.5, then we cannot make the previous judgment. Rather, this persistence mean the price movement of this commodity is consistent, i.e. large volatility is more likely to be followed by another huge disturbance in price data. Meanwhile, a Hurst exponent smaller than 0.5 for these measure could be the traders has mixed predictions for the commodity price thus the it changes rapidly and unpredictably.

9.3.2 Practical Considerations

It is obvious that the infinite filter in step 5 from the algorithm above converges to a small enough number that can be ignored eventually, as a mater of fact, in our practice we find $s = 150$ is an adequate choice if not too large (See Fig. 9.3 for a comparsion on $b(s)$'s convergence with differnce d values). However at step 6, we reckon it is not feasible to choose the best ARIMA model for the time series data manually even if we follow the Box-Jenkins method. This is essentially because we are dealing with a data set with 40 different types of commodities, it will be time consuming to compare several models for each product and never mention people's perspective varies. To solve this problem, after individual studying the data samples, we realize usually a model smaller than $ARMA(6, 6)$ is usually good enough in terms of modeling, simulation and prediction within our experience. Thus, this problem can be solved by adopting a range of models from $ARMA(0, 0)$ to $ARMA(6, 6)$ for the filtering procedures, more specifically, for each commodity, the algorithm runs separately on each of these possible combinations for a totally of 49 times. All fitting results and statistical testing parameters are saved in local files.

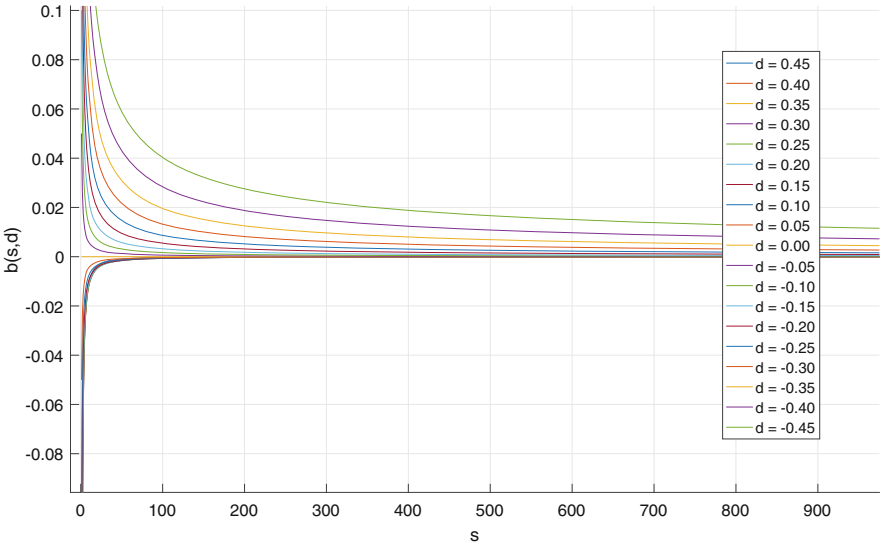


Fig. 9.3 The convergence of $b(s)$ for $d = -0.45$ to $d = 0.45$

After all models are estimated, a third program compares the Akaike or Bayesian information criteria (AIC, BIC) from each realization of the estimation and choose the model according to the AIC and BIC under the condition that convergence in d happens. One consideration of this implementation is the calculation efficiency and running time. However, we think this approach does not necessarily hold when examining the log-difference running-time for the whole algorithm, as each model filter is estimated separately and does not require any information from the previous steps, so one can implement them running in parallel thus saving a signification amount of running time (Table 9.6).

Table 9.7 is the output example for "CBOT Corn Futrue" increment data. In total, 49 models are estimated using the algorithm above. The "NoS" stands for number of steps, it indicates the recursions times for d to converge to a stable value. A minimum boundary of 9 is set for stability concerns. One can see, all models converge with relatively fast converge speed and different models do provide significantly different estimation of the fraction parameter. Eventually, an AFRIMA (5,−0.0218,5) is chosen by sorting the "AICBIC" in an ascending order, and the Hurst exponent is estimated as the average of the best 3 models. We also provide Fig. 9.4 for some of the convergence test. In these figures, the logarithm of the convergence ratios calculated from both increments data and log difference are presented, we select some of the most representative commodities and one can see they all converge very fast usually within 9 steps.

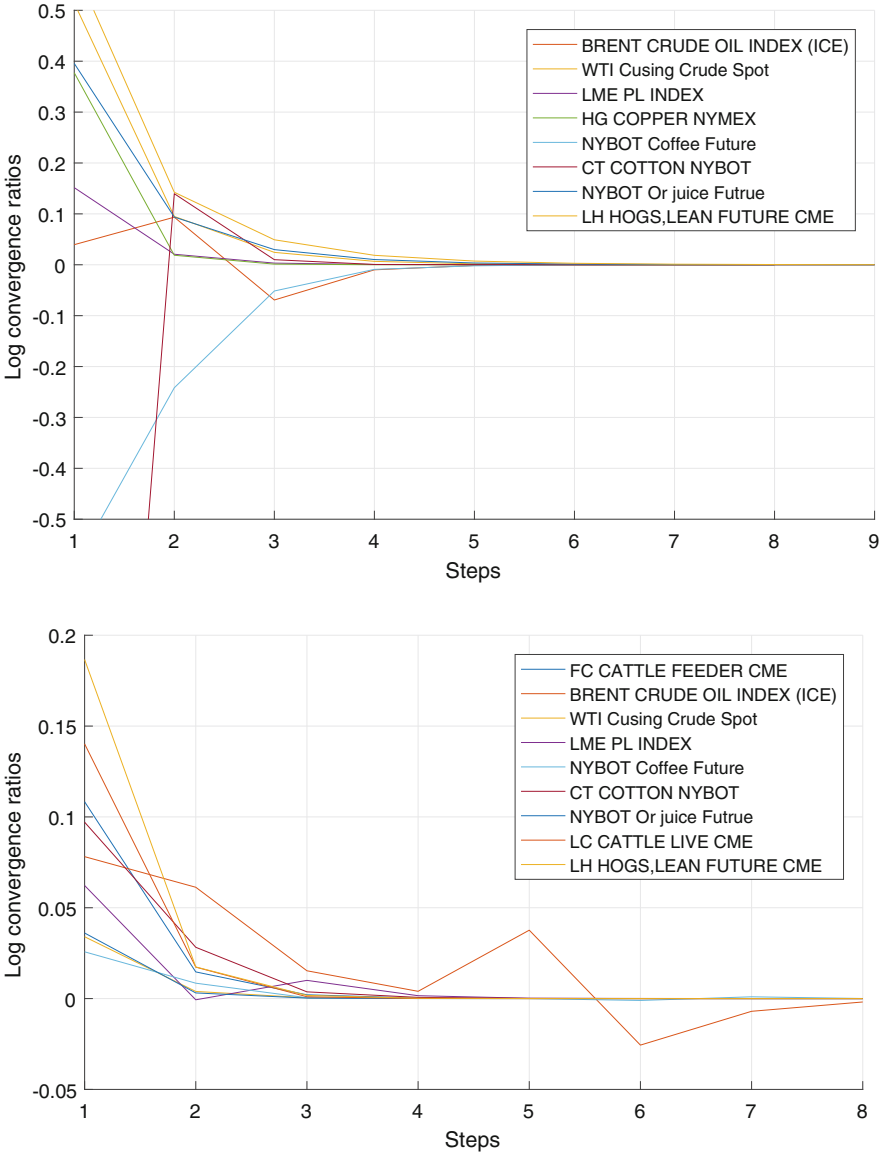


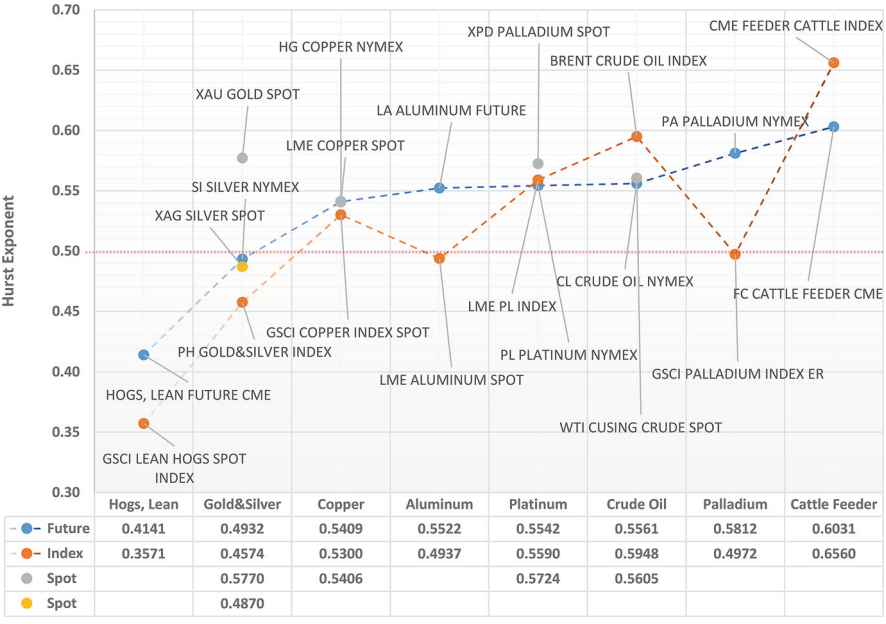
Fig. 9.4 Log plots of convergence ratios (Left: increments Right: Log difference)

9.3.3 Hurst Exponents and ARFIMA Model Results

Tables 9.8 and 9.9 provide the final results from us. In these tables, we list the best ARFIMA model for each of the 40 commodities and report their model parameters in the increments form and Log difference form. We see the algorithm works with

all commodities and they usually converge very fast (a minimum step of 9 is used in all recursions for stability reasons). The details of the estimation of each commodity is saved in the local files (provide on request) with the convergence test as well as statistical test values for each parameters. In Fig. 9.6, we give an overview of the Hurst exponent for all commodities and the chosen optimal model for each of them. As one can see in the vertical bar plot, the orange bars indicate the Hurst exponent for the absolute values log difference data and the blue bar are for the first difference data. We find the log return generates larger Hurst exponent than the increments for all data, this results consist with the results from [31, 32], where the Hurst exponents for agriculture commodities are calculated in another two difference methods. Furthermore, there is no obvious pattern in the optimal ARFIMA models our algorithm picked, but we see an consideration for a total number of 49 models is indeed needed, as some of the commodities require pretty large model to cooperate their dynamics.

For those commodities which have index, future or spot prices, we compare their Hurst exponent in Fig. 9.5. Despite the facts that a future contact and an index for the same commodity may not necessary written on the exactly same product, we can still observe a trend where future prices tend to have a closer value to 0.5 when compare to index prices, and spot prices always have a larger Hurst exponent than the future and index prices. Recall a Hurst exponent close to 0.5 means more liquidity in the contract based on the assumptions from financial mathematics. It is not hard to understand future prices as derivatives prices are trading more frequently



	Hogs, Lean	Gold&Silver	Copper	Aluminum	Platinum	Crude Oil	Palladium	Cattle Feeder
Future	0.4141	0.4932	0.5409	0.5522	0.5542	0.5561	0.5812	0.6031
Index	0.3571	0.4574	0.5300	0.4937	0.5590	0.5948	0.4972	0.6560
Spot		0.5770	0.5406		0.5724	0.5605		
Spot		0.4870						

Fig. 9.5 Hurst exponent comparison among future and index

than indexes, thus behaves more randomly. And since the spot price of a product is mostly related to the product itself and all the seasonal facts affects it, the larger values in Hurst exponent, which indicate high persistence, for the spot data also seems reasonable to us.

Another interesting discovery for commodity categorization is that agriculture product tends to have smaller Hurst than industry products if we sort the commodity with respect to their Hurst exponent calculated from increments. This pattern does not necessary holds when exam the log difference data or if the Hurst exponent is calculated by RS method without the recursion. See Table 9.6, where we summarize the estimation results for Hurst exponent with increments data and log difference data for both our method and the RS method in on run, this pattern only shows in "Hurst(Diff)". In Fig. 9.6, we color-coded all commodities with respect to their categories. From bottom to top, red is for livestock or animal, green is for non-livestock or crops, grey is for metal and orange is or industrial oil. Averagely speaking, despite feeder cattle, one can observe a general trend where Hurst exponent goes from small to large following the order of livestock, non-livestock, metal and eventually industrial oil. And for most of the cases, live and non-live stock commodities have Hurst exponent that is smaller than 0.5 for their first difference while metals and oils have Hurst exponent larger than 0.5. We do not endeavour to explain the reason behind this observation, as this is not within the scope of this paper. But we believe this differences between commodities could be cased partly by their demand elasticity. As similar discoveries have been found by [33] for CEV model parameters on agriculture product data, we think the for industrial goods, the demand is harder to be substitute, for example, certain products and industrial processes require copper as resources, it is not feasible and economical to find a substitution for copper at a short time if the price of copper goes up, thus, the demand for copper is more elastic. The same argument goes for other metals and even for oils. However, on the other hand, agriculture commodities like crops and particular meats are easy to be substituted.

9.4 Conclusion

In conclusion, this paper has demonstrated the application of a recursive method to the agricultural prices for the determination of their Hurst exponent and optimal Autoregressive fractionally integrated moving average (ARFIMA) models. We initially discussed the concept of long-range dependence in time series modeling and presented a recursive algorithm for estimating the Hurst Exponent. This algorithm was empirically validated using simulated data to assess its stability and convergence before being applied to real commodity datasets. Our findings reveal the identification of optimal ARFIMA models for each commodity under study, along with estimates of their corresponding Hurst exponent and ARFIMA parameters. These results provide a reliable approach for estimating the Hurst index and fitting stationary long memory processes to ARFIMA models.

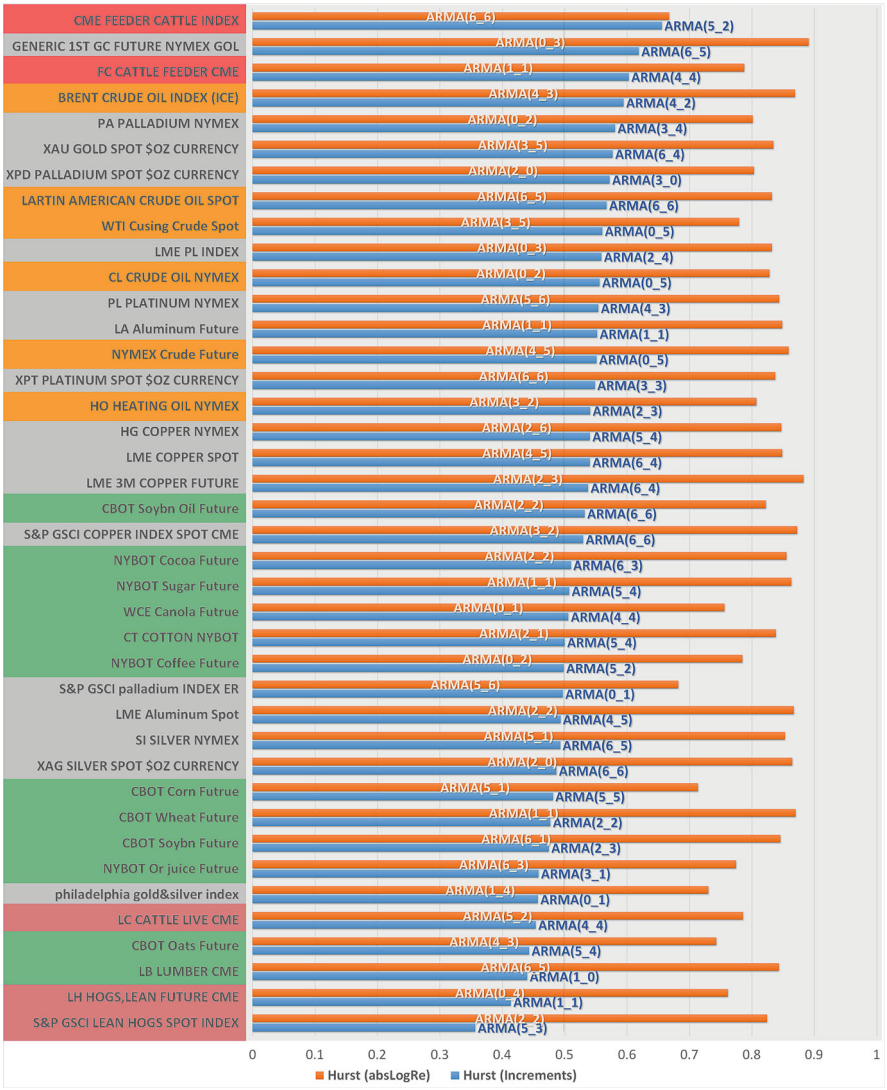


Fig. 9.6 Fitting model results and Hurst exponent for all data

9.5 Tables

The following are tables mentioned in the article above.

Table 9.1 Accuracy test

		AR	FI	MA	Average err.
ARFIMA	Real	**0.5**	**0.3**		
0.5,0.3,0	RS	0.5535	0.2548		0.0350
	AV	0.5654	0.2431		0.0433
	OT	0.5575	0.2508		0.0378
	Average	0.5591	0.2492		0.0390
ARFIMA	Real	**0.5**	**−0.3**		
0.5,−0.3,0	RS	0.3545	−0.1599		0.1010
	AV	0.5040	−0.3122		0.0064
	OT	0.4651	−0.2735		0.0219
	Average	0.4401	−0.2483		0.0395
ARFIMA	Real		**0.2**	**0.2**	
0,0.2,0.2	RS		0.1802	0.2125	0.0117
	AV		0.1386	0.2484	0.0391
	OT		0.1639	0.2266	0.0224
	Average		0.1607	0.2293	0.0245
ARFIMA	Real		**−0.2**	**0.2**	
0,−0.2,0.2	RS		−0.1276	0.1239	0.0525
	AV		−0.2344	0.2182	0.0195
	OT		−0.2262	0.2110	0.0142
	Average		−0.1964	0.1849	0.0078
ARFIMA	Real	**0.5**	**0.1**	**0.2**	
0.5,0.1,0.2	RS	0.4107	0.1828	0.2141	0.0409
	AV	0.4702	0.1322	0.2055	0.0148
	OT	0.4742	0.1288	0.2050	0.0130
	Average	0.4524	0.1475	0.2078	0.0226
ARFIMA	Real	**0.5**	**−0.1**	**0.2**	
0.5,−0.1,0.2	RS	0.3936	−0.0017	0.2271	0.0491
	AV	0.5054	−0.0982	0.2125	0.0046
	OT	0.5209	−0.1120	0.2112	0.0089
	Average	0.4740	−0.0706	0.2157	0.0141

Table 9.2 Table of hurst estimation methods stabilities

d	Model	ARFIMA method	0,d,0	10,000	ARFIMA method	0.5,d,0	10,000	ARFIMA method	0,d,0.3	10,000	ARFIMA method	0.5,d,0.2	10,000	fBM method	d+1 first difference	10,000
	Stats	RS	AV	OT	RS	AV	OT	RS	AV	OT	RS	AV	OT	RS	AV	OT
−0.5	Mean	−0.279968	−0.423363	−0.402521	−0.212742	−0.387203	−0.320102	−0.257004	−0.411348	−0.372613	−0.201618	−0.382799	−0.304920	−0.305253	−0.437374	−0.430768
	Std.	0.008970	0.014603	0.029589	0.009791	0.015963	0.029862	0.009160	0.015372	0.029168	0.009983	0.016310	0.028688	0.008373	0.013014	0.029327
−0.4	Mean	−0.234204	−0.363694	−0.343588	−0.163616	−0.324702	−0.257297	−0.210269	−0.350438	−0.313828	−0.153634	−0.320417	−0.242862	−0.256719	−0.376678	−0.368953
	Std.	0.010218	0.017184	0.028535	0.011385	0.018571	0.028536	0.010834	0.017762	0.028663	0.011467	0.018729	0.028907	0.009945	0.016031	0.028848
−0.3	Mean	−0.177702	−0.287244	−0.270438	−0.107324	−0.250307	−0.185323	−0.154015	−0.275308	−0.240844	−0.098468	−0.247510	−0.172931	−0.195200	−0.296527	−0.290274
	Std.	0.012168	0.020113	0.028096	0.013218	0.021136	0.027981	0.012522	0.020402	0.028473	0.013450	0.021532	0.027593	0.011772	0.019179	0.027641
−0.2	Mean	−0.112445	−0.199786	−0.188175	−0.045193	−0.168250	−0.107885	−0.089908	−0.190219	−0.159039	−0.037227	−0.165777	−0.095980	−0.123552	−0.205396	−0.200271
	Std.	0.014491	0.023289	0.027763	0.015269	0.023431	0.026714	0.014589	0.023308	0.027473	0.015331	0.023805	0.027815	0.014262	0.022976	0.027413
−0.1	Mean	−0.040111	−0.106369	−0.098209	0.021633	−0.081374	−0.026382	−0.019612	−0.099609	−0.073008	0.028611	−0.078497	−0.015416	−0.044546	−0.108198	−0.103433
	Std.	0.016888	0.026127	0.027479	0.017406	0.026289	0.027675	0.017225	0.026336	0.027375	0.017094	0.025912	0.027151	0.017106	0.026079	0.027407
0	Mean	0.037691	−0.011074	−0.006004	0.091437	0.007469	0.056708	0.054800	−0.007138	0.015244	0.096370	0.009498	0.065634	0.037607	−0.011592	−0.006626
	Std.	0.019328	0.029161	0.028472	0.019481	0.028474	0.027994	0.019680	0.029313	0.028179	0.019757	0.028965	0.028084	0.019377	0.029039	0.028166
0.1	Mean	0.117959	0.082196	0.085353	0.162628	0.097227	0.140054	0.131476	0.085973	0.104074	0.165103	0.098236	0.145844	0.122339	0.084286	0.090286
	Std.	0.021879	0.032085	0.029640	0.021386	0.031054	0.028563	0.022057	0.032305	0.029815	0.021682	0.031927	0.029858	0.021638	0.031865	0.029290
0.2	Mean	0.199294	0.172899	0.175925	0.230719	0.182208	0.218204	0.208509	0.175302	0.190802	0.231876	0.182470	0.222715	0.206664	0.175834	0.184005
	Std.	0.023508	0.034515	0.030617	0.023460	0.034654	0.031087	0.023409	0.034566	0.031175	0.023128	0.034979	0.031003	0.024103	0.034468	0.030930
0.3	Mean	0.276544	0.254610	0.259019	0.294540	0.261357	0.291414	0.281279	0.256042	0.270628	0.294510	0.261887	0.294296	0.287409	0.260532	0.271124
	Std.	0.025167	0.036576	0.032294	0.024219	0.036274	0.031515	0.024826	0.037494	0.032784	0.024274	0.036355	0.031631	0.025743	0.035364	0.031711
0.4	Mean	0.347163	0.327971	0.334992	0.351028	0.331667	0.355215	0.345771	0.327742	0.341323	0.349404	0.331373	0.356389	0.358304	0.332396	0.346412
	Std.	0.025945	0.037171	0.032610	0.023931	0.036567	0.031289	0.025397	0.037093	0.032332	0.023891	0.036176	0.030782	0.026420	0.034008	0.030523
0.5	Mean	0.403169	0.384160	0.393219	0.395565	0.386959	0.404183	0.398019	0.385324	0.397667	0.393529	0.387018	0.404992	0.413473	0.386695	0.402586
	Std.	0.024060	0.034560	0.029745	0.022737	0.034371	0.028729	0.023627	0.034818	0.029743	0.023060	0.034661	0.028778	0.024246	0.029042	0.026283
Average	Err. mean	**0.031039**	**0.015544**	**0.016221**	**0.044614**	**0.018528**	**0.028196**	**0.035524**	**0.016309**	**0.019214**	**0.046519**	**0.019023**	**0.030796**	**0.027227**	**0.014231**	**0.012679**
Average	Std.	**0.018420**	**0.027762**	**0.029531**	**0.018389**	**0.027889**	**0.029086**	**0.018484**	**0.028070**	**0.029562**	**0.018465**	**0.028123**	**0.029117**	**0.018453**	**0.026460**	**0.028867**

Bold values are simple averages of the estimated statistics.

Table 9.3 Table of the estimation convergence

	Models	ARFIMA		ARFIMA		ARFIMA		
	p,d,q	0.5,0.3,0	0.5,−0.3,0	0,0.2,0.2	0,−0.2,0.2	0.5,0.1,0.2	0.5,−0.1,0.2	Average
Method	Steps	0.3	−0.3	0.2	−0.2	0.1	−0.1	error
RS	1	0.282508	−0.120719	0.188014	−0.117429	0.212298	0.044687	**0.045017**
	2	0.259761	−0.153673	0.180409	−0.126846	0.187427	0.007140	**0.036472**
	3	0.255753	−0.158888	0.180165	−0.127535	0.183607	0.000085	**0.035167**
	4	0.254986	−0.159766	0.180157	−0.127585	0.182974	−0.001351	**0.034939**
	5	0.254837	−0.159916	0.180157	−0.127588	0.182868	−0.001648	**0.034897**
	10	0.254801	−0.159946	0.180157	−0.127589	0.182847	−0.001725	**0.034887**
	15	0.254801	−0.159946	0.180157	−0.127589	0.182847	−0.001725	**0.034887**
	20	0.254801	−0.159946	0.180157	−0.127589	0.182847	−0.001725	**0.034887**
	25	0.254801	−0.159946	0.180157	−0.127589	0.182847	−0.001725	**0.034887**
	30	0.254801	−0.159946	0.180157	−0.127589	0.182847	−0.001725	**0.034887**
	45	0.254801	−0.159946	0.180157	−0.127589	0.182847	−0.001725	**0.034887**
AV	1	0.253092	−0.273630	0.140896	−0.226125	0.142337	−0.072987	**0.016325**
	2	0.243741	−0.305453	0.138629	−0.234143	0.132896	−0.094750	**0.016018**
	3	0.243174	−0.310923	0.138612	−0.234419	0.132289	−0.097701	**0.016115**
	4	0.243139	−0.311952	0.138612	−0.234429	0.132249	−0.098133	**0.016136**
	5	0.243137	−0.312149	0.138612	−0.234429	0.132246	−0.098197	**0.016140**
	10	0.243136	−0.312196	0.138612	−0.234429	0.132246	−0.098208	**0.016141**
	15	0.243136	−0.312196	0.138612	−0.234429	0.132246	−0.098208	**0.016141**
	20	0.243136	−0.312196	0.138612	−0.234429	0.132246	−0.098208	**0.016141**
	25	0.243136	−0.312196	0.138612	−0.234429	0.132246	−0.098208	**0.016141**
	30	0.243136	−0.312196	0.138612	−0.234429	0.132246	−0.098208	**0.016141**
	45	0.243136	−0.312196	0.138612	−0.234429	0.132246	−0.098208	**0.016141**

Bold values are simple averages of the estimated statistics

OT	1	0.293221	−0.193101	0.176548	−0.204196	0.182024	−0.025177	**0.026017**
	2	0.258793	−0.252295	0.164501	−0.224179	0.139805	−0.087479	**0.014491**
	3	0.252460	−0.267381	0.163916	−0.226019	0.131268	−0.104351	**0.013228**
	4	0.251173	−0.271671	0.163887	−0.226188	0.129355	−0.109562	**0.013049**
	5	0.250907	−0.272926	0.163886	−0.226204	0.128917	−0.111232	**0.013013**
	10	0.250837	−0.273451	0.163886	−0.226206	0.128787	−0.112031	**0.013002**
	15	0.250837	−0.273452	0.163886	−0.226206	0.128787	−0.112034	**0.013002**
	20	0.250837	−0.273452	0.163886	−0.226206	0.128787	−0.112034	**0.013002**
	25	0.250837	−0.273452	0.163886	−0.226206	0.128787	−0.112034	**0.013002**
	30	0.250837	−0.273452	0.163886	−0.226206	0.128787	−0.112034	**0.013002**
	45	0.250837	−0.273452	0.163886	−0.226206	0.128787	−0.112034	**0.013002**
Average	1	0.276274	−0.195817	0.168486	−0.182583	0.178887	−0.017826	**0.026713**
	2	0.253173	−0.238164	0.160906	−0.195481	0.152101	−0.059895	**0.018172**
	3	0.249850	−0.246228	0.160683	−0.196336	0.148224	−0.068243	**0.016915**
	4	0.249335	−0.247898	0.160676	−0.196393	0.147621	−0.070068	**0.016671**
	5	0.249254	−0.248250	0.160676	−0.196396	0.147526	−0.070474	**0.016619**
	10	0.249239	−0.248344	0.160676	−0.196397	0.147508	−0.070590	**0.016605**
	15	0.249239	−0.248344	0.160676	−0.196397	0.147508	−0.070590	**0.016605**
	20	0.249239	−0.248344	0.160676	−0.196397	0.147508	−0.070590	**0.016605**
	25	0.249239	−0.248344	0.160676	−0.196397	0.147508	−0.070590	**0.016605**
	30	0.249239	−0.248344	0.160676	−0.196397	0.147508	−0.070590	**0.016605**
	45	0.249239	−0.248344	0.160676	−0.196397	0.147508	−0.070590	**0.016605**
	p,d,q	**0.5,0.3,0**	**0.5,-0.3,0**	**0,0.2,0.2**	**0,-0.2,0.2**	**0.5,0.1,0.2**	**0.5,-0.1,0.2**	

Table 9.4 Data statistical properties

Product name	Days	Mean	Variance	Skweness	Kurtosis	KPSS	ADF
BRENT CRUDE OIL INDEX	4383	62.83	1213	0.2473	1.7393	0.01	0.57
CBOT CORN FUTURE	11,356	295.51	16,622	1.8275	6.4470	0.01	0.40
CBOT OATS FUTURE	11,452	180.07	6459	1.2847	4.2692	0.01	0.29
CBOT SOYBN FUTURE	11,457	713.43	79,087	1.3174	4.4276	0.01	0.44
CBOT SOYBN OIL FUTURE	10,191	26.94	110	1.5061	4.8304	0.01	0.27
CBOT WHEAT FUTURE	10,992	403.76	23,451	1.5125	5.5088	0.01	0.35
CL CRUDE OIL NYMEX	6624	46.11	979	0.8023	2.2901	0.01	0.41
CME FEEDER CATTLE INDEX	4805	109.85	1518	1.5069	5.1426	0.01	1.00
CT COTTON NYBOT	10,080	68.15	368	2.6049	16.5470	0.01	0.34
FC CATTLE FEEDER CME	6446	102.04	1283	1.7922	6.2853	0.01	0.99
GENERIC 1ST FUTURE GOL	10,093	534.29	153,494	1.6741	4.7814	0.01	0.84
HG COPPER NYMEX	6685	181.52	13,005	0.7429	1.9573	0.01	0.50
HO HEATING OIL NYMEX	7296	125.22	8782	0.9644	2.5198	0.01	0.46
LA ALUMINUM FUTURE	4486	1874.19	217,237	0.7035	2.5746	0.01	0.47
LA CRUDE OIL SPOT	6163	41.47	1024	0.8416	2.2766	0.01	0.46
LB LUMBER CME	6646	282.54	4487	0.2529	2.4577	0.01	0.39
LC CATTLE LIVE CME	8347	81.99	539	1.6916	5.4931	0.01	0.88
HOGS, LEAN FUTURE CME	7385	61.15	301	1.0170	4.1324	0.01	0.43
LME 3M COPPER FUTURE	7411	3777.42	6,095,279	0.8835	2.1950	0.01	0.53
LME ALUMINUM SPOT	7034	1785.20	224,536	0.9541	3.6849	0.01	0.35
LME COPPER SPOT	7395	3805.70	6,104,507	0.8680	2.1900	0.01	0.52
LME PL INDEX	5873	884.98	248,848	0.5268	1.9135	0.01	0.58
NYBOT COCOA FUTURE	11,209	1825.89	674,508	0.7549	3.2154	0.01	0.48
NYBOT COFFEE FUTURE	10,648	125.17	2660	0.7929	3.6491	0.01	0.21
NYBOT OR JUICE FUTRUE	11,348	109.78	1521	0.2347	2.4749	0.01	0.36
NYBOT SUGAR FUTURE	11,312	12.01	48	2.1007	9.2338	0.01	0.13
NYMEX CRUDE FUTURE	8094	41.83	892	1.0768	2.8560	0.01	0.35
PA PALLADIUM NYMEX	7269	326.72	55,634	0.9770	2.6874	0.01	0.57
PL PLATINUM NYMEX	7941	107.68	1542	0.8977	3.2269	0.01	0.55
GSCI COPPER INDEX SPOT	7354	807.26	227,179	0.8662	2.3940	0.01	0.50
GSCI LEAN HOGS SPOT INDEX	9730	230.98	27,862	1.2550	3.0455	0.01	0.36
GSCI PALLADIUM INDEX ER	9983	98.53	376	0.8783	5.6554	0.01	0.70
SI SILVER NYMEX	1723	415.07	17,772	−0.9263	2.5756	0.01	0.21
WCE CANOLA FUTRUE	10,123	9.74	61	1.9738	6.6083	0.01	0.59
WTI CUSING CRUDE SPOT	8421	377.23	9224	0.9768	3.6844	0.01	0.35
XAG SILVER SPOT	8020	41.96	895	1.0688	2.8372	0.01	0.15
XAU GOLD SPOT	11,520	9.06	60	2.0231	6.9884	0.01	0.85
XPD PALLADIUM SPOT	10,521	527.46	156,339	1.6412	4.7355	0.01	0.60
XPT PLATINUM SPOT	5572	398.22	54,108	0.6520	2.1052	0.01	0.56
PH GOLD&SILVER INDEX	7441	813.35	228,838	0.8320	2.3314	0.01	0.23

Table 9.5 Stationarity test for the increments and log difference data

Product name	Diff KPSS	Diff ADF	Log KPSS	Log ADF
BRENT CRUDE OIL INDEX	0.0497	0.0010	0.1000	0.0010
CBOT CORN FUTURE	0.1000	0.0010	0.1000	0.0010
CBOT OATS FUTURE	0.1000	0.0010	0.1000	0.0010
CBOT SOYBN FUTURE	0.1000	0.0010	0.1000	0.0010
CBOT SOYBN OIL FUTURE	0.1000	0.0010	0.1000	0.0010
CBOT WHEAT FUTURE	0.1000	0.0010	0.1000	0.0010
CL CRUDE OIL NYMEX	0.1000	0.0010	0.1000	0.0010
CME FEEDER CATTLE INDEX	0.0162	0.0010	0.0345	0.0010
CT COTTON NYBOT	0.1000	0.0010	0.1000	0.0010
FC CATTLE FEEDER CME	0.1000	0.0010	0.1000	0.0010
GENERIC 1ST FUTURE GOL	0.1000	0.0010	0.1000	0.0010
HG COPPER NYMEX	0.1000	0.0010	0.1000	0.0010
HO HEATING OIL NYMEX	0.1000	0.0010	0.1000	0.0010
LA ALUMINUM FUTURE	0.1000	0.0010	0.1000	0.0010
LA CRUDE OIL SPOT	0.0825	0.0010	0.1000	0.0010
LB LUMBER CME	0.1000	0.0010	0.1000	0.0010
LC CATTLE LIVE CME	0.1000	0.0010	0.1000	0.0010
HOGS, LEAN FUTURE CME	0.1000	0.0010	0.1000	0.0010
LME 3M COPPER FUTURE	0.1000	0.0010	0.1000	0.0010
LME ALUMINUM SPOT	0.1000	0.0010	0.1000	0.0010
LME COPPER SPOT	0.1000	0.0010	0.1000	0.0010
LME PL INDEX	0.1000	0.0010	0.1000	0.0010
NYBOT COCOA FUTURE	0.1000	0.0010	0.1000	0.0010
NYBOT COFFEE FUTURE	0.1000	0.0010	0.1000	0.0010
NYBOT OR JUICE FUTRUE	0.1000	0.0010	0.1000	0.0010
NYBOT SUGAR FUTURE	0.1000	0.0010	0.1000	0.0010
NYMEX CRUDE FUTURE	0.1000	0.0010	0.1000	0.0010
PA PALLADIUM NYMEX	0.1000	0.0010	0.1000	0.0010
PL PLATINUM NYMEX	0.1000	0.0010	0.1000	0.0010
GSCI COPPER INDEX SPOT	0.1000	0.0010	0.1000	0.0010
GSCI LEAN HOGS SPOT INDEX	0.1000	0.0010	0.1000	0.0010
GSCI PALLADIUM INDEX ER	0.1000	0.0010	0.1000	0.0010
SI SILVER NYMEX	0.1000	0.0010	0.1000	0.0010
WCE CANOLA FUTRUE	0.1000	0.0010	0.1000	0.0010
WTI CUSING CRUDE SPOT	0.1000	0.0010	0.1000	0.0010
XAG SILVER SPOT	0.1000	0.0010	0.1000	0.0010
XAU GOLD SPOT	0.1000	0.0010	0.0100	0.0010
XPD PALLADIUM SPOT	0.1000	0.0010	0.1000	0.0010
XPT PLATINUM SPOT	0.1000	0.0010	0.0867	0.0010
PH GOLD&SILVER INDEX	0.1000	0.0010	0.1000	0.0010

Table 9.6 Hurst exponent for increments and log difference for our algorithm and the RS method

Product	Hurst (Diff)	Hurst (Log)	RS (Diff)	RS (Log)
GSCI LEAN HOGS SPOT INDEX	0.3571	0.8248	0.5267	0.6102
HOGS, LEAN FUTURE CME	0.4141	0.7618	0.5170	0.6938
LB LUMBER CME	0.4403	0.8436	0.5168	0.7831
CBOT OATS FUTURE	0.4432	0.7433	0.5111	0.7714
LC CATTLE LIVE CME	0.4540	0.7863	0.4911	0.7664
PH GOLD&SILVER INDEX	0.4574	0.7307	0.5720	0.8187
NYBOT OR JUICE FUTRUE	0.4583	0.7747	0.4956	0.7481
CBOT SOYBN FUTURE	0.4748	0.8460	0.5237	0.8569
CBOT WHEAT FUTURE	0.4775	0.8703	0.5265	0.8206
CBOT CORN FUTURE	0.4815	0.7140	0.5225	0.8234
XAG SILVER SPOT	0.4870	0.8652	0.5713	0.8575
SI SILVER NYMEX	0.4932	0.8535	0.5409	0.7506
LME ALUMINUM SPOT	0.4937	0.8674	0.5599	0.8439
GSCI PALLADIUM INDEX ER	0.4972	0.6825	0.5840	0.8604
NYBOT COFFEE FUTURE	0.4989	0.7853	0.5472	0.7741
CT COTTON NYBOT	0.5001	0.8385	0.5607	0.7909
WCE CANOLA FUTRUE	0.5062	0.7566	0.5477	0.8458
NYBOT SUGAR FUTURE	0.5074	0.8636	0.5781	0.8276
NYBOT COCOA FUTURE	0.5102	0.8558	0.5232	0.8161
GSCI COPPER INDEX SPOT	0.5300	0.8726	0.5201	0.7907
CBOT SOYBN OIL FUTURE	0.5325	0.8225	0.5531	0.8265
LME 3M COPPER FUTURE	0.5377	0.8831	0.5549	0.8650
LME COPPER SPOT	0.5406	0.8488	0.5463	0.8664
HG COPPER NYMEX	0.5409	0.8473	0.5362	0.8190
HO HEATING OIL NYMEX	0.5412	0.8075	0.5542	0.8184
XPT PLATINUM SPOT	0.5487	0.8378	0.5066	0.7927
NYMEX CRUDE FUTURE	0.5512	0.8590	0.5470	0.8426
LA ALUMINUM FUTURE	0.5522	0.8488	0.5616	0.7610
PL PLATINUM NYMEX	0.5542	0.8440	0.5389	0.8776
CL CRUDE OIL NYMEX	0.5561	0.8287	0.5601	0.8280
LME PL INDEX	0.5590	0.8324	0.5667	0.8091
WTI CUSING CRUDE SPOT	0.5605	0.7799	0.5793	0.8755
LA CRUDE OIL SPOT	0.5676	0.8323	0.5815	0.8053
XPD PALLADIUM SPOT	0.5724	0.8036	0.5803	0.8157
XAU GOLD SPOT	0.5770	0.8346	0.5962	0.8081
PA PALLADIUM NYMEX	0.5812	0.8017	0.5967	0.7898
BRENT CRUDE OIL INDEX	0.5948	0.8693	0.6111	0.7835
FC CATTLE FEEDER CME	0.6031	0.7878	0.5740	0.7312
GENERIC 1ST FUTURE GOL	0.6192	0.8914	0.5745	0.8676
CME FEEDER CATTLE INDEX	0.6560	0.6680	0.6556	0.6883

Table 9.7 Output example for "CBOT Corn Futrue"

Model	NoS	d	C	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	MA_1	MA_2	MA_3	MA_4	MA_5	MA_6	Var	AICBIC
ARMA(0_0)	9	−0.0124	0.0260													39.3506	147,885
ARMA(0_1)	9	−0.0143	0.0263							0.0341						39.3128	147,874
ARMA(0_2)	9	−0.0109	0.0258							0.0309	−0.0358					39.2620	147,856
ARMA(0_3)	9	−0.0136	0.0261							0.0325	−0.0342	0.0215				39.2473	147,859
ARMA(0_4)	9	−0.0151	0.0263							0.0340	−0.0337	0.0217	0.0099			39.2447	147,869
ARMA(0_5)	9	−0.0169	0.0266							0.0360	−0.0331	0.0224	0.0101	0.0110		39.2409	147,878
ARMA(0_6)	9	−0.0166	0.0265							0.0358	−0.0334	0.0224	0.0101	0.0108	-0.0019	39.2406	147,889
ARMA(1_0)	9	−0.0143	0.0254	0.0317												39.3155	147,876
ARMA(1_1)	9	−0.0130	0.0478	−0.8256						0.8571						39.2305	147,838
ARMA(1_2)	9	−0.0132	0.0480	−0.8311						0.8644	0.0027					39.2304	147,849
ARMA(1_3)	22	−0.0292	0.0061	0.7834						−0.7353	−0.0641	0.0441				39.2353	147,864
ARMA(1_4)	22	−0.0296	0.0055	0.8034						−0.7549	−0.0649	0.0473	−0.0048			39.2346	147,875
ARMA(1_5)	22	−0.0297	0.0060	0.7869						−0.7384	−0.0641	0.0469	−0.0072	0.0043		39.2342	147,886
ARMA(1_6)	9	−0.0150	0.0477	−0.8046						0.8403	−0.0066	−0.0063	0.0287	0.0182	-0.0115	39.1930	147,873
ARMA(2_0)	9	−0.0108	0.0260	0.0292	−0.0361											39.2647	147,858
ARMA(2_1)	9	−0.0132	0.0479	−0.8279	0.0025					0.8611						39.2303	147,849
ARMA(2_2)	9	−0.0133	0.0668	−1.1926	−0.2933					1.2288	0.3113					39.2294	147,860
ARMA(2_3)	22	−0.0297	0.0062	0.6703	0.1094					−0.6217	−0.1678	0.0392				39.2345	147,875
ARMA(2_4)	23	−0.0301	0.0098	−0.0548	0.7070					0.1043	−0.7304	−0.0037	0.0248			39.2043	147,868
ARMA(2_5)	22	−0.0289	0.0123	−0.0918	0.6538					0.1406	−0.6766	−0.0079	0.0269	0.0115		39.2006	147,877
ARMA(2_6)	22	−0.0286	0.0091	−0.0135	0.6889					0.0624	−0.7160	−0.0077	0.0338	0.0080	-0.0148	39.1942	147,885
ARMA(3_0)	9	−0.0139	0.0256	0.0331	−0.0353	0.0236										39.2468	147,859
ARMA(3_1)	21	−0.0288	0.0064	0.7886	−0.0632	0.0451				−0.7412						39.2369	147,865
ARMA(3_2)	23	−0.0298	0.0065	0.6322	0.0993	0.0384				−0.5835	−0.1556					39.2345	147,875
ARMA(3_3)	9	−0.0236	0.0019	1.1124	0.5938	−0.7739				−1.0759	−0.6573	0.8086				39.1138	147,816
ARMA(3_4)	29	−0.0236	0.0087	0.1848	0.6859	−0.1890				−0.1416	−0.7240	0.1919	0.0271			39.2014	147,878
ARMA(3_5)	22	−0.0292	0.0217	−0.9283	0.5713	0.5910				0.9778	−0.5527	−0.6202	0.0225	0.0387		39.1786	147,876

(continued)

Table 9.7 (continued)

Model	NoS	d	C	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	MA_1	MA_2	MA_3	MA_4	MA_5	MA_6	Var	AICBIC
ARMA(3_6)	23	−0.0293	0.0152	−0.6464	0.6341	0.4732				0.6962	−0.6293	−0.4959	0.0296	0.0269	−0.0177	39.1753	147,886
ARMA(4_0)	9	−0.0153	0.0255	0.0344	−0.0343	0.0239	0.0088									39.2448	147,869
ARMA(4_1)	23	−0.0296	0.0054	0.8298	−0.0658	0.0506	−0.0076			−0.7814						39.2353	147,875
ARMA(4_2)	23	−0.0301	0.0101	−0.0536	0.6723	−0.0027	0.0255			0.1036	−0.6963					39.2033	147,868
ARMA(4_3)	22	−0.0295	0.0091	0.1516	0.6568	−0.1591	0.0272			−0.1023	−0.6914	0.1602				39.2018	147,878
ARMA(4_4)	26	−0.0295	0.0210	−0.7063	−0.2627	0.5370	0.7013			0.7521	0.2792	−0.4971	−0.7063			39.1409	147,854
ARMA(4_5)	20	−0.0293	0.0252	−1.1929	0.0979	0.8248	0.3712			1.2426	−0.0651	−0.8331	−0.3598	0.0338		39.1664	147,880
ARMA(4_6)	22	−0.0298	0.0356	−1.5318	−0.1915	0.9762	0.5314			1.5821	0.2416	−0.9819	−0.5301	0.0474	0.0205	39.1627	147,890
ARMA(5_0)	9	−0.0172	0.0254	0.0362	−0.0336	0.0249	0.0089	0.0098								39.2419	147,879
ARMA(5_1)	22	−0.0298	0.0060	0.8062	−0.0649	0.0502	−0.0102	0.0049		−0.7577						39.2346	147,886
ARMA(5_2)	9	−0.0155	0.0786	−1.4302	−0.5348	−0.0068	0.0253	0.0330		1.4662	0.5520					39.1842	147,868
ARMA(5_3)	23	−0.0303	0.0232	−1.0292	0.5802	0.6413	0.0224	0.0340		1.0799	−0.5553	−0.6700				39.1788	147,876
ARMA(5_4)	24	−0.0300	0.0260	−1.2257	0.0382	0.8103	0.4221	0.0322		1.2761	−0.0024	−0.8154	−0.4104			39.1667	147,881
ARMA(5_5)	9	−0.0218	0.0049	0.3750	0.4367	0.7561	0.0398	−0.7882		−0.3414	−0.4740	−0.7267	−0.0720	0.8131		39.0154	147,804
ARMA(5_6)	12	−0.0203	0.0055	0.3384	0.4537	0.7355	0.0723	−0.8035		−0.3001	−0.4920	−0.7167	−0.1040	0.8262	0.0142	39.0127	147,814
ARMA(6_0)	9	−0.0178	0.0253	0.0368	−0.0333	0.0251	0.0092	0.0098	0.0029							39.2419	147,890
ARMA(6_1)	9	−0.0152	0.0473	−0.8129	−0.0054	−0.0047	0.0283	0.0172	−0.0103	0.8482						39.1951	147,874
ARMA(6_2)	9	−0.0150	0.0763	−1.3785	−0.4963	−0.0071	0.0243	0.0286	−0.0039	1.4141	0.5108					39.1835	147,879
ARMA(6_3)	23	−0.0294	0.0161	−0.7189	0.6249	0.4818	0.0297	0.0257	−0.0153	0.7685	−0.6162	−0.5048				39.1770	147,887
ARMA(6_4)	23	−0.0298	0.0367	−1.5488	−0.2478	0.9289	0.5493	0.0470	0.0196	1.5991	0.2987	−0.9319	−0.5458			39.1630	147,890
ARMA(6_5)	15	−0.0241	0.0056	−0.2772	1.4067	0.8868	−0.6423	−0.5562	−0.0224	0.3188	−1.4263	−0.9307	0.6657	0.5970		39.0517	147,837
ARMA(6_6)	11	−0.0228	0.0134	0.6885	−0.2067	0.2231	−0.3203	−0.5150	0.6510	−0.6485	0.1563	−0.1697	0.2802	0.5642	−0.6525	39.0821	147,866

This is an estimation sample results from our program for the increments of "CBOT Corn Futrue" data. "Model" column is for the models we estimated, here the size of the ARMA filter is used as an indicator of the model size. " NoS" is short for number of steps, it gives the number of steps the program needs to find the convergence of d, a minimum boundary of 9 is used for stability purpose. The value under "d, C, AR_1, AR_2, AR_3, AR_4, AR_5, AR_6, MA_1, MA_2, MA_3, MA_4, MA_5, MA_6 and Var" are models parameters for the fraction term, the constant, the AR and MA parameters and the variance of the error term, normal distribution error term is used here. The "AICBIC" is calculated by sum up AIC and BIC for model selection

Table 9.8 Best ARFIMA models for increments data

Product	NoS	Hurst	C	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	MA_1	MA_2	MA_3	MA_4	MA_5	MA_6	Var	AIC	BIC	Model
GSCI LEAN HOGS SPOT INDEX	45	0.3571	0.0001	1.7113	−1.5000	0.8920	−0.1077	−0.0473		−1.4371	1.2042	−0.6366				2.2036	36,236	36,316	ARMA(5_3)
HOGS, LEAN FUTURE CME	17	0.4141	0.0019	0.7227						−0.6232						1.6454	24,642	24,677	ARMA(1_1)
LB LUMBER CME	9	0.4403	0.0167	0.1137												38.2584	43,082	43,110	ARMA(1_0)
CBOT OATS FUTURE	9	0.4432	0.0374	−0.2377	0.7572	−0.4588	−0.9427	0.1343		0.3702	−0.7129	0.3590	0.9867			20.7434	67,243	67,331	ARMA(5_4)
LC CATTLE LIVE CME	11	0.4540	0.0037	0.6211	0.1967	0.6370	−0.7550			−0.5570	−0.2102	−0.6546	0.7219			0.7793	21,625	21,703	ARMA(4_4)
PH GOLD&SILVER INDEX	9	0.4574	−0.0069							0.0384						6.7933	37,753	37,781	ARMA(0_1)
NYBOT OR JUICE FUTRUE	12	0.4583	0.0030	0.7170	−0.0609	0.0408				−0.6392						5.0197	50,469	50,520	ARMA(3_1)
CBOT SOYBN FUTURE	9	0.4748	0.0610	1.1328	−0.9591					−1.0364	0.8646	0.0917				188.4235	92,541	92,600	ARMA(2_3)
CBOT WHEAT FUTURE	9	0.4775	0.0500	0.6032	−0.9352					−0.5968	0.9076					86.0859	80,174	80,225	ARMA(2_2)
CBOT CORN FUTURE	9	0.4815	0.0049	0.3750	0.4367	0.7561	0.0398	−0.7882		−0.3414	−0.4740	−0.7267	−0.0720	0.8131		39.0154	73,854	73,950	ARMA(5_5)
XAG SILVER SPOT	46	0.4870	0.0007	−0.3020	−0.4523	0.2692	0.2003	0.2935	0.5139	0.3235	0.4380	−0.1436	−0.1373	−0.2905	−0.5852	0.1261	8866	8976	ARMA(6_6)
SI SILVER NYMEX	12	0.4932	0.0014	0.8824	−0.5539	−0.6320	0.7851	−0.8875	0.0614	−0.8116	0.5398	0.6588	−0.7589	0.8369		0.1232	7560	7661	ARMA(6_5)
LME ALUMINUM SPOT	9	0.4937	−0.0303	−0.4664	0.7696	−0.3440	−0.8716			0.4432	−0.7705	0.3251	0.8663	0.0341		1171.3974	69,678	69,760	ARMA(4_5)
GSCI PALLADIUM INDEX ER	12	0.4972	0.1617							0.0968						59.5537	11,932	11,954	ARMA(0_1)
NYBOT COFFEE FUTURE	56	0.4989	0.0034	1.3817	−0.9266	0.0472	−0.0735	0.0284		−1.3832	0.9159					11.9550	56,444	56,517	ARMA(5_2)

(continued)

Table 9.8 (continued)

Product	NoS	Hurst	C	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	MA_1	MA_2	MA_3	MA_4	MA_5	MA_6	Var	AIC	BIC	Model
CT COTTON NYBOT	38	0.5001	0.0075	−0.9929	−0.5517	−0.9342	−0.6888	0.1164		1.1121	0.6730	0.9928	0.7883			1.8064	34,587	34,674	ARMA(5_4)
WCE CANOLA FUTRUE	20	0.5062	0.0350	0.1940	−0.3723	0.3440	−0.6707			−0.1457	0.3200	−0.3437	0.6149			27.5791	51,693	51,770	ARMA(4_4)
NYBOT SUGAR FUTURE	9	0.5074	0.0008	−0.1172	0.9133	−0.2854	−0.8390	−0.0128		0.1447	−0.9186	0.2804	0.8743			0.1684	11,932	12,020	ARMA(5_4)
NYBOT COCOA FUTURE	99	0.5102	0.1229	1.0011	−1.1982	0.7877	−0.0893	0.0353	−0.0603	−0.9582	1.1563	−0.7196				1670.8560	114,586	114,674	ARMA(6_3)
GSCI COPPER INDEX SPOT	9	0.5300	0.0687	0.0293	−0.1408	0.2269	−0.6932	−0.2605	−0.3342	−0.1101	0.1227	−0.2409	0.7311	0.2105	0.2618	24.9348	58,931	59,039	ARMA(6_6)
CBOT SOYBN OIL FUTURE	9	0.5325	−0.0015	−0.5404	0.2072	0.4884	0.2638	−0.5212	−0.8577	0.5549	−0.2136	−0.4858	−0.2666	0.5324	0.8791	0.2249	13,742	13,851	ARMA(6_6)
LME 3M COPPER FUTURE	10	0.5377	0.4899	0.2071	0.0744	0.2135	−0.8684	−0.1007	−0.0585	−0.3267	−0.0764	−0.2236	0.9199			5995.0634	85,512	85,602	ARMA(6_4)
LME COPPER SPOT	9	0.5406	0.5708	0.1588	0.1347	0.1918	−0.8905	−0.0931	−0.0633	−0.2756	−0.1491	−0.1870	0.9305			6679.6387	86,127	86,217	ARMA(6_4)
HG COPPER NYMEX	9	0.5409	0.0151	0.6902	−0.6924	0.7449	−0.8107	−0.1020		−0.7955	0.7516	−0.8241	0.8809			14.8029	37,005	37,086	ARMA(5_4)
HO HEATING OIL NYMEX	9	0.5412	0.0356	−0.7678	−0.7566					0.7010	0.6843	−0.0947				9.4157	37,076	37,132	ARMA(2_3)
XPT PLATINUM SPOT	9	0.5487	0.0458	1.7174	−1.1514	0.0949				−1.6853	1.0923	−0.0567				166.5297	59,189	59,251	ARMA(3_3)
NYMEX CRUDE FUTURE	14	0.5512	0.0015							−0.0998	−0.0529	−0.0203	0.0239	−0.0525		1.1718	24,266	24,322	ARMA(0_5)
LA ALUMINUM FUTURE	18	0.5522	0.0002	0.2923						−0.4553						881.9653	43,156	43,188	ARMA(1_1)
PL PLATINUM NYMEX	13	0.5542	0.0259	1.6915	−1.4509	0.3930	−0.0618			−1.6617	1.3506	−0.2916				186.4083	59,328	59,397	ARMA(4_3)
CL CRUDE OIL NYMEX	16	0.5561	0.0033							−0.1077	−0.0560	−0.0222	0.0223	−0.0540		1.4016	21,047	21,102	ARMA(0_5)

LME PL INDEX	9	0.5590	0.2358	−1.6029	−0.9235					1.5754	0.8304	−0.0933	−0.0500			224.0440	48,460	48,520	ARMA(2_4)
WTI CUSING CRUDE SPOT	13	0.5605	0.0014							−0.1000	−0.0444	−0.0177	0.0245	−0.0691		1.2074	24,284	24,340	ARMA(0_5)
LA CRUDE OIL SPOT	9	0.5676	0.0012	−0.0200	0.5322	−0.0073	0.6932	0.0621	−0.7063	0.0037	−0.5634	0.0084	−0.6753	−0.0695	0.7201	0.9351	17,103	17,204	ARMA(6_6)
XPD PALLADIUM SPOT	9	0.5724	0.0595	0.0111	−0.0680	−0.0790										80.4283	40,264	40,304	ARMA(3_0)
XAU GOLD SPOT	9	0.5770	0.1570	−0.4553	0.8986	−0.2638	−0.9075	−0.0840	−0.0467	0.3969	−0.9584	0.3083	0.9100			69.1736	74,450	74,544	ARMA(6_4)
PA PALLADIUM NYMEX	9	0.5812	0.0991	−1.4496	−0.3201	0.3072				1.4446	0.2401	−0.4829	−0.1067			64.4031	50,918	50,987	ARMA(3_4)
BRENT CRUDE OIL INDEX	9	0.5948	0.0123	−1.4340	−0.8352	−0.0374	−0.1026			1.5578	0.9511					1.0513	12,673	12,730	ARMA(4_2)
FC CATTLE FEEDER CME	9	0.6031	0.0029	0.1132	1.2041	0.2946	−0.8585			−0.1288	−1.2402	−0.2935	0.8735			0.8173	17,012	17,086	ARMA(4_4)
GENERIC 1ST FUTURE GOL	11	0.6192	0.0877	0.3229	−0.4136	−0.2123	0.5577	−0.3361	−0.0751	−0.3753	0.4201	0.1839	−0.6205	0.3752		71.2587	71,724	71,825	ARMA(6_5)
CME FEEDER CATTLE INDEX	9	0.6560	0.0209	−0.2430	−0.2890	0.1330	0.0400	−0.1381		0.2867	0.4261					0.3847	9064	9129	ARMA(5_2)

Table 9.9 Best ARFIMA models for log difference data

Product	NoS	Hurst	C	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	MA_1	MA_2	MA_3	MA_4	MA_5	MA_6	Var	AIC	BIC	Model
CME FEEDER CATTLE INDEX	9	0.6680	0.0001	0.7330	0.1604	−0.1850	0.2726	0.6150	−0.6737	−0.7809	−0.1425	0.1792	−0.2484	−0.5211	0.5911	0.0000	−39, 517	−39, 420	ARMA(6_6)
GSCI PALLADIUM INDEX ER	9	0.6825	0.0003	0.0562	1.0583	−0.6937	−0.2091	0.7351		−0.1682	−1.0616	0.8483	0.1859	−0.7864	0.0460	0.0002	−9993	−9916	ARMA(5_6)
CBOT CORN FUTURE	45	0.7140	0.0001	0.8754	0.0306	0.0038	0.0322	0.0316		−0.9520						0.0001	−68, 405	−68, 339	ARMA(5_1)
PH GOLD&SILVER INDEX	99	0.7307	0.0002	0.9765						−1.0788	0.0861	0.0016	0.0380			0.0002	−43, 705	−43, 649	ARMA(1_4)
CBOT OATS FUTURE	9	0.7433	0.0002	0.7887	0.8204	−0.6075	−0.0689			−0.9216	−0.7600	0.7434				0.0002	−63, 720	−63, 647	ARMA(4_3)
WCE CANOLA FUTRUE	9	0.7566	0.0021							−0.1674						0.0001	−55, 903	−55, 875	ARMA(0_1)
HOGS, LEAN FUTURE CME	9	0.7618	0.0026							−0.2873	−0.1026	−0.0503	−0.0316			0.0003	−39, 114	−39, 066	ARMA(0_4)
NYBOT OR JUICE FUTRUE	9	0.7747	0.0004	1.4779	−0.2954	−0.2791	−0.0033	−0.0118	−0.0346	−1.6113	0.4371	0.2865				0.0002	−64, 489	−64, 401	ARMA(6_3)
WTI CUSING CRUDE SPOT	99	0.7799	0.0003	−0.5738	0.5579	0.9815				0.5478	−0.5210	−0.9011	0.0497	−0.0121		0.0003	−43, 292	−43, 215	ARMA(3_5)
NYBOT COFFEE FUTURE	9	0.7853	0.0031							−0.1441	−0.0684					0.0003	−57, 152	−57, 116	ARMA(0_2)
LC CATTLE LIVE CME	9	0.7863	0.0000	1.6689	−0.6143	−0.0233	−0.0032	−0.0294		−1.8808	0.8823					0.0001	−57, 430	−57, 360	ARMA(5_2)
FC CATTLE FEEDER CME	9	0.7878	0.0005	0.4417						−0.6705						0.0000	−47, 282	−47, 248	ARMA(1_1)
PA PALLADIUM NYMEX	9	0.8017	0.0023							−0.1565	−0.0396					0.0002	−42, 217	−42, 183	ARMA(0_2)
XPD PALLADIUM SPOT	9	0.8036	0.0030	−0.1565	−0.0630											0.0002	−31, 690	−31, 657	ARMA(2_0)
HO HEATING OIL NYMEX	9	0.8075	0.0001	1.4596	−0.3940	−0.0864				−1.6815	0.6998					0.0003	−39, 675	−39, 620	ARMA(3_2)

CBOT SOYBN OIL FUTURE	9	0.8225	0.0001	1.2391	−0.2781					−1.5202	0.5516					0.0001	−64, 196	−64, 146	ARMA(2_2)
GSCI LEAN HOGS SPOT INDEX	9	0.8248	0.0004	0.9652	−0.1812					−1.1925	0.3217					0.0001	−63, 866	−63, 816	ARMA(2_2)
CL CRUDE OIL NYMEX	9	0.8287	0.0024							−0.2615	−0.0370					0.0003	−35, 949	−35, 915	ARMA(0_2)
LA CRUDE OIL SPOT	9	0.8323	0.0061	−1.6434	−1.2267	−0.1237	0.5491	0.3867	0.0496	1.4307	0.7750	−0.3382	−0.7801	−0.3876		0.0003	−32, 809	−32, 715	ARMA(6_5)
LME PL INDEX	9	0.8324	0.0013							−0.2143	−0.0430	−0.0362				0.0001	−38, 396	−38, 356	ARMA(0_3)
XAU GOLD SPOT	10	0.8346	0.0001	0.7861	0.8102	−0.6496				−0.9233	−0.7242	0.7867	−0.0837	0.0128		0.0001	−69, 014	−68, 935	ARMA(3_5)
XPT PLATINUM SPOT	88	0.8378	0.0001	0.2724	−0.0048	−0.0071	0.0002	0.9814	−0.2716	−0.3796	0.0363	0.0179	−0.0165	−0.9231	0.3542	0.0001	−49, 059	−48, 955	ARMA(6_6)
CT COTTON NYBOT	9	0.8385	0.0008	0.4713	0.0393					−0.7452						0.0002	−57, 666	−57, 622	ARMA(2_1)
LB LUMBER CME	9	0.8436	0.0014	0.3092	1.0956	−0.6553	−0.9090	0.4610	0.0140	−0.5768	−1.1111	0.9169	0.8426	−0.7152		0.0002	−37, 330	−37, 235	ARMA(6_5)
PL PLATINUM NYMEX	9	0.8440	0.0034	−1.2567	−1.0821	−0.1917	0.3626	0.6639		0.9986	0.7383	−0.1575	−0.5153	−0.6575	0.1306	0.0001	−47, 428	−47, 331	ARMA(5_6)
CBOT SOYBN FUTURE	18	0.8460	0.0001	0.7525	0.1137	0.0134	0.0417	0.0544	−0.0051	−0.9583						0.0001	−71, 787	−71, 713	ARMA(6_1)
HG COPPER NYMEX	9	0.8473	0.0003	−0.0549	0.8837					−0.2324	−0.9791	0.2514	0.0445	−0.0040	0.0480	0.0001	−40, 921	−40, 846	ARMA(2_6)
LA ALUMINUM FUTURE	9	0.8488	0.0008	0.4041						−0.6774						0.0001	−29, 309	−29, 277	ARMA(1_1)
LME COPPER SPOT	9	0.8488	0.0002	0.0504	1.1019	0.4229	−0.7187			−0.2890	−1.1632	−0.1865	0.8611	−0.1114		0.0001	−44, 771	−44, 688	ARMA(4_5)
SI SILVER NYMEX	9	0.8535	0.0001	0.7421	0.0649	0.0425	0.0151	0.0634		−0.9441						0.0002	−57, 456	−57, 391	ARMA(5_1)
NYBOT COCOA FUTURE	9	0.8558	0.0006	0.6958	−0.0322					−1.0230	0.1735					0.0002	−64, 984	−64, 933	ARMA(2_2)
NYMEX CRUDE FUTURE	9	0.8590	0.0000	0.6262	1.1769	−0.6278	−0.1827			−0.9121	−1.0607	0.9906	0.0645	−0.0747		0.0003	−44, 114	−44, 030	ARMA(4_5)
NYBOT SUGAR FUTURE	9	0.8636	0.0012	0.4201						−0.6666						0.0003	−58, 643	−58, 606	ARMA(1_1)
XAG SILVER SPOT	9	0.8652	0.0021	−0.2309	−0.1122											0.0002	−64, 893	−64, 856	ARMA(2_0)

(continued)

Table 9.9 (continued)

Product	NoS	Hurst	C	AR_1	AR_2	AR_3	AR_4	AR_5	AR_6	MA_1	MA_2	MA_3	MA_4	MA_5	MA_6	Var	AIC	BIC	Model
LME ALUMINUM SPOT	9	0.8674	0.0002	1.0289	−0.2256					−1.3163	0.4387					0.0001	−44, 602	−44, 554	ARMA(2_2)
BRENT CRUDE OIL INDEX	9	0.8693	0.0004	1.4208	−0.9879	0.2338	0.0358			−1.7298	1.3560	−0.4761				0.0001	−27, 026	−26, 962	ARMA(4_3)
CBOT WHEAT FUTURE	9	0.8703	0.0009	0.4417						−0.7054						0.0002	−64, 989	−64, 953	ARMA(1_1)
GSCI COPPER INDEX SPOT	9	0.8726	0.0001	1.3500	−0.3664	−0.0873				−1.6625	0.7314					0.0001	−60, 923	−60, 865	ARMA(3_2)
LME 3M COPPER FUTURE	9	0.8831	0.0001	1.5663	−0.6066					−1.8544	0.9855	−0.1015				0.0001	−46, 712	−46, 656	ARMA(2_3)
GENERIC 1ST FUTURE GOL	9	0.8914	0.0008							−0.2977	−0.0564	−0.0338				0.0001	−67, 870	−67, 827	ARMA(0_3)

References

1. Granger, C. W. J. (1966). The typical spectral shape of an economic variable. *Econometrica, 34*(1), 150–161.
2. Tieslau, M. A. (1992). *Strongly dependent economic time series: Theory and applications*. Springer.
3. Baillie, R. T., Chung, C. F., & Tieslau, M. A. (1996). Analysing inflation by the fractionally integrated ARFIMA-GARCH model. *Journal of Applied Econometrics, 11*(1), 23–40.
4. Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics, 73*(1), 5–59.
5. Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis, 1*(1), 15–29.
6. Whittle, P. (1951). *Hypothesis testing in time series analysis* (Vol. 4). Almqvist & Wiksells.
7. Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics, 53*(1), 165–188.
8. Crato, N., & Rothman, P. (1994). Fractional integration analysis of long-run behavior for US macroeconomic time series. *Economics Letters, 45*(3), 287–291.
9. Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis, 4*(4), 221–238.
10. Robinson, P. M. (1994). Semiparametric analysis of long-memory time series. *The Annals of Statistics, 22*(2), 515–539.
11. Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers, 116*, 770–808.
12. Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena, 31*(2), 277–283.
13. Lo, A. W. (1989). *Long-term memory in stock market prices*. National Bureau of Economic Research.
14. Smith, J., Taylor, N., & Yadav, S. (1997). Comparing the bias and misspecification in ARFIMA models. *Journal of Time Series Analysis, 18*(5), 507–527.
15. Boutahar, M., Marimoutou, V., & Nouira, L. (2007). Estimation methods of the long memory parameter: Monte Carlo analysis and application. *Journal of Applied Statistics, 34*(3), 261–301.
16. Peters, E. E. (1994). *Fractal market analysis: Applying chaos theory to investment and economics* (Vol. 24). John Wiley & Sons.
17. Racine, R. (2011). *Estimating the hurst exponent* (pp. 1–30). Mosaic Group.
18. Rea, W., Oxley, L., Reale, M., & Brown, J. (2009). Estimators for long range dependence: An empirical study. *Fractals, 17*(4), 405–419.
19. Reisen, V. A., & Fajardo, F. A. (2011). Robust estimation in time series with long and short memory properties. Preprint. arXiv:1112.6308.
20. Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics, 23*(5), 1630–1661.
21. Shimotsu, K., & Phillips, P. C. B. (2005). Exact local Whittle estimation of fractional integration. *The Annals of Statistics, 33*(4), 1890–1933.
22. Shimotsu, K. (2010). Exact local Whittle estimation of fractional integration with unknown mean and time trend. *Econometric Theory, 26*(2), 501–540.
23. Reisen, V., Abraham, B., & Lopes, S. (2001). Estimation of parameters in ARFIMA processes: A simulation study. *Communications in Statistics: Simulation and Computation, 30*(4), 787–803.
24. Butler, J. (1999). Simulation and estimation of fractionally integrated time series. *SAS Users Group International Conference, Miami Beach, Florida* (pp. 0–5).
25. Isham, V., Keiding, N., Louis, T., Reid, N., Tibshirani, R., Tong, H., Hammersley, J. M., & Handscomb, D. C. (2004). *Monographs on statistics and applied probability* (p. 452). Hand The 101.

26. Montanari, A., Rosso, R., & Taqqu, M. S. (1997). Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. *Water Resources Research, 33*(5), 1035–1044.
27. Mandelbrot, B. B., & Taqqu, M. S. (1979). *Robust R/S analysis of long run serial correlation*. IBM Thomas J. Watson Research Division.
28. Beran, J. (1994). *Statistics for long-memory processes* (Vol. 61). CRC Press.
29. Beran, J. (1989). A test of location for data with slowly decaying serial correlations. *Biometrika, 76*(2), 261–269.
30. Jensen, A. N., & Nielsen, M. Ø. (2014). A fast fractional difference algorithm. *Journal of Time Series Analysis, 35*(5), 428–436.
31. Power, G. J., & Turvey, C. G. (2010). Long-range dependence in the volatility of commodity futures prices: Wavelet-based evidence. *Physica A: Statistical Mechanics and its Applications, 389*(1), 79–90.
32. Turvey, C. G. (2007). A note on scaled variance ratio estimation of the Hurst exponent with application to agricultural commodity prices. *Physica A: Statistical Mechanics and its Applications, 377*(1), 155–165.
33. Assa, H. (2016). Financial engineering in pricing agricultural derivatives based on demand and volatility. *Agricultural Finance Review, 76*(1).

Chapter 10
Examining the Impact of Weather Factors on Agricultural Market Price Risk: An XAI Approach

Muhathaz Gaffoor and Hibob Assa

Abstract In this chapter, we explore the application of machine learning (ML) and deep learning (DL) techniques to forecast commodity price volatility, emphasizing the integration of climatic data and financial variables. We use an XAI method, namely the Shapley interpretation method, to explain the impact of different variables on the agricultural price risk. As a preliminary consideration, agricultural businesses are supposed to be significantly influenced by environmental factors, particularly climatic anomalies such as El Niño and La Niña. Therefore, understanding their impact is crucial for effective market prediction and risk management. We discuss various predictive models, including time series analysis, machine learning models, and recurrent neural networks (RNNs) , highlighting their ability to handle large datasets and complex patterns. This chapter provides a comprehensive overview of how advanced computational methods can enhance the accuracy of volatility forecasts, to show the substantial benefits for farmers, investors, and policymakers. By integrating diverse data sources, including historical price data and environmental indicators, while illustrating the potential of ML and DL to study commodity trading and financial planning, we observe that climate features do not persistently rank among the top predictors of agricultural price risk in the US market. This might look surprising at first, as the common belief is the great influence of climate on any aspect of agriculture. This can be interpreted as a sign of adequately manageable risk in commodity market prices against natural phenomena.

M. Gaffoor
University of Essex, Colchester, UK

H. Assa (✉)
Model Library, London, UK
e-mail: h.assa@essex.ac.uk

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_10

10.1 Introduction

The financial markets are ever-evolving, influenced by many factors contributing to various asset price volatility, including commodities. Understanding and forecasting commodity price volatility is crucial for stakeholders such as farmers, insurance companies, financial institutions, investors, and policymakers, as it impacts decision-making across the agricultural and financial sectors. This chapter investigates the application of machine learning (ML) and deep learning (DL) techniques to predict the volatility of commodity prices. Historically, commodity price predictions were based on fundamental analysis considering factors like supply and demand, weather conditions, and geopolitical events. However, the advent of ML and DL had a substantial impact on the field, providing new methodologies that enhance prediction accuracy and efficiency. These advanced computational techniques are capable of handling large datasets and extracting complex patterns that are often not apparent through traditional statistical methods.

In this chapter, we explore how ML and DL models can be specifically tuned to forecast commodity volatility as the major component of risk management by incorporating a wide array of inputs, including historical price data (lagged, spillover, etc), weather information, and more nuanced data like sentiment indices. We discuss various models including but not limited to Time Series Analysis, Machine Learning Models, and Long Short-Term Memory (LSTM) networks.

As climate variability becomes more pronounced, understanding its impact on commodity prices and their risk management aspect i.e., volatility, has never been more important. We analyze the impact of weather patterns, especially those linked to large-scale climatic phenomena like El Niño and La Niña, on market volatility. Particularly, we employ the Shapley interpretation methods to understand the impact of the predictors in our models. An interesting finding, despite the stereotypical perception of the weather's impact on agricultural prices, is that they do not emerge persistently as the main influencing factor on the risk of agricultural prices in the US market. This can be because prices are determined in an open market where the equilibrium is reached in the open trades, and adverse weather conditions while influencing the production, cannot be the main price risk drivers.

10.2 Literature Review

Volatility is a measure of price fluctuation over a given period, mirroring the uncertainties surrounding price change magnitudes [1]. The range of the volatility in commodity prices has serious consequences for risk managers, insurance companies, producers, consumers, traders, and policymakers [2]. With a changing global climate, increasingly frequent extreme weather events, and the rise of climate risk, it is important to understand and predict the volatility. This literature review aims to focus on this complicated issue by concentrating on the impact of the Niño-Southern Oscillation (ENSO) on agricultural commodity realized volatility [3].

ENSO, a coupled atmosphere-ocean phenomenon in the Pacific Ocean can cause substantial climate variability globally. ENSO affects shifts in rainfall patterns, temperatures, and frequency of extreme weather events [2]. These changes have far-reaching impacts on agricultural production. However, it is not very clear how they would directly lead to considerable fluctuations in commodity price risk.

To have a good understating of the topic, this study not only focuses on the ENSO phenomenon and its effect, but we also investigate methodologies involved in forecasting such as Long short-Term Memory networks [4], volatility models, and various other mathematical features used in the modeling by exploring the complex relationship between climatic phenomena, agricultural productivity, and financial volatility [5].

El Niño-Southern Oscillation (ENSO) is a periodic phenomenon that occurs every 3–7 years and involves changes in the sea surface temperature of the Central and Eastern Pacific Ocean [3]. ENSO consists of 3 phases namely El Niño, La Niña, and neutral which typically last 1 year each [6]. El Niño is the warm phase where there is higher than average sea surface temperature in the Central and Eastern Pacific Ocean while La Niña is the cold phase where there is lower than average sea surface temperature in the Central and Eastern Pacific Ocean. The Neutral phase also called La Nad̃a is the time between El Niño phase and La Niña phase [7].

Though ENSO is a phenomenon that occurs in the Pacific Ocean, its effects are not limited to Pacific Ocean areas but have global effects due to the fluctuations in the Sea surface temperature (SST) which in turn affects tradewinds [8]. For example, increased rainfall can be observed in the eastern Pacific while drier conditions can be seen in the western Pacific areas during the El Niño phase and the opposite can be observed during the La Niña phase [9]. Similarly, the Northern side of North America experiences warmer winters while the Southern side experiences cooler and wetter winters during El Niño phase [10]. On the other hand during the El Niño phase, South America sees increased rainfall and experiences flooding [11]. The best-known case study for ENSO on Weather patterns and Extreme events is the 1997–1998 El Niño event. 1997–1998, North America has experienced one of the warmest and wettest winters with heavy rainfall resulting in flooding of California and Northern Gulf [12] while Severe drought and wildfires in Southeast Asia and Australia [13]. Similarly, 2010–2011 La Niña is one of the strongest ENSO events in recent history. It had caused heavy rainfall and flooding in Australia which was termed as the 'Big Wet' by the Australian Bureau of Meteorology [14].

ENSO has a huge role in controlling global weather patterns, whose impacts are often reflected as extreme weather events. The importance of understanding these events has increased as ENSO can become a huge climate risk in the future. As a clear implication of that, ENSO has a huge impact on Climate patterns which will affect agricultural productivity. As agricultural productivity directly impacts food security, international trade, and economic stability, we need to understand the complex relationship between ENSO, Weather patterns, and agricultural risk.

Temperature and rainfall are the key determinants of crop growth and yields. Studies show that as ENSO impacts both temperature and rainfall, it impacts commodity yield and prices both directly and indirectly. El Niño led drought conditions are associated with a decrease in Maize yield in Zimbabwe [15]

which shows a direct impact on ENSO on agricultural yields. ENSO-led climatic phenomena can also create a favorable environment for crop pests and diseases and thus can affect agricultural yields indirectly [15]. It is also important to note that the effects of ENSO show a high degree of variability. For example, During an El Niño event countries like India, the United States of America, and Australia experience increased yields of maize, soybean, and rice, while regions experience a decrease in the yield [16]. Similarly, the tropical region experiences more intense effects as compared to temperate regions [17]. The timing of the ENSO events also has an impact on the yield. If an ENSO event is considered a critical period like flowering or graining, then the impact may be higher, as compared to other less sensitive periods [18]. In addition, extreme weather-related ENSO events can disrupt post-harvest activities, leading to an overall decrease in the yield. For instance, flooding incidents can affect the quality of the harvested crops, increase the probability of post-harvest diseases, and sometimes even destroy the entire storage [19]. In addition to that, the effects of ENSO are not confined to crops alone but also affect the livestock as extreme weather conditions can affect Animal health and productivity. Drought associated with ENSO events due to an increase in temperature and decrease in precipitation can lead to insufficient water and feed resources, thus affecting livestock productivity negatively. On the other hand, increased rainfall and decreased temperature can increase the risk of diseases thus also affecting livestock productivity negatively [19].

10.2.1 Volatility of Agricultural Commodity Prices Due to ENSO

As we discussed earlier, ENSO affects agriculture globally indicating that it would in turn affect commodity price risk which is manifested as volatility in price. Volatility due to uncertainties can be induced directly or indirectly by the ENSO as discussed earlier. The reduction in the yield can cause a supply shock which leads to a sharp increase in the prices if the commodity holds a substantial market share. These supply shocks can also be compounded by other secondary effects such as speculative trading. Speculative trading includes activities such as hoarding the commodities expecting a price rise in the future, which can increase the price even before the supply shock occurs [20]. Farmers observing the increased price for a particular commodity often adjust their cultivation which causes imbalances in the market due to oversupply of some commodities while underproduction in some other commodities, leading to further volatility in the future [1].

10.2.2 Overview of Modeling Techniques Used to Predict Agricultural Commodity Prices and Volatility

Researchers have been using various modeling methods to predict commodity prices and volatility, especially considering the effect of climate. However, only a few of

them have given satisfactory results. Some methods employed by the researcher are discussed in the following paragraphs.

10.2.2.1 ARIMA, VAR and GARCH Models

Time series models such as ARIMA (Autoregressive integrated moving average) and VAR (Vector Auto Regression) have been extensively used in the modeling and predicting the commodity prices [21]. Though ARIMA has been used in modeling, because of its univariate nature, its application in modeling the effect of climate has been limited. VAR as an extension of the ARIMA model, captures linear interdependencies among multiple time series by modeling each variable as a function of the lagged values of all other variables [22].

GARCH (Generalized Autoregressive Conditional Heteroskedasticity) is a powerful model to forecast changing volatility over time. It is econometric modeling in particular used in financial time series data to address volatility clustering, which is a common phenomenon observed in financial markets. It assumes that larger volatility is to be followed by larger volatility while small volatility by smaller volatility [23].

In [24], the authors use ARIMA and GARCH to forecast the prices of major export crops of vegetables and fruits in Egypt from 2016 to 2030. They used the augmented dickey fuller (ADF) test and also found that the model was suitable for forecasting the prices for different crops. The authors have primarily used the ARIMA and GARCH models to forecast agricultural commodity prices. They employed producer daily price data for crops mostly from 1967 to 2015, using annual time series data. Their benchmark models were ARIMA(1,1,1), ARIMA(2,1,2), GARCH(1,1), etc., and the results showed that foretasted prices for these crops would increase from 2016 to 2030. The work is limited to traditional time series models and does not consider external factors like climate data. Using machine learning and deep learning methods to predict commodity volatility by incorporating price data and climate data could potentially fill this gap by capturing the complex non-linear relationships and external factors influencing price volatility, leading to more accurate forecasting.

10.2.2.2 Event Study Approaches

The event study methodology is anchored in pinpointing notable abnormal returns that are divergences from the expected returns in a normal circumstance. This approach combines an event window which is a predetermined period that captures the expected impact of the event estimation Window which is a period before the event window that acts as a reference point to determine the asset's 'normal' return and abnormal returns derived by contrasting the actual returns during the event window with the expected returns from the estimation window to make the prediction. This method has been significantly used In the agricultural commodities

sector, especially to understand the impact of significant climatic events, like El Niño, on market behaviors.

For instance, [25] studies the Impact of El Niño Phenomenon on the Volatility of the U.S. Soybean Futures Price Yield shows that the ENSO had a huge impact on the Soybean futures in some years between 2007–2018. The authors employed the GARCH-M model to estimate returns and the event study approach with parametric tests to examine the effect of El Niño events on the volatility of U.S. soybean futures prices. They utilized monthly soybean futures data from the Chicago Stock Exchange spanning January 2007 to June 2022, covering six El Niño occurrences. The benchmark models included GARCH(1,1), and the findings suggested that most El Niño events did not significantly influence the volatility of U.S. soybean futures prices. However, this study focused solely on traditional time series techniques and a single commodity, overlooking external factors like climate data and spillover effects.

10.2.2.3 STL Decomposition and GARCH-MIDAS Frameworks

STL Decomposition which stands for Seasonal-Trend decomposition using LOESS (locally estimated scatterplot smoothing) breaks down the time series into three primary components which are namely trend component, seasonal component, and remainder component. The decomposition in this method is achieved by iterative applying the LOESS methods to the time series in the following order. The trend component is extracted first and then followed by the seasonal component. Finally, the estimated trend and seasonal component are subtracted from the time series to obtain the remainder component thus allowing this method to not only capture the non-linear trends but also the change in the seasonal patterns over time . Due to the advantages of this method, it has widely been used in forecasting commodity prices and volatility.

GARCH-MIDAS Model stands for Generalized Autoregressive Conditional Heteroskedasticity- Mixed Data Sampling is an advanced econometric model that combines high-frequency data (e.g., daily or weekly) with low-frequency data (e.g., monthly or quarterly) to capture the volatility dynamics in time series data. The MIDAS component also allows the incorporation of macroeconomic variables which are normally measured in a higher time frame. This feature of this model made it suitable for its application in forecasting commodity price volatility too.

This method is used to break down the Southern Oscillation Index (SOI) used GARCH-MIDAS models and their extensions to forecast the volatility of major U.S. grain commodity futures using the ENSO data from the SOI, see [26]. They applied daily futures price data from January 1990 to October 2021 and monthly SOI data. Their benchmark model was the standard GARCH-MIDAS. They employed the STL decomposition method to extract trend, seasonality, and remainder components from the SOI and incorporated them into the GARCH-MIDAS framework. The results showed that models involving ENSO information, particularly the seasonal component, outperformed the benchmark in forecasting volatility.

10.2.2.4 Machine Learning and Deep Learning Methods

The recent advancement of Machine learning (ML) and deep learning (DL) have taken prime spots in forecasting various financial and economic phenomena. For instance, [27] used machine learning algorithms like Decision Trees, Random Forest, and Support Vector Machines to forecast corn prices using a dataset spanning from 1980 to 2018 having variables like production, consumption, exports, and ending stocks. They had not only found that the Random Forest algorithm outperformed other models in terms of prediction accuracy. They observe that variables like production and consumption were more influential in predicting corn prices than the others [27]. Similarly, [28] used DL algorithms like Long Short-Term Memory (LSTM) networks to forecast Soybean prices using a dataset spanning from 1990 to 2019 having variables like weather patterns, global demand, and geopolitical events. The study found that weather patterns, especially unexpected changes, played a crucial role in influencing soybean prices. In addition, the study also reflects LSTM models' superior predictive capabilities compared to traditional time series models [28]. It is also important to note that this development is limited to just forecasting prices but also can be seen for the forecasting of volatility. In [29], the authors used Machine learning techniques like Gradient Boosting and Neural Networks to predict the volatility of commodity prices in India. They used a dataset that comprised daily price data of various commodities from 2005 to 2020 and found that Gradient Boosting showed better performance in predicting volatility compared to Neural Networks. The study also emphasized the role of external shocks, such as policy changes and international events, inducing volatility in the agricultural commodity market.

Similarly, [5] used the Random Forest algorithm to examine the predictive value of El Niño and La Niña weather episodes for the subsequent realized variance of 16 agricultural commodity prices. They used high-frequency data between 2009 and 2020 and also estimated the realized variance, realized skewness, realized kurtosis, realized jumps, realized upside, and downside tail risks to capture potential nonlinear links between El Niño and La Niña and the subsequent realized variance. They found that El Niño and La Niña can help in predicting the volatility in a longer horizon. This work is the most similar to our paper so it warrants some comparison between the two to further emphasize the differences in methodology and also the results. It is important to note that unlike our paper they used climate data as a categorical feature to identify whether it was an El Niño phase or La Niña phase in their machine learning model thus limiting the models' ability to capture non-linear relationships. In addition, unlike us using more comprehensive weather data, they used only Niño 3.4 index representing the sea surface temperature (SST) anomalies in a specific region of the central equatorial Pacific Ocean as their climate feature also limited the model performance. We also used different machine learning methods and found that linear regression can outperform random forests for live cattle, emphasizing the differences when using live-stock data. Finally, we used XAI methodology to rank the features and found out weather factors are not consistently ranked as the top explanatory features.

10.3 Methodology

In this study, we aim to examine the predictive value of a large set of predictors including spill-over effects, volatility lags, and particularly El Niño and La Niño weather episodes for the subsequently realized variance of 6 agricultural commodity prices. For that purpose, we use an XAI method called Shapley values to rank the predictors in our study.

10.3.1 Data Collection

The study used two main types of data namely Commodity price data and Climate data.Yahoo Finance API was used to fetch the price data of commodities 01/01/2000 to 20/01/2020 to ensure a comprehensive overview of the commodity market as well as to factor in the cross-volatility effect. To get a holistic understanding of the ENSO data the following geographical coordinates (5, -120), (5, −160), (5, −150), (5, −90), (5, 160), (0, −160), (0, −150), (0, −140), (0, −130), (0, −120), (0, −110), (0, −100), (0, −90), and (0, −80) will be fetched. This data is sourced from MultiScale Ultra High Resolution (MUR) Sea Surface Temperature (SST) dataset which is a comprehensive dataset that is created by merging multiple Level 2 satellite SST data sets such as NASA Advanced Microwave Scanning RadiometerEOS (AMSRE), JAXA Advanced Microwave Scanning Radiometer 2 (AMSR2) on GCOM-W1, Moderate Resolution Imaging Spectroradiometers (MODIS) on the NASA Aqua and Terra platforms etc (Chin,2017). which is available from 01/06/2002 to 20/01/2020. Initial data analysis was performed before the data prepossessing and modeling in order to grasp the fundamental characteristics of the dataset and statistical measures such as mean, median, and standard deviation were obtained. Various feature engineering techniques have been applied to both the commodity price dataset and climate dataset to get the final dataset to calculate the features used. These techniques were used to calculate both statistical and climate features. Statistical features used include realized rolling weekly realized volatility using the Garman-Klass estimator method, and rolling monthly and quarterly weekly realized volatility [28]. In the case of climatic data, unlike the traditional method of just using the Niño 3.4 Index, our study analyzes the anomalies between different regions such as the North Pacific, South Pacific, East Pacific, and West Pacific, and finally, two regions in the equatorial area to calculate the ENSO intensity.

10.3.2 Exploratory Data Analysis

The dataset includes weekly records of commodity prices across various sectors, including agricultural products, metals, and energy resources. Each entry captures

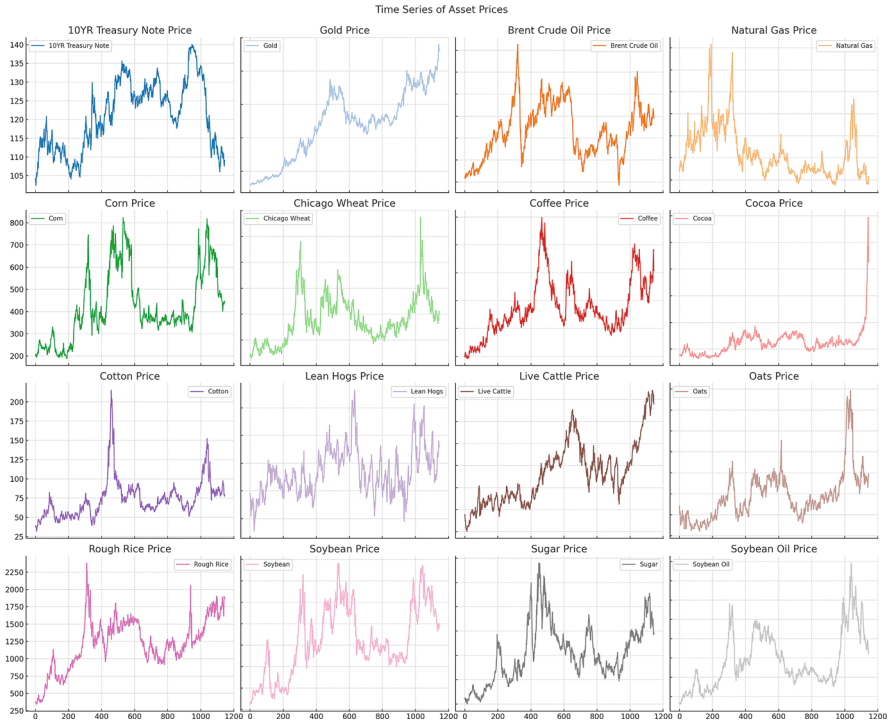


Fig. 10.1 Time series plot of prices

the closing prices for commodities such as corn, soybeans, gold, crude oil etc. at the end of each week, reflecting the dynamics of global trading activities. Spanning from January 1, 2000, to December 31, 2020, the dataset provides two decades of insights into market trends, highlighting seasonal variations, price volatility, and economic cycles that significantly impact commodity prices. This comprehensive dataset serves as a crucial resource for analyzing long-term trends, formulating risk management strategies, and exploring potential investment opportunities in the commodities market (Fig. 10.1).

In the exploration of price data across various commodities and financial instruments, we observe significant fluctuations indicative of diverse market conditions and economic factors. Gold, notably, exhibits an extensive price range from approximately $290 to over $2398, reflecting its sensitivity to global economic uncertainties and shifts in investor sentiment. Such broad fluctuations underscore gold's status as a "safe haven" during times of economic turmoil. The standard deviation for Gold prices stands at $511.61, highlighting its high volatility relative to other commodities. Natural Gas and Brent Crude Oil also demonstrate notable price variability. Natural Gas has ranged from $1.495 to $14.312, with a standard deviation of $2.23, and Brent Crude Oil has prices ranging from $16.94 to $145.29, with a standard deviation of $24.05. These figures suggest high variability

influenced by factors such as seasonal demand fluctuations, geopolitical tensions, and changes in global economic policies. Agricultural commodities like Corn and soybeans display considerable price variations influenced by environmental conditions, supply chain disruptions, and shifts in global demand patterns. Corn prices have varied from $189.75 to $824.50, with a standard deviation of $157.69, and Soybean prices from $451 to $1764.50, with a standard deviation of $316.28.

Volatility analysis of prices reveals distinct patterns of risk and stability across different commodities. Brent Crude Oil and Natural Gas are characterized by high annualized volatility, at approximately 0.446 each, emphasizing their susceptibility to rapid price changes influenced by external market shocks and geopolitical events. Conversely, more stable assets like the 10YR Treasury Note exhibit lower volatility, with an annualized volatility of approximately 0.053. This lower volatility reflects its role as a safer investment haven during volatile market periods. Live Cattle also shows lower volatility, with an annualized value of around 0.144, indicating less price fluctuation and thus a lower risk profile. Seasonal patterns are particularly evident in commodities like Natural Gas, which shows a marked increase in price and volatility during the winter months due to higher heating demands. This seasonal trend is crucial for developing predictive models that anticipate fluctuations in commodity prices and volatility based on seasonal changes. Extremes and anomalies in volatility data, such as the maximum observed values (e.g., Natural Gas volatility peaking at about 0.446 and Brent Crude Oil at about 0.446), show the impact of extraordinary market events and are critical in training predictive models to recognize and react to similar future incidents (Fig. 10.2).

The dataset comprises weekly observations of commodity prices, processed using the Garman-Klass method to calculate the volatility based on daily price data. This method provides a more accurate and robust measure of price volatility by utilizing the high, low, opening, and closing prices of commodities such as wheat, corn, soybeans, crude oil, and gold. Spanning from January 1, 2000, to December 31, 2020, the dataset offers 20 years of insights into market fluctuations, capturing economic influences, supply-demand shifts, and periodic trends that shape commodity prices. This extensive historical data, enhanced by the Garman-Klass volatility estimate, is invaluable for conducting robust trend analysis, understanding market cycles, and developing strategic approaches to commodity trading and risk management.

The dataset reveals distinct volatility profiles for each commodity. Corn shows mean volatility of about 0.014 with peaks up to 0.093, influenced by factors like seasonal growth cycles and market demands. Coffee stands out with a high mean volatility of 0.018 and significant spikes up to 0.179, highly reactive to climatic changes and international trade shifts. Lean Hogs and Live Cattle display volatility shaped by agricultural practices and market demands, with means around 0.023 and 0.014, respectively, and maximum values reaching 0.109 and 0.092. Oats and Rough Rice illustrate the impact of agricultural conditions on volatility, with Oats peaking at 0.144 and Rough Rice at 0.149. Their volatility is driven by factors like weather conditions and global supply chains. Gold, Brent Crude Oil, and Natural

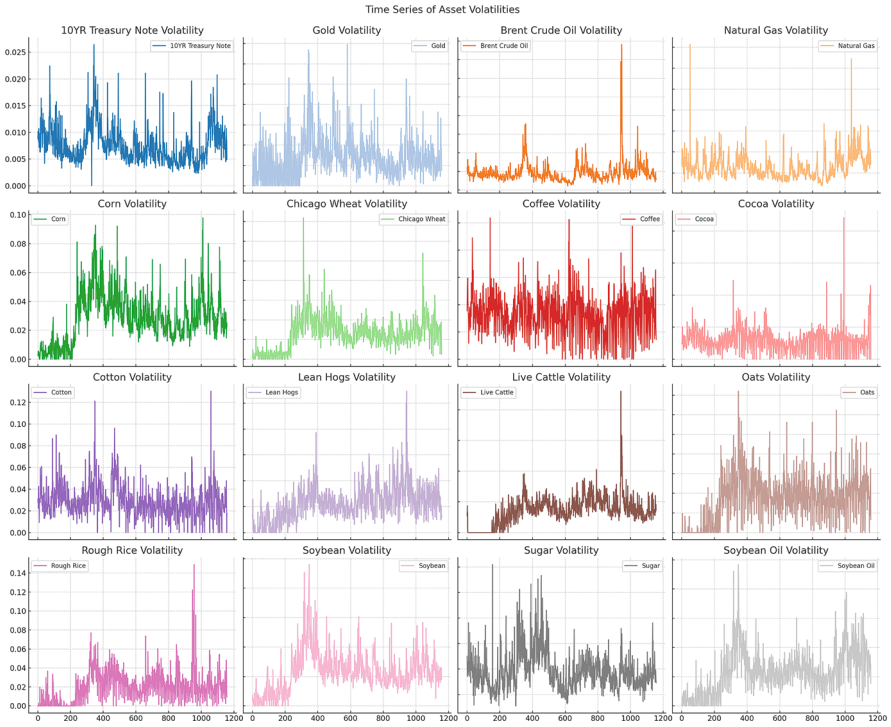


Fig. 10.2 Time series plot of Realised Volatility

Gas are particularly notable for their high volatility levels, reflecting their sensitivity to global economic fluctuations and geopolitical events.

Autocorrelation (ACF) and Partial Autocorrelation (PACF) analyses highlight the temporal dependencies in the volatility data. Agricultural commodities like Corn, Coffee, and Oats exhibit extended autocorrelations, suggesting that historical volatility significantly influences future volatility, indicating a persistent memory effect. In contrast, Lean Hogs and Live Cattle show shorter autocorrelation spans, pointing to a more immediate influence of recent events. Energy commodities such as Natural Gas and Brent Crude Oil also demonstrate strong autocorrelation over longer periods, underscoring the lasting impact of past volatility on future predictions. Gold shows significant autocorrelations across multiple lags, reinforcing its role as a safe haven during economic uncertainties.

The dataset contains daily measurements of sea surface temperature (SST) anomalies derived from various points across the Pacific Ocean, specifically designed to monitor changes in regions critical to understanding the El Niño-Southern Oscillation (ENSO) dynamics. It integrates SST data from multiple coordinates, including key Niño regions (5° latitude) and equatorial points (0° latitude), spanning a broad longitudinal range from 160° East to 80° West. The data covers the period from June 1, 2002, to July 31, 2023, providing over 21 years

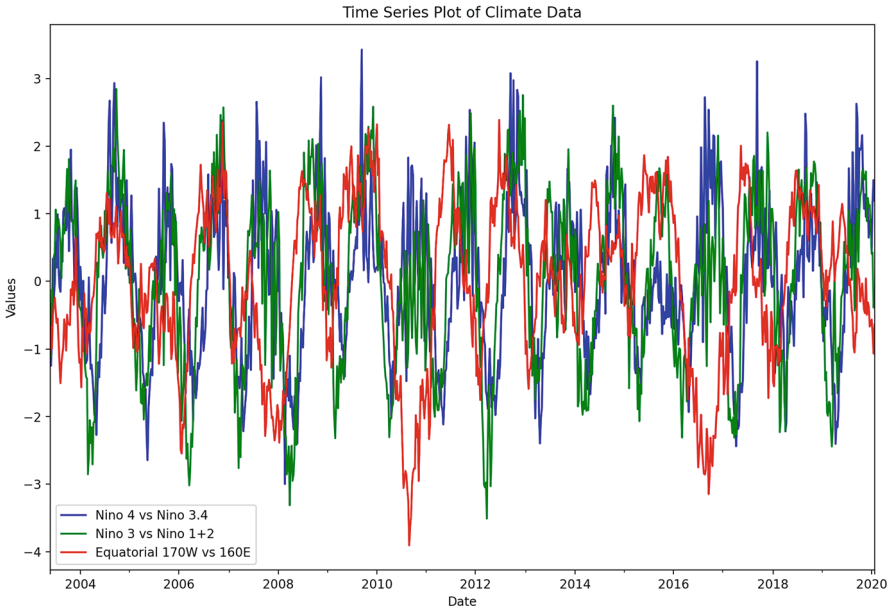


Fig. 10.3 Time series plot of Climatic Features

of temperature anomaly readings. Utilizing this comprehensive SST dataset, differences in anomalies are computed across significant ENSO monitoring regions: Niño 4 vs Niño 3.4, Niño 3 vs Niño 1+2, and across the equatorial Pacific from 170W to 160E (Fig. 10.3).

An initial examination reveals moderate positive correlations between the three temperature anomaly measurements: 0.632 between Niño 4 vs Niño 3.4 and Niño 3 vs Niño 1+2, 0.325 between Niño 4 vs Niño 3.4 and Equatorial 170W vs 160E, and 0.366 between Niño 3 vs Niño 1+2 and Equatorial 170W vs 160E. This suggests the anomaly patterns are related across regions, but not perfectly synchronous. The data also exhibits significant volatility, with standard deviations of 1.302 for Niño 4 vs Niño 3.4, 1.385 for Niño 3 vs Niño 1+2, and 1.318 for Equatorial 170W vs 160E. Visually examining the time series plots reveals prolonged periods of large positive and negative anomaly values. For instance, Niño 4 vs Niño 3.4 reached a peak of 3.430 on September 13, 2009, and a trough of −3.998 on February 24, 2008. Niño 3 vs Niño 1+2 hit 2.847 on September 26, 2004 and −3.510 on March 25, 2012. These prolonged deviations could signal El Niño or La Niña events capable of disrupting commodity supplies. However, no obvious cyclical seasonality can be detected at this stage.

Overall, this Pacific Ocean temperature anomaly data exhibits the characteristics relevant to predicting commodity volatility. With its long history, high volatility, correlation across regions, and ability to capture potential supply disruption scenarios, the data shows promise as a predictive signal for commodity markets.

10.3.3 Feature Engineering

In this study, we employed several feature engineering techniques to reprocess the data and select the most relevant features for the regression models. These techniques helped improve model performance by reducing noise, mitigating the curse of dimensionality, and enhancing the interpretability of the models.

10.3.3.1 Removing Highly Correlated Features

We removed features that were highly correlated with each other to mitigate the problem of multicollinearity, which can adversely affect the model's performance and interpretability [30]. Highly correlated features introduce redundancy in the data and can lead to unstable coefficient estimates and inflated standard errors [31]. The correlation between two features X and Y can be measured using the Pearson correlation coefficient (r), which is defined as:

$$r = \frac{\sum[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\sum(X - \mu_X)^2 \cdot \sum(Y - \mu_Y)^2}} \tag{10.1}$$

where μ_X and μ_Y are the means of X and Y, respectively. We removed features that had a correlation coefficient greater than a specified threshold (e.g., 0.7) with any other feature in the dataset.

10.3.3.2 Recursive Feature Elimination (RFE)

RFE is a feature selection technique that recursively eliminates features based on their importance or contribution to the target variable [32]. We used RFE with an XGBoost model as the estimator to identify and select the most informative features for each commodity. The RFE algorithm works as follows:

- Train a model on the initial set of features
- Compute the feature importance scores
- Remove the least important features
- Repeat the process with the remaining features until the desired number of features is reached

This approach helps reduce the dimensionality of the feature space, which can improve model performance, reduce overfitting, and enhance interpretability [33].

10.3.3.3 Data Scaling

Before fitting the models, we scaled the features using the MinMaxScaler from the scikit-learn library. This technique rescales the features to a common range, typically between 0 and 1, [34]. The MinMaxScaler transforms a feature X using the following formula:

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{10.2}$$

Scaling the features is particularly important for algorithms that are sensitive to the scale of the input variables, such as neural networks and tree-based models [35]. It can also improve the convergence speed and stability of gradient-based optimization algorithms.

By employing these feature engineering techniques, we aimed to enhance the quality of the input data and improve the performance of the regression models. Removing highly correlated features reduced multicollinearity and improved interpretability, RFE helped identify the most relevant features for each commodity, and data scaling ensured that all features contributed equally to the model's predictions.

10.3.3.4 Train-Test Split

We adopted a train-test split approach to separate the data into training and testing sets. This technique is widely used in machine learning to assess the performance of models on unseen data and mitigate overfitting. The train-test split was performed as follows:

- The data was sorted chronologically based on the date index.
- The last year of data was reserved as the test set.
- The remaining data was used as the training set.

In this study, we employed a strategic method to prepare our test data. Thus we isolated the subsequent week's data as our target variable. This approach is essential in time series forecasting as it allows our models to be tested against the very next sequence of events that follows. Testing the models on data that simulate future conditions ensures that the predictions are not only robust but also reflective of the models' ability to generalize to new, unseen data. This methodology underpins the reliability of our forecasting models in predicting future commodity price volatility, which is crucial for practical applications such as risk management and strategic planning in volatile markets. Our dataset contains weekly data from 2003–05-25 to 2020-01-26 of which data till 2018-01-25 is used as in-sample data and data from 2018-01-26 to 2020-01-26 is used as out-of-sample data.

10.3.4 Model Selection

1. **Linear Regression**: This is a simple yet powerful model that assumes a linear relationship between the input features and the target variable. Linear Regression models are interpretable and can provide insights into the relationship between features and the target variable [36]. However, they may struggle with capturing non-linear patterns in the data [37]. The linear regression model can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \tag{10.3}$$

   This is Eq. (10.3).
   where:

   - y is the target variable,
   - $x_1, x_2, \ldots, x_n$ are the input features,
   - β_0 is the intercept,
   - $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and
   - ϵ is the error term.

   [36]. The coefficients are typically estimated using the ordinary least squares (OLS) method, which minimizes the sum of squared residuals between the observed and predicted values.
2. **Random Forest**: This is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce overfitting [38]. Random Forest models are known for their ability to handle non-linear relationships and high-dimensional data, as well as their robustness to outliers and noise [39]. They are also relatively easy to tune and parallelize for efficient training [40]. The Random Forest algorithm can be represented as:
   The final prediction, $\hat{f}(x)$, is given by:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x) \tag{10.4}$$

   where B is the number of trees, and $\hat{f}_b(x)$ is the prediction of the bth tree [40].
3. **XGBoost** (Extreme Gradient Boosting): XGBoost is a powerful and efficient implementation of gradient-boosted decision trees [41]. It has been widely adopted in various machine-learning competitions and real-world applications due to its excellent performance and ability to handle a variety of data types and distributions [42]. XGBoost models are particularly effective when dealing with complex, non-linear relationships and high-dimensional data [43]. The objective function of XGBoost can be represented as:

The loss function $L(\Phi)$ is defined as:

$$L(\Phi) = \sum_i l(y_i, f_i(x_i)) + \sum_k \Omega(f_k) \tag{10.5}$$

where $l(y_i, \hat{y}_i)$ is the loss function that measures the difference between the true label y_i and the predicted label $\hat{y}_i$, and $\Omega(f_k)$ is the regularization term that penalizes the complexity of the model [44].

4. **LSTM** (Long Short-Term Memory): LSTM is a type of recurrent neural network (RNN) that is particularly well-suited for sequential data, such as time series [44]. Unlike traditional feedforward neural networks, LSTMs can effectively capture long-term dependencies and patterns in the data, making them a popular choice for forecasting tasks [45].LSTM architecture with Bidirectional layers and Dropout regularization was used to improve the model's performance and prevent overfitting [46] [47]. The LSTM cell can be represented as:
The equations for the LSTM cell are as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ \tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\ c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t \\ h_t &= o_t * \tanh(c_t) \end{aligned} \tag{10.6}$$

where f_t, i_t, o_t, and $\tilde{c}_t$ are the forget gate, input gate, output gate, and candidate cell state, respectively. W_f, W_i, W_o, and W_c are the weight matrices, and b_f, b_i, b_o, and b_c are the bias vectors [47].

10.3.5 Evaluation Metrics

In the we had analyzed the performance of metrics using R^2 value. Further, we had also used the Shapley value to understand how each feature is affecting the prediction.

10.3.5.1 R^2

The R^2 value which is often referred to as the coefficient of determination is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by independent variables in a regression model. It provides

an indication of the goodness of fit of a set of predictions to the actual values. R^2 value normally ranges between 0 and 1 and out of which 0 indicates the model does not explain any of the variability of the response data around its mean while 1 indicates that the model explains all the variability of the response data around its mean or regression predictions perfectly fit the data. Values of R^2 outside the range 0 to 1 can occur when the model fits the data worse than a horizontal hyperplane which indicates that the model is performing worse than random guessing, [48].

In mathematical terms let us introduce the following:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{10.7}$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{(Residual Sum of Squares)}$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(Total Sum of Squares)}$$

y_i : Actual value

$\hat{y}_i$: Predicted value

$\bar{y}$: Mean of the actual values

10.3.5.2 Shapley Value

Shapley values are a concept from cooperative game theory that has been adapted to explain the output of machine learning models as they provide a unified measure of feature importance by attributing the difference be- between the model's prediction and the average prediction to each feature in a fair and consistent manner, [49]. Thus they provide values of the average marginal contribution of that feature across all possible feature combinations. Considering feature i, the value associated to this feature is given by:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{10.8}$$

f : The model under consideration

S : A subset of features

N : The set of all features

$|S|$: Number of features in set S

$|N|$: Total number of features

$\phi_i(f)$ shows for the regression function or classifier f, how much the Shapley value gives to feature i, representing the magnitude of i's explainablity. We will use these values to rank our features. We use the SHAP library in Python for this purpose.

10.4 Result

10.4.1 Analysis of R^2 Values

In this part, we evaluate the performance of four machine learning models—Linear Regression, Random Forest, XGBoost, and LSTM—across six commodities namely corn, coffee, lean hogs, live cattle, oats, and rough rice by examining their R^2 values, which measure the proportion of variance in the dependent variable that can be explained by the independent variables. The results are presented in Fig. 10.5.

Corn The predictive performance for Corn shows a clear preference for ensemble and boosting methods over simpler models. XGBoost leads with an R^2 of 0.497, suggesting it most effectively captures the variability in Corn prices, possibly due to its proficiency in handling non-linear relationships among features like Corn Volatility and related commodities (e.g., Soybean and Wheat). Linear Regression follows closely with 0.483, while Random Forest slightly lags at 0.486. LSTM underperforms significantly, managing only 0.292, which might indicate its lesser capability in capturing the patterns in Corn price data over time or the need for more complex feature engineering.

Coffee For Coffee, Random Forest outperforms other models with an R^2 of 0.577, possibly benefiting from its ability to handle the interactions between different types of volatility and external market factors (e.g., Natural Gas prices and other commodities like Cocoa). XGBoost also shows a strong fit at 0.443, while Linear Regression has a moderate R^2 of 0.389. The LSTM's lower score of 0.336 may reflect challenges in modeling the sequential dependencies within the volatile coffee market.

Lean Hogs Lean Hogs appear to be challenging for all models, with relatively low R^2 values. Random Forest achieves the best among them at 0.163, potentially due to its capacity to model complex dependencies between features like Lean Hogs Volatility and external influences (e.g., Soybean Oil prices). Linear Regression provides limited predictability at 0.106. Surprisingly, XGBoost performs poorly, with an R^2 slightly negative at -0.014, indicating possible overfitting or inadequate model specification for this particular commodity. LSTM, with a mere 0.043, also struggles, likely due to the complex nature of agricultural market data.

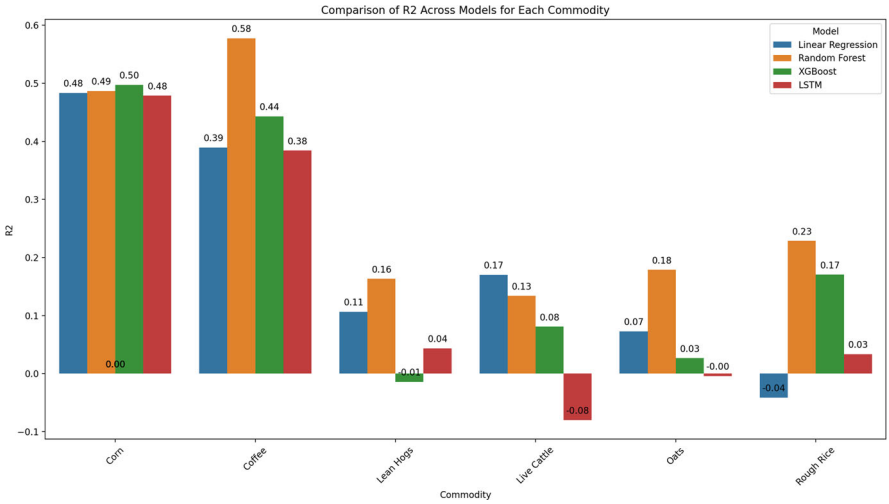


Fig. 10.4 Analysis of R^2 values

Live Cattle All models generally struggle with Live Cattle predictions. LSTM shows a marginally better R^2 of 0.177, which suggests some benefits from its ability to process sequential data, possibly capturing cyclic patterns not as apparent to other models. Linear Regression and Random Forest record R^2 values of 0.170 and 0.134, respectively, while XGBoost falls behind at 0.081. This could reflect the impact of external economic factors or feed prices (like Corn and Soybean), which may not be sufficiently modeled in simpler or even advanced ensemble techniques without specialized treatments.

Oats and Rough Rice The models exhibit limited effectiveness in predicting Oats and Rough Rice prices. For Oats, Random Forest performs the best at 0.179, with XGBoost again trailing at 0.027. The negative R^2 value for LSTM indicates substantial issues, either with the model fitting or underlying data inconsistencies. Rough Rice sees a slightly better performance from Random Forest at 0.229 and XGBoost at 0.171, while Linear Regression and LSTM display limited to poor predictive power, with R^2 values of -0.041 and 0.034, respectively.

This analysis demonstrates that while Random Forest regression generally offers superior accuracy for most commodities, specific contexts like Corn and Coffee likely due to its robust handling of heterogeneous data and complex interactions between features. The poor performance across the board with LSTM highlights potential mismatches in model applications or challenges in capturing the dynamic dependencies within commodity price series. This differential performance emphasizes the necessity of tailoring the choice of modeling techniques to the specific characteristics of the commodity and the economic context, ensuring the optimal alignment of model capabilities with the predictive requirements of commodity price data (Fig. 10.4).

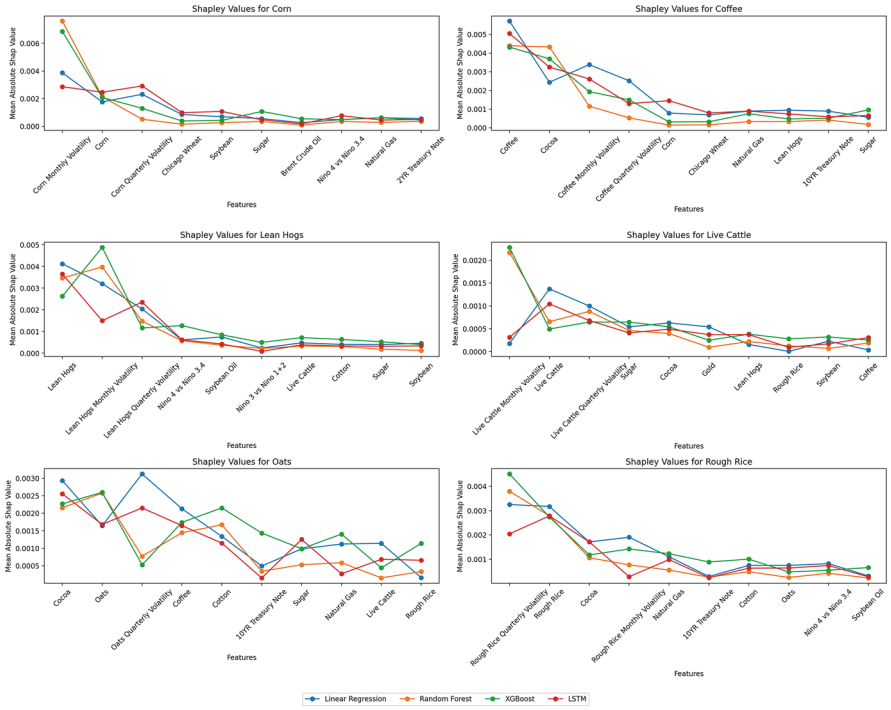


Fig. 10.5 Analysis of Shapley values

10.4.2 Analysis of Shapley Values

In this section, we investigate a detailed analysis of Shapley values to assess the influence of various predictors across multiple machine learning models for different commodities. The predictors are methodically grouped into three primary categories—market volatility indicators, climate and environmental factors, and economic factors. This structured approach aids in dissecting their impact on predicting commodity price volatility, offering a deeper comprehension of the intricate dynamics that control commodity markets. The analysis is based on results shown in Fig. 10.5.

Market Volatility Indicators remain pivotal in forecasting price movements, consistently displaying the highest Shapley values across all commodities. For example, Corn's Monthly Volatility in the Random Forest model records a notable Shapley value of 0.007600, signifying its paramount role in capturing immediate market reactions to supply-demand changes. Coffee's Monthly Volatility, with a Shapley value of 0.001162 in Random Forest, and Lean Hogs' Monthly Volatility at 0.004880 in XGBoost, also emphasize the acute sensitivity of these commodities to rapid market shifts. These indicators are crucial for accurately predicting price fluctuations driven primarily by market dynamics.

Climate and Environmental Factors also play a significant role, especially in agricultural commodities where output and quality are directly impacted by weather patterns. Corn's exposure to global climate variations is evidenced by the Niño 4 vs Niño 3.4 index, which holds a Shapley value of 0.000491 in XGBoost, pointing to its influence on Corn prices due to climate anomalies. While Lean Hogs and Live Cattle are less directly affected by typical environmental factors, the importance of Soybean Oil in Lean Hogs, with a Shapley value of 0.000747 in Linear Regression, suggests indirect climate impacts through feed costs. Oats and Rough Rice respond distinctly to environmental factors; for instance, Rough Rice's Quarterly Volatility shows a significant impact from seasonal weather changes, marked by the highest Shapley value of 0.003551 in XGBoost.

Economic Factors generally exhibit lower Shapley values compared to market volatilities and climate effects but still significantly impact commodity prices by linking them to broader economic conditions. For instance, the influence of Cocoa on Coffee prices, where Cocoa has a Shapley value of 0.002444 in Linear Regression, demonstrates how interconnected markets can influence pricing strategies. Economic indicators such as the 10YR Treasury Note also mirror broader economic conditions affecting commodities like Live Cattle, which in turn influences consumer spending patterns on meat products.

While in the majority the ranking of predictors is coherent, we also can see some inconsistency across different rankings by different methods. For example, for Live Cattle, it is clear that the two best prediction models i.e., linear regression and random forest show the impact of the lagged volatility, the other two models show the lagged prices are better predictors. This emphasized the importance of choosing the right model for the predictions.

10.5 Conclusion

This book chapter investigates the power of machine learning and deep learning techniques for forecasting volatility in commodity prices, with a keen focus on the impact of climatic events like El Niño-Southern Oscillation (ENSO). Through meticulous analysis, we have witnessed how diverse modeling approaches, from Random Forests and linear regression to XGBoost and LSTM networks, can effectively capture and predict the intricate patterns of market volatility driven by climatic shifts.

From a technical standpoint, Random Forests consistently demonstrate superior performance across various commodities. However, an intriguing finding emerges—for all commodities except rough rice, the sophisticated LSTM models fail to surpass the performance of simple linear regression.

By integrating machine learning with Explainable AI (XAI) techniques such as Shapley values, we can assess the predictive power of climatic data on price risk. Interestingly, our investigation reveals that lag and spill-over effects hold a significantly higher explanatory power compared to immediate weather conditions.

This could be attributed to the efficient hedging practices employed in commodity markets, like the US agricultural market, which mitigate the direct impact of weather fluctuations on prices.

It is crucial to recognize that the integration of production and price risk forms the foundation of revenue risk. Managing revenue risk effectively necessitates separate modeling approaches for price and production risks. While Shapley values demonstrate consistency across most models, for specific commodities like Live Cattle, we observe variations in feature ranking. This underlines the importance of selecting the optimal model for precise prediction in each scenario.

References

1. Zeng, D., Bobenrieth, E., & Wright, B. (2013). Stocks-to-use ratios and prices as indicators of vulnerability to spikes in global cereal markets. *Agricultural Economics*, *44*.
2. Wang, J., Liu, Z., Pan, X., & Peng, J. (2017). Volatility spillovers between agricultural commodity prices. *Sustainability*, *9*(12), 2247.
3. Timmermann, A., Oberhuber, J., Bacher, A., Esch, M., Latif, M., & Roeckner, E. (1999). Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature*, *398*(6729), 694–697.
4. Fischer, T., & Krauss, C. (2018). Deep volatility: Forecasting volatility with recurrent neural networks. *Journal of Empirical Finance*, *47*, 1–22.
5. Bonato, M., Cepni, O., Gupta, R., & Pierdzioch, C. (2022). El Niña, La Niña, and forecastability of the realized variance of agricultural commodity prices: Evidence from a machine learning approach. *Journal of Forecasting*, *42*.
6. McPhaden, M. J., Zebiak, S. E., & Glantz, M. H. (2006). ENSO as an integrating concept in earth science. *Science*, *314*(5806), 1740–1745.
7. Trenberth, K. E., & Fasullo, J. T. (2013). An apparent hiatus in global warming? *Earth's Future*, *1*(1), 19–32.
8. Philander, S. G., Holton, J. R., & Dmowska, R. (1989). *El Nino, La Nina, and the Southern Oscillation* (Vol. 46, 1st ed.). ISBN: 9780125532358. eBook ISBN: 9780080570983.
9. Ropelewski, C. F., & Halpert, M. S. (1987). Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, *115*(8), 1606–1626.
10. Ropelewski, C. F., & Halpert, M. S. (1986). North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Monthly Weather Review*, *114*(12), 2352–2362.
11. Wolter, K., & Timlin, M. S. (1998). Measuring the strength of ENSO events: How does 1997/1998 rank? *Weather*, *53*(9), 315–324.
12. Kovats, R. S., Bouma, M. J., Hajat, S., Worrall, E., & Haines, A. (2003). El Niño and health. *The Lancet*, *362*(9394), 1481–1489.
13. Glantz, M. H. (2001). *Currents of Change: Impacts of El Niño and La Niña on Climate and Society*. Cambridge University Press.
14. Funk, C., Dettinger, M. D., Michaelsen, J. C., Verdin, J. P., Brown, M. E., Barlow, M., & Hoell, A. (2008). Warming of the Indian Ocean threatens eastern and southern African food security but could be mitigated by agricultural development. *Proceedings of the National Academy of Sciences*, *105*(32), 11081–11086.
15. Cane, M. A., Eshel, G., & Buckland, R. W. (1994). Forecasting Zimbabwean maize yield using eastern equatorial Pacific sea surface temperature. *Nature*, *370*(6486), 204–205.

16. Iizumi, T., Luo, J. J., Challinor, A. J., Sakurai, G., Yokozawa, M., Sakuma, H., Brown, M. E., & Yamagata, T. (2014). Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nature Communications*, *5*, 4712.
17. Acevedo, M. F. (2016). Impact of El Niño on world crops. *Nature Climate Change*, *6*(10), 933–938.
18. Slingo, J. M., Challinor, A. J., Hoskins, B. J., & Wheeler, T. R. (2005). Introduction: food crops in a changing climate. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1463), 1983–1989.
19. Deutsch, C. A., Tewksbury, J. J., Huey, R. B., Sheldon, K. S., Ghalambor, C. K., Haak, D. C., & Martin, P. R. (2008). Impacts of climate warming on terrestrial ectotherms across latitude. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(18), 6668–6672.
20. Liu, J., & Devereux, J. (2008). Agricultural production, weather anomalies and rural poverty: evidence from smallholder farmers in Southeast Asia. *Journal of Development Studies*, *44*(8), 1141–1157.
21. Brockwell, P. J., & Davis, R. A. (2013). *Introduction to time series and forecasting*. Springer New York.
22. Lütkepohl, H. (2006). *New introduction to multiple time series analysis*. Springer.
23. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.
24. Agbo, H. M. S. (2023). Forecasting agricultural price volatility of some export crops in Egypt using ARIMA/GARCH model. *Review of Economics and Political Science*, *8*(2), 123–133. https://doi.org/10.1108/REPS-06-2022-0035.
25. Guo, W. (2023). Research on the impact of El Nino phenomenon on the volatility of the U.S. soybean futures price yield. *SHS Web of Conferences*, *170*, 03007. International Conference on Digital Economy and Management Science (CDEMS 2023) (pp. 1–6).
26. Su, Y., Liang, C., Zhang, L., & Zeng, Q. (2022). Uncover the response of the U.S grain commodity market on El Niño–Southern Oscillation. *International Review of Economics & Finance*, *81*, 98–112.
27. Shahzad, F., Wei, Z., & Khan, M. A. (2019). Forecasting agricultural commodity prices using machine learning algorithms: A case study of corn. *Journal of Agricultural Economics and Development*, *7*(2), 1–10.
28. Fan, K., Hu, Y., Liu, H., & Liu, Q. (2023). Soybean futures price prediction with dual-stage attention-based long short-term memory: A decomposition and extension approach. *Journal of Intelligent & Fuzzy Systems*, *45*(6), 10579–10602.
29. Gupta, R., & Jain, A. (2021). Volatility forecasting in agricultural commodities using machine learning: Evidence from Indian markets. *Journal of Agribusiness in Developing and Emerging Economies*, *11*(3), 234–250.
30. Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, *6*(2).
31. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46.
32. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422.
33. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer New York.
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.
35. Žliobaitė, I., Pechenizkiy, M., & Grama, I. (2016). Handling unconstrained drift in model-based regression with model selection techniques. *Data Mining and Knowledge Discovery*, *30*(4), 859–900.
36. Yan, X., & Su, X. (2009). *Linear regression analysis: Theory and computing*. World Scientific.

37. Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data mining, inference, and prediction*. Springer New York.
38. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
39. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
40. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958.
41. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).
42. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.
43. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.
44. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
45. Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A comparative analysis of forecasting financial time series using ARIMA, LSTM, and BiLSTM. Preprint. arXiv:1805.09851.
46. Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. In *International conference on artificial neural networks* (pp. 799-804). Springer.
47. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.
48. Dessain, J. (2023). *Improving the prediction of asset returns with machine learning by using a custom loss function*. *Advances in Artificial Intelligence Machine Learnning 3*(4), 1640–1653.
49. Li, M., Sun, H., Huang, Y., & Chen, H. (2024). Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, *4*(1), 2. Retrieved from https://doi.org/10.1007/s43684-023-00060-8.

Chapter 11
Textual Analysis in Agriculture Commodities Market

Navid Parvini

Abstract This chapter is concerned with textual and sentiment analysis in agriculture commodities market using the natural language processing (NLP) methods. There are extensive research on textual and sentiment analysis in financial markets however, most of them are focusing on equity market and a minority on other commodities like energy commodities. Therefore, this chapter first reviews research works on textual and sentiment analysis in agriculture market in general. Then, presents textual analysis methods that can be carried out to study the effect of textual data and sentiment in agriculture market. Finally, it presents an example of implementing a topic modelling task and textual regression for forecasting realized volatility of corn returns. To the best of the author's knowledge, there is no study focusing on textual regression in agriculture market. Additionally, the studies conducting textual sentiment analysis are very limited. In this spirit, this study tries to fill this gap by introducing both well established and new textual and sentiment analysis methods to the agricultural researchers community. The limited experiment carried out with these methods in the present research testifies the superiority of the text-based models in explaining future movements of corn's volatility. More specifically, the results of one-month-ahead realized volatility regression indicates statistically significant superior performance of both direct textual regression and sentiment regression compared to traditional methods like HAR and ARIMA. In addition, as the most accurate method, textual regression's accuracy stands higher above that of the sentiment regression model.

N. Parvini (✉)
Department of Accounting and Finance, Kent Business School, University of Kent, Canterbury, UK
e-mail: n.parvini@kent.ac.uk

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_11

11.1 Introduction

Throughout history, understanding sentiment has played a significant role in economic modeling. Traditionally, sentiment was measured using surveys or indirect methods. However, in recent times, there's been a shift towards examining sentiment within text, audio, and visual data. With the digital transformation of communication, vast amounts of such data are available, offering valuable insights for financial and economic analysis. This shift has spurred a new area of finance and econometric research focused on translating qualitative sentiment data into measurable variables. These variables are then used to explore the connections between sentiment and other economic factors.

There are commonly three approaches for measuring sentiment. First, it can be collected by means of surveys. This method is expensive, time-consuming, hardly replicative, and not timely—suffering from release lags. Another way is to buy sentiment measures or indices from a third-party provider, like Thomson Reuters MarketPsych Indices. However, these indices are not domain-specific, rather they measure the overall sentiment of the economy/market. The third way is to estimate the sentiment from a collection of quantitative and qualitative data like market variables (e.g. trading volume and volatility), news, social media, blogs, and search engines [1]. The latter way is relatively cheap, timely, and flexible, meaning that one can collect information about any market or even market sub-categories over various time spans. If the underlying data for measuring sentiment is text, it is usually referred to as textual sentiment analysis.

One challenge of textual sentiment analysis is the huge amount of data available to process. The daily amount of information produced in the twenty-first century is beyond the processing power of human beings. Hopefully, the advent of artificial intelligence (AI) methods along with advancement in the computer's processing power has enabled us to overcome this issue. Nowadays, the development of AI-based machines has reached a zenith that experts raise alerts about the dangers of super-intelligent machines. One of the focus targets of AI is to provide tools that can help us extract information from big data. Textual analysis tools are very specific AI-based algorithms that can analyze a body of text, understand the context, and provide us with meaningful and easy-to-understand information from the text.

There are many attempts to categorize textual analysis according to the tasks carried out, the underlying methodology, etc. However, without going deep into the categorizing complexities, and focusing merely on the methods that are used in financial studies more frequently, we divide the textual analysis tasks into sentiment analysis, textual regression, topic analysis, and named entity recognition. Similarly, the methods can be roughly divided into word (token) frequency methods (a.k.a. count-based methods) and word embedding methods. Each of the above-mentioned tasks can be carried out using the two methods.

This chapter is interested in a branch of textual analysis methods called sentiment analysis. Sentiment analysis methods use algorithms to measure, estimate, or extract domain-specific sentiment from a body of text or from sentiment proxies. The

procedure usually starts by quantifying the sentiment in a text using tools and methods from the natural language processing (NLP) field. Then, use the sentiment score or index to study the relationship between textual sentiment and economic factors. However, thanks to advent of more sophisticated NLP models, a shortcut can be regressing financial and economics time series directly using text. Following [1] sentiment can be defined as:

> The disposition of an entity toward an entity, expressed via a certain medium.

Potential sources for textual data whose sentiment might affect financial and economic factors are media news, social media, blogs and microblogging platforms, and official reports. Text data is a source of timely information for estimating the future of the market using which market participants try to gain an edge over their competitors [2]. This source of novel information enables: traders to place their orders faster and make more profitable trades [3], investors to actively manage their portfolios [4], and policymakers to have a deeper insight into the market determinants. Moreover, it helps the market to work more efficiently and increases the speed of the price discovery process [3, 5].

Depending on the targeted market, the data sources can have different weights. For example, corporate reports might have the highest effect on corporate share price fluctuation, followed by media news. Social media sentiment is another major player in affecting financial markets because it is a significant source of investors' and traders' opinions regarding companies. However, commodities usually follow macro-level information and what we are calling "hard facts", rather than individuals' opinions on social media. In special cases like crude oil, though, information about big suppliers' and cartels' opinions or decisions that is usually reflected on social media can transmit shocks to the market. But other commodities like agriculture, tend to follow hard facts like information about weather, production, planting area, fertilizers, shipping, trading policy information and various macroeconomic variables.

The rest of the chapter is organized as follows: Section 11.2 provides a background on the research works on sentiment analysis in agriculture market; Sect. 11.3 explains how the text data should be prepared for the NLP methods explained in Sect. 11.4; Sect. 11.5 is an example of implementation of NLP models for real-world problems; and Sect. 11.6 concludes the chapter.

11.2 Backgrounds

The current sentiment analysis studies in agriculture can be generally divided into three main categories:

- Sentiment proxy analysis: where a (usually) non-text-related variable is proxied for market sentiment

- Event analysis: in which the effect of an event on the market is studied. The reaction of the market to the event is attributed to the sentiment carried by the event.
- News volume: where the count of news (social media posts) with regard to a market is proxied for the sentiment of the market.

In sentiment proxy analysis, researchers assume a time series like VIX, trading volume, etc. as a proxy for sentiment, and study the effect of this proxy on different aspects of the agriculture market. For example, Zheng [6] studied the predictive power of stock market investor sentiment on forecasting the returns of commodity future (agriculture commodity included). The author followed Baker and Wurgler [7] in constructing the sentiment index. Baker and Wurgler [7] proposed a measure that uses a combination of market turnover, closed-end fund discount, new equity issuances, number of IPOs, first-day return on IPOs, and difference in book-to-market ratios between dividend payers and dividend nonpayers as a proxy for investor sentiment. Zheng [6] documented a persistent negative relationship between commodity future returns and investor sentiment. The result also indicated a more strong relationship in periods with high conditional volatility. Hamadi et al. [8] explores the effect of macroeconomic announcements and news on the measurement of integration among the commodities. They used the consumer price index, federal funds rate, unemployment rate, and non-farm payroll announcements as the macroeconomic announcements, and Bloomberg economic surprise index (BESI) as the proxy for news surprises. The authors found evidence in favor of the strong influence of news surprises on the agriculture commodity variances. Moreover, Akyildirim et al. [9] studied the connectedness of the agriculture commodity market to the news-driven investor sentiment. They used Thomson Reuters Market Psych Indices (TRMI) as their proxy for market sentiment. TRMI is a sentiment index reported by Thomson Reuters and is estimated using NLP applications based on news, social media, and press releases. They indicated that around the first cycle of the COVID-19 pandemic in 2020, the pandemic had a significant impact on agriculture commodity returns. Additionally, they found that financial market uncertainty and economic policy are the determinants of the relationship between agriculture commodity returns and sentiment. Borgards and Czudaj [10] took the equidirectional trading of long and short agricultural commodity futures of long-short speculators as a proxy for their market sentiment to study if the long-short speculators can generate short-term investment returns in agricultural commodities. They showed evidence that commodity returns in the sentiment period were highly positive and had a significant difference from those of the neutral sentiment period. They also documented that a sentiment-based momentum strategy yields high returns. Balcilar et al. [11] utilized the Federal Reserve Bank of San Francisco's sentiment index, constructed following research by Shapiro et al. [12] as a news-based sentiment index to study the impact of the COVID-19 pandemic on major agriculture commodities. The results indicate that the news-based COVID-19 sentiment is the cause (by means of the Granger causality measure) of drastic volatility and price changes in agriculture commodities.

In event analysis, scholars study the effects of an event like USDA reports release on the agriculture market. Karali et al. [13] investigate the effects of USDA reports on the crop markets considering the private estimation of the report outcomes. They reveal that although the accuracy in private sector analysts' estimation of the reports has increased due to competition between private firms, they have not affected the surprise of the USDA report to the market. By exploring the impact of the change in the USDA news announcement regime in 2018, Adjemian and Irwin [14] revealed that the new release regime leads to an increase in trading volatility at announcement time, but vanishes quickly. In contrast, Indriawan et al. [15] examined the effect of the same release regime change on the agriculture market liquidity, volatility, information asymmetry, and high-frequency trading activity and found no significant difference in the aforementioned market qualities before and after the announcement regime change. Confirming the findings of Adjemian and Irwin [14], Bian et al. [16] found an increase in the volatility and the market microstructure friction, as well as a more efficient price discovery process resulting from the change in the news release regime. Cao and Robe [17] proxied the degree of analysts' optimism/pessimism about the upcoming announcement information for the commodity-specific sentiment to explore commodity price implied volatility (IVol) expectations around the USDA report release. They reveal that the way IVol responds to the reports depends on the sentiment before the report is released.

The third group of studies, i.e. news volume, takes the count of the news flow over a period of time as a proxy for sentiment and studies its impact on the agriculture market. Caporale et al. [18] investigated the effects of the macroeconomic news on the commodities returns (including agriculture commodities). They used the number of macroeconomic news (i.e. GDP unemployment, retail sales, and durable goods) articles published on Bloomberg news feed and identified spillover from news volume to agriculture commodity returns and a bigger spillover effect from news volume to agriculture price volatility. Phan and Zurbruegg [19] examined the time-to-maturity on the sensitivity of commodity future prices to news flows. They used the number of daily news headlines reported on Thomson Reuters News Analytics (TRNA) database and included agriculture future prices in their commodity price datasets. They found a U-shape relationship between unexpected news volume and the realized volatility of the commodities. Klomp [20] tried to explore the level of surprise in Russian retaliation sanctions for the agriculture commodity market. The author used the number of sanction-related news posted in major European newspapers containing sanction-related keywords as the proxy for retaliation measures by Russia. The results indicate that the boycott was partly anticipated by the investors. While still, it was the reason behind a drastic drop in the commodities returns. Sun et al. [21] investigated the effects of the trade policy uncertainty index (TPU) on agriculture commodity prices. Developed by Davis et al. [22], TPU is the monthly count of news articles that include keywords related to economic, trade policy, and uncertainty categories. They uncovered both positive and negative impacts of TPU on the agriculture commodity prices, which is due to the effect of TPU on supply and demand in the agriculture commodity market.

Moreover, they showed a bidirectional relationship between TPU and the prices where in some periods commodity prices affect the fluctuations in TPU.

11.3 Data Preparation

Data preparation is a task-specific, data-specific, and language-model-specific part of NLP and there are no strict rules to follow. However, there are general steps that can be applied to many types of data in various tasks, with different language models. For example, data cleansing and Text normalization are part of any text preparation in finance.

11.3.1 Data Cleansing

One challenge of working with textual data is that they tend to be very noisy and contaminated by huge amounts of redundant data and information. Data cleansing is the process of removing any redundant pieces of text before normalizing it. It helps the model to understand the text better, generalize easier and be trained faster. Therefore, the first and most time-consuming part of any textual analysis project is data cleansing. The problem lies within the text extraction task and has various methods to tackle with. One of the most widely used and efficient, yet complex, method is regular expressions (also called regex). This method searches the text to find pre-designed patterns. These patterns are designed by the user and usually target the most frequent part of the text, i.e. the noise. Email addresses, URLs, headers and footers are examples of noise that can be captured by the regex. Interestingly, despite the advancement of textual analysis tools, tables still fall into the category of noise information for the NLP algorithms. The way language models understand the tables and finding a solution for extracting information from tables using NLP models is currently one of the hot topics among NLP scholars.

Removing email addresses, URLs, authors, usernames, dates, credentials, etc. are usually part of text cleansing for finance and economics applications. However, it ultimately depends on what data is used, for what purpose and using what language models. For example, using news articles for financial market analysis using LLMs requires, in addition to abovementioned steps, removing tables. Because today's state-of-the-art NLP models have no way of interpreting tables yet. A counter intuitive example is the role of exclamation marks, question marks, and emojis that are more common in texts like social media posts. While researchers used to eliminate them from the text in the past, nowadays the experts suggest keeping this data as they might represent some information about the sentiment of the text.

11.3.2 Text Normalization

Text normalization is the task of preparing the text for a language model. According to Jurafsky and Martin [23] it comprises of at least three steps: Tokenization, Normalizing word formats, Segmenting sentences.

11.3.2.1 Tokenization

Tokenization is the task of dividing a body of text into separate words or subwords that are generally called **tokens**. The sentence "*The low volatility may remain as one of the features of the following move*" can be simply divided into 14 words. Transforming the example above, only the word "*volatility*" is divided into subwords: "*vol*", "*ati*", "*lity*".[1]

11.3.2.2 Normalizing Words

Normalizing words is the process that deals with the inconsistencies in different versions of the same word. It can start with translating (or removing) parts of text whose language is different form the rest of the corpus and be continued with typo correction. Then, case folding, stop word removing, lemmatization or stemming are the steps that might or might not be necessary.

Case folding is the process of changing upper case letters into lower case ones. While it can be helpful to generalize the model in some tasks, it is not generally done in tasks like sentiment analysis or text regression because it eliminates useful information in similar words with different cases like *US* and *us.*

While there is no strict definition on **Stop words**, they are usually the most frequent words in a text document carrying little value regarding the contextual meaning of the text. For some examples, "a", "an", "the", "be", "by", "that", "will", etc. are usually stop words [25]. There are "stop lists" for different approaches. The length of these stop lists may vary between 7 to 300 terms. Most of the NLP algorithms (except transformer-based LLMs) either ignore stop words or have no idea how to process them. Therefore, if the LM is not and LLM, stop words are considered noise and should be cleaned from the text.

Lemmatization is to find the root of a word. For example, among the words "go", "goes", "went", and "going", the root for all of them is "go". Lemmatization will change the sentence "*gains in soybean, corn and wheat futures were capped by the disappointing U.S. weekly export inspections*" into "*gain in soybean, corn and wheat future be cap by the disappointing U.S. weekly export inspection*". The direct result of lemmatization is a reduction in the length of the vocabulary $|V|$, and simpler

[1] The tokenization is carried out using the tokenizer of pretrained BERT package that follows WordPiece algorithm [24].

NLP algorithms will become more efficient. **Stemming** on the other hand, is similar to lemmatization, but more cruder, chopping off the final affixes of words. There are different stemmer schemes, among which Porter Stemmer [26] is one of the most widely used. Using the Porter Stemmer, the above phrase will become: "*gain in soybean, corn and wheat futur were cap by the disappoint U.S. weekli export inspect*".

11.3.2.3 Sentence Segmentation

Sentence segmentation (a.k.a. sentence tokenization) is the process of splitting a text into individual sentences. It is usually done with regard to the position of punctuations like exclamation marks, question marks, and periods withing the text. However, it can sometimes be difficult to distinguish a period that is the sentence boundary from the one that is the mark of abbreviations like *Mrs.* or *corp.* In practice, sentence segmentation methods first decide if the period character is part of a word or sentence boundary. This can be achieved by rule-based systems or machine learning. A dictionary of abbreviations can help facilitating the task.

Nowadays, with the advent of transformer-based LLMs and pre-trained LLMs, some of the normalization steps are skipped, while text cleansing is still widely practiced. Most of the pre-trained LLMs are designed with tokenizer modules that circumvent stop word removal and stemming and lemmatization by transforming the words into sub words with regards to a pre-defined dictionary of sub words.

11.4 Methods

11.4.1 Lexicon-Based Methods

During the recent past years, language processing methods have witnessed a drastic improvement both in their performance and their applications. Generally speaking models that assign probabilities to sequences of words are called language models or LMs. The most simple LMs are called **Bag-of-words (BoW)** methods because they regard a document a bunch of independent words (tokens) without considering their order. This methods assign a probability to each word based on the frequency of the word in the corpus. Based on this simple concept, the more complex models like **n-grams** and **Bayesian models** are developed.

Using BoW for text classification is simple. The user needs a dictionary of words categorized in different classes. These dictionaries are called **lexicon**. For example, a lexicon containing words that are tagged with positive, negative, or neutral sentiments. Then, in a sentiment classification task, the ratio of the positive, negative, and neutral words relative to the document length is considered positive, negative, and neutral sentiment score, respectively, for the document. Let S_{Pos}, S_{Neg},

and S_{Neu} be the positive, negative, and neutral sentiment scores, respectively, of the document d with length $|d|$. Then, the sentiment scores are:

$$S_{Pos} = \frac{Number\ of\ positive\ words}{|d|}, \tag{11.1}$$

$$S_{Neg} = \frac{Number\ of\ negative\ words}{|\ d\ |}, \tag{11.2}$$

$$S_{Neu} = \frac{Number\ of\ neutral\ words}{|\ d\ |}. \tag{11.3}$$

The most widely used general purpose lexicons are General Inquirer [27], LIWC [28], the opinion lexicon of Hu and Liu (2004a) and the MPQA Subjectivity Lexicon [29]. However, the efficacy of general purpose lexicon for domain-specific task like financial and economics textual analysis is questionable. In response to this shortcoming, several finance and economics domain-specific lexicons are developed, the most prominent among them are Henry lexicon [30] and the Loughran and McDonald lexicon [31].

11.4.2 Linear Classifiers

Developed on the word independence assumption of BoW, linear classifier **naïve Bayes** assigns class $\hat{c}$ to document d from a set of pre-defined classes C following

$$\hat{c}_d = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P\left(w_i | c\right), \tag{11.4}$$

where i and n are the position of word w_i in the document and length of the document, respectively. The model is first trained using a manually labeled set of documents (training set), also called **Gold labels** or **Gold standard**. The probabilities $P(c)$ and $P(w_i|\ c)$ that are determined during training phase, are applied to the test set to estimate the class of each document using Eq. (11.4).

11.4.3 Term and Document-Based Matrices

Another way of modeling a language is to utilize **frequency matrices** also called **term-document matrices**. In these matrices, each column represents a document, and the rows are the words within the corpus (the dataset of all documents). The cells are filled with the frequency of the word t in each document d ($tf_{t,\,d}$). If

the corpus is large enough, the length of the rows of the underlying frequency matrix is approximately equal to the vocabulary size $|V|$ of the language. This very sparce matrix is a collection of column vectors for each document. In other words, each document is represented by a vector in $|V|$-dimensional space. Note that the dimensionality of vectors, i.e. the vocabulary size, is often between 10,000 and 50,000.[2]

The term-document matrix is not as informative because of presence of highly frequent words like *the, a, it, they*, etc. A better measure would be to adjust the term-document matrix with the frequency of occurrence of each word in the documents. First, let us redefine term-frequency of a word t in document d as:

$$tf_{t,d} = \log_{10}\left(\text{count}\left(t, d\right) + 1\right). \tag{11.5}$$

Then, we can argue that the words that appear in less documents can be much more informative regarding distinguishing between documents. Therefore, we can define a scaling value for term-frequency that puts more weight on less occurring terms and less weight on more occurring terms, and call it inverse document frequency (idf):

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right), \tag{11.6}$$

where N is the total number of documents in the corpus D, and df_t is the number of documents containing word t. Now, the scaling measure, called term frequency inverse document frequency or **tf-idf**, for a word t in document d would be $w_{t,d}$ and is defined as:

$$w_{t,d} = tf_{t,d} \times idf_t. \tag{11.7}$$

Presenting an example form [23] Table 11.1 shows the tf-idf weights of four words *battle*, *good*, *fool*, and *wit* in four Shakespeare plays. As presented in the table, because the word *good* is a ubiquitous word in the four plays, the vector for word *good* is not discriminative. While the word *battle* is quite discriminative with high values in dimensions *Julius Ceasar* and *Henry V*, plays with battle components, and low values in dimensions *As You Like it* and *Twelfth Night*.

As an example of calculating tf-idf values, the term frequency of *wit* in *As You Like It* is 20 and its document frequency is 34, meaning it is appeared in 34 documents out of 37 sampled documents. Therefore, $tf_{wit,\ \textit{As You Like It}} = \log_{10}20 + 1 = 1.322$, and $idf_{wit} = \log_{10}\left(\frac{37}{34}\right) = 0.037$.

[2] Sorting the words according to their frequency, keeping the most frequent words after 50,000 does not help the performance of the model.

Table 11.1 The tf-idf values for four sample words in a sample of four Shakespeare plays

	As You Like It	Twelfth Night	Julius Caesar	Henry V
Battle	0.074	0	0.22	0.28
Good	0	0	0	0
Fool	0.019	0.021	0.0036	0.0083
Wit	0.049	0.044	0.018	0.022

Because using Eq. (11.7), $w_{wit,\ As\ You\ Like\ It} = 1.322 \times 0.037 = 0.049$, the tf-idf score of *wit* and *As You Like It* in Table 11.1 is 0.049.

The tf-idf matrix can be used directly as the input for classification and regression methods (in finance and economics), or to model the language using a probabilistic language modelling. Please refer to [23] for more information on probabilistic language models.

When it comes to quantifying the semantic and syntactic role of an individual word in document (corpus), another matrix appears to be more useful. Usually called **term-term matrix,**[3] it is a $|V| \times |V|$ matrix containing the information regarding the co-occurrence of words in a document. The idea is based on the fact that the words that appear in the same document tend to have similar meanings. Therefore, a word can be defined by a simple function of counts of nearby words. In a term-term matrix the cells record the co-occurrence of the words in rows (target word) and columns (context word) in some context in a training corpus.

Similar to term-document matrix, term-term matrix can be improved by adjusting the uninformative raw frequencies. The method for adjusting this matrix is called **positive pointwise mutual information (PPMI)**. PPMI for target word w_i and context word $c_j$is calculated as:

$$PPMI_{ij} = \max\left(\log_2 \frac{p_{ij}}{p_{i*}p_{*j}}, 0\right), \tag{11.8}$$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}, \tag{11.9}$$

$$p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}, \tag{11.10}$$

$$p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}, \tag{11.11}$$

[3] word-word matrix and term-context matrix are two other names for term-term matrix.

Table 11.2 co-occurrence count matrix for four words in five context in the Wikipedia corpus

	Computer	Data	Result	Pie	Sugar	Count(w)
Cherry	2	8	9	442	25	486
Strawberry	0	0	1	60	25	80
Digital	1670	1683	85	5	4	3447
Information	3325	3982	378	5	13	7703
Count(context)	4997	5673	473	512	61	11,716

Table 11.3 The PPMI matrix of the word-context pairs reported in Table 11.2

	Computer	Data	Result	Pie	Sugar
Cherry	0	0	0	4.83	3.30
Strawberry	0	0	0	4.10	5.51
Digital	0.18	0.01	0	0	0
Information	0.02	0.09	0.28	0	0

where f_{ij} gives the number of co-occurrence of target word w_i and context word c_j. The following example from the [23] clarifies the procedure. Suppose that we have co-occurrence count table like Table 11.2.

Assuming that the information in Table 11.2 encompasses all the relevant word contexts, the $PPMI(\text{information}, \text{data})$ is calculated as follows:

$$P\,(w = \text{information}, c = \text{data}) = \frac{3982}{11716} = 0.3399,$$

$$P\,(w = \text{information}) = \frac{7703}{11716} = 0.6575,$$

$$P\,(c = \text{data}) = \frac{5673}{11716} = 0.4842,$$

$$PPMI\,(\text{information}, \text{data}) = \log_2 \frac{0.3399}{0.6575 \times 0.4842} = 0.0944.$$

The *PPMI* matrix of the association between word-context pairs of the Table 11.2 is presented in Table 11.3. From the results in Table 11.3 it can be seen that the vectors for the words *cherry* and *strawberry*, [0, 0, 0, 4.83, 3.30] and [0, 0, 0, 4.10, 5.51], are very similar, and the vectors for *digital* and *information* are very similar too. While the vector for *cherry* and *information* are very dissimilar. The similarity between two vectors are measured by cosine distance between the two vectors that is discussed in Sect. 11.4.7.

Here, we have used only the immediate neighbor of the target words as the context for the sake of simplicity. However, the equations can be easily generalized to consider L words around the target word as the context. Due to growing

computation cost as L increases, it is common in practice to consider L in the range of 2–10.

It is a ground breaking achievement to model qualitative information like text as multidimensional vectors. Because now, these word-representing vectors can be used as the inputs to any quantitative model and draw quantitative inferences from the text. In other words, these vectors are the raw data for classification and regression methods that can be used for extracting financial and economic conclusions.

11.4.4 Word2vec

The methods presented above have many shortcomings. One problem is that the vectors they produce are very sparce, containing many zero dimensions. In addition, the term-term matrix cannot decode the syntactic and semantic properties of words very accurately. It is limited to the surrounding words. For some other reasons that are not completely clear yet, shorter and more dense vectors tend to result in models with higher performance. These dense vectors usually have dimensions in the range of 50 to 1000, instead of the whole vocabulary size $|V|$ in term-term matrix. Despite the term-term matrix the dimensions of dense vectors don't have a clear interpretations. It means that, despite word vectors of term-term matrix, in dense word vectors dimension n of the vector of word w cannot be attributed to the relationship between the word w and context word c at dimension n.

Nowadays, there are various methods for estimating dense vector representations, skip-gram based methods, also known as word2vec, recurrent neural networks, and transformer-based algorithms are the most prominent ones.

Word2vec methods introduced by Mikolov et al. [32] are fast and efficient methods to assign static dense vectors to words using **skip-gram with negative sampling method (SGNS).** It is a so-called self-supervised method that looking at a document, it treats target words and their neighboring context words as positive (true) examples. While it randomly samples other context words in the lexicon to construct a negative (false) examples for the same target word, hence negative sampling in SGNS. Then, it uses a logistic regression to train a classifier to distinguish those two cases, and uses the trained weights as the vector representations called **embeddings**. In other words, the intuition behind the word2vec is to train a classifier to answer "is target word w likely to show up in the vicinity of context c?". When the classifier managed to learn the answer optimally, with lowest possible error, then it ignore the classifier's answers and takes the classifier's weights as the word embeddings. These embeddings represent the relationship of the target word w with all the context words c that have appeared in the vicinity of w in the whole training corpus, hence the semantic embedding of w.

The word2vec classifier tries to minimize the following loss function:

$$L_{CE} = -\left[\log P\left(pos|w, c_{pos}\right) + \sum_{i=1}^{k} \log P\left(neg|w, c_{neg_i}\right)\right], \tag{11.12}$$

where $P(pos|\, w, c_{pos})$ is the probability that c_{pos} is a real context (neighboring) word for target word w, $P\left(neg|w, c_{neg_i}\right)$ are the probabilities that c_{neg_i} are not the real context words for w, i.e. they are the randomly sampled words, for all the noise words $c_{neg_1}, \ldots, c_{neg_k}$.

During training, Word 2vec learns two sets of weights, $\boldsymbol{w_i}$ when the i is a target word and $\boldsymbol{c_i}$ when i is a context word. It is common to add these two sets of vectors to produce the final embeddings. Alternatively, $\boldsymbol{c_i}$ can be ignored and i can be represented with $\boldsymbol{w_i}$ only.

Similar to the count based methods like term-term matrix, the window L around the target word can be expanded to incorporate more information about a target word.

As an extension to word2vec, **fasttext** [33] introduced to address the unknown word problem of word2vec. Unknown words are the words that are not present in the training dataset vocabulary. Therefore, the word2vec algorithm ignores any new word in the test dataset if it is not in the training set. fasttext, on the other hand, utilizes subword model that breaks down the words into subwords, assigns an embedding to each subword, and adds the embeddings to construct the final static embedding for a word. This way, any new word in the test set can be split into subwords familiar to the trained algorithm.

GloVe [34] is another widely used static embedding representation. It is short for Global Vectors and incorporates both the linear structure used by methods like word2vec and the probabilities of count-based co-occurrence methods like PPMI to capture the so-called global statistics of the corpus.

All of the above methods have the context window size parameter L that needs to be tuned. The choice of the L depends on the application of the embeddings and it is usually between 1 and 10 words on each side of the target word, meaning a total of 2–20 context words per target. The magnitude of L slightly changes the embeddings with shorter windows capturing more syntactic properties of the target, while larger windows result in embeddings closer to the topic of the document.

In practice, instead of training a task-specific model for every application, it is more common to use the embeddings that are pretrained on a huge general corpus and are available online. The benefit of using these pretrained embeddings is to save the cost of the training an NLP model, as well as having a good generalization. However, it is important to note that the NLP models performance can deteriorate drastically if the training corpus and the test corpus contain very different texts.

11.4.5 Neural Networks and Deep Learning

While frequency-based LMs and linear language models like naïve Bayes struggle to incorporate the dependence between the words in a document and the role of the previous words on its contextual decoding, more modern non-linear LMs can handle more distant histories (previous context words) and generalize better. Thus, they produce more accurate representations, however, at the cost of higher complexity, training time, and lower interpretability.

The most widely used deep neural network models for modeling language, prior to introducing Transformers, was RNN-based[4] model. Like the previous language models, RNNs try to approximate the probability of a word, given the prior context $P(w_t | w_{1:t-1})$. In their most basic form, RNNs take an initial embeddings (usually random vectors) for input words and try to optimize the embeddings during the training. They have a recursive mechanism, hence the name recurrent, that feeds the previous step's hidden state containing information about the previous word (token) h_{t-1} as input to the current hidden state h_t. Therefore, passing information about the previous words to the next words in a sequence, they can capture the relationship between the words in a sequence. For example, processing "*it*" in the sentence "*The equity market experienced a significant downturn yesterday, while it has managed to recover some of those losses today.*", an NLP model needs to preserve all the information from the beginning of the sequence to relate "*it*" to the "*the equity market*".

More specifically, RNN language models [35] process one word at a time trying to predict the next word using the information from the current word and the previous words of an input sequence in a self-supervised manner. This way, RNNs overcome the problem of fixed context length of the previous models.

In essence, an RNN language model tries to minimize the following loss function:

$$L_{CE}\left(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t\right) = -\log \hat{\boldsymbol{y}}_t \left[w_{t+1}\right], \tag{11.13}$$

where $\hat{\boldsymbol{y}}_t \left[w_{t+1}\right]$ is the probability distribution over the possible next word. It is computed by applying softmax function to the information form the last hidden state. Let $\boldsymbol{E}$ be the matrix of weights from the last hidden state to the output layer, then $\hat{\boldsymbol{y}}_t$ can be calculated as:

$$\hat{\boldsymbol{y}}_t = \textbf{softmax}\left(\boldsymbol{E}^T \boldsymbol{h}_t\right). \tag{11.14}$$

An advantage of the RNNs is that they can be trained directly as the classifiers and regressors for high-end tasks in finance and economics. While, they can also

[4] The set of RNNs represent the all the NN variants that incorporate a recursive mechanism, including simple\vanilla RNN, LSTM, GRU, etc.

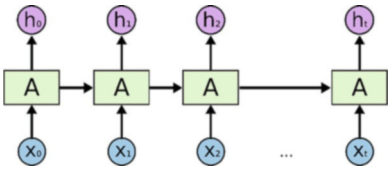


Fig. 11.1 A schematic of an unrolled RNN [36]

be trained on a large corpus of text and the trained embeddings be extracted as the word embeddings for the following financial and economics applications. Figure 11.1 shows a schematic of an unrolled RNN.

11.4.5.1 Transformers

In simple words, Transformers are NN-based encoders-decoders. These are currently the state-of-the-art algorithms [37]. The most important features of Transformers are self-attention mechanism and positional encoding. Before introducing self-attention mechanism, the attention mechanism should be explained. Attention mechanism introduced before the advent of Transformers as a way to mitigate the information loss in RNNs. It is a way for the network to get the information form all the previous hidden states and not only the final hidden state in an RNN. In this mechanism, instead of providing the previous hidden state to the final processing unit, a context vector $\boldsymbol{c}$ is provided, which is a function of all the previous hidden states, i.e. $\boldsymbol{c} = f(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n)$. This function f is usually a weighted average and the weights are the attentions that the final context vector $\boldsymbol{c}$ pays to each hidden state $\boldsymbol{h}_n$. This is the way for the algorithm to have an idea about the importance of different hidden states (containing information about previous words) in the sequence. Note that in RNNs, the context vector $\boldsymbol{c}$ is calculated at the end of the encoding (when the algorithm processed the final word in the sequence) and is passed to the next module, i.e. decoder, for using the encoded information.

Exploiting the attention idea, the Transformers algorithm developed self-attention mechanism in which it assigns weights to the previous words of the sequence when processing another word in the same sequence. Self-attention helps the algorithm relate each word in the sequence to the other words in the same sequence.

We discussed that the RNNs processing words of sequence in series, meaning that they process a word at a time and pass the information about the word to the next processing step. This series processing preserves the order of the words in the sequence, but the disadvantage is that the algorithm needs the result of the previous step to process the current step. Transformers address this issue by introducing positional encoding. In this mechanism, each word is assigned a vector relative to its position in the sequence. This vector is added to the embedding of the word before feeding it to the Transformer block. Using positional encoding the algorithm knows the position of word in the sequence and can process each word independently in a faster parallel processing fashion.

As with the overall Transformers, they are made up of stacks of transformer block to map a sequence of input vectors $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ to a sequence of output vectors of the same length $[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]$. Each block contains layers of self-attention, feed forward NNs, and simple linear layers. The process starts by assigning a position to each word in the sequence. Then, the initial embeddings of each word is added to the positional vectors. These positioned embeddings then are passed through the self-attention layer. The self-attention layer calculates the weights that represent the relationship between a words and its previous words in the sequence. In a vanilla Transformer, the self-attention does not have access to information about the words beyond the currently processing word. The output vectors of the self-attention, i.e. positional embeddings and the self-attention vectors, are fed to several normalization and feed forward NN layers. Finally, the final layer produces an embedding vector for each input word. These embeddings are similar to the ones produced by word2vec and RNN, a high-dimensional vectors, but more rich in information about the word. These vectors can be extracted for the subsequent tasks or fed into the decoder block of the Transformer.

Another revolutionary advantage of Transformers is that, despite the static embedding produced by word2vec and RNNs, they can output a context related dynamic embeddings called **contextual embeddings**. Consider the word "*bank*" in two sentences "*I deposit my money in the bank*" and "*I walked along the river bank*". A static embedding representation for the word "*bank*" have to incorporate both meanings of the word at the same time because it is static and once produced cannot adapt to the context of the sentence. While a dynamic contextual embeddings will change depending on the context, i.e. the words around it. While this was an example of two different meaning for the same word, some other words might convey even more meanings from one context to the other. For example, the word "*book*" can have up to six different meanings in different contexts.

Transformers achieve this feat by using **multi-attention-head** mechanism in which not one self-attention vector, but multiple vectors are trained for one word simultaneously, each representing one aspect of the word depending on its context. Therefore, a Transformer algorithm produces two different embeddings for the same word "*bank*" using the two sentences.

Despite the static embeddings where a model is trained on a corpus then the learned weights are shared as the embedding vectors of the words, producing contextual embeddings requires the user to have the trained model in possession and run it for any target sentence. Therefore, in Transformer-based LMs it is the trained model, i.e. model structure and trained weight, that is shared. The end user can download a trained model and feed it with sentences to extract the embedding for the words. An example of an encoder-decoder Transformer block is presented in Fig. 11.2.

Although initially introduced as an encoder-decoder architecture, Transformers can be found in encoder only models, like BERT, and decoder only architecture, like generative pretrained transformer (GPT).

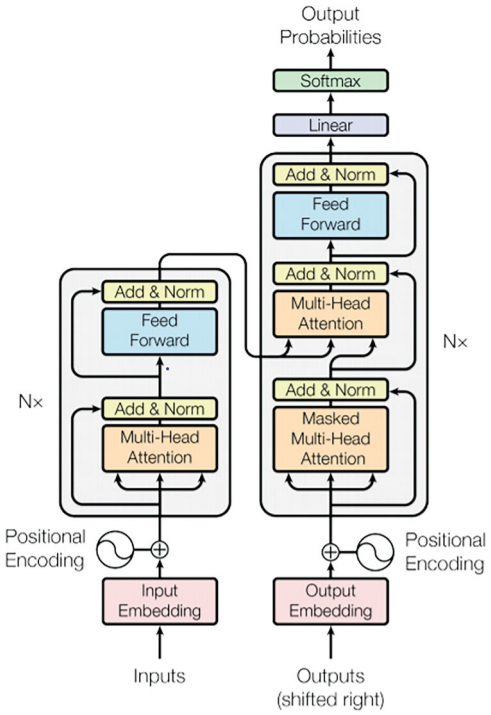


Fig. 11.2 An encoder-decoder Transformer block. [37]

11.4.5.2 BERT

The model introduced in the previous section was vanilla Transformer or so-called causal or left-to-right Transformer models. In these models, the Transformer has only access to the previous words when it process the current word. A new generation of Transformers, called bidirectional Transformer encoders, provide the access to the entire sequence, both left and right context, while learning about the current word. This helps the model to learn the role of the word in the sentence more deeply with regard to what comes before and after that. They also use a training technique called **masked language modelling (MLM)** in which, instead of predicting the next word, a word in the sequence is masked randomly, i.e. its value is unknown to the algorithm. The model tries to predict the masked word using the entire context of the sequence. In other words, instead of "guess-the-next-word" task, the algorithm tries to learn "fill-in-the-blank" task.

BERT [24], short for Bidirectional Encoder Representations from Transformers, is one of the most widely used bidirectional Transformer that utilizes MLM. In BERT, 15% of the inputs are sampled for learning. Unsing the learning sample, MLM is performed in three ways:

- 80% are masked, i.e. replaces with the unique token [MASK],
- 10% are replaced with random tokens from the vocabulary,
- 10% are left unchanged.

The objective of MLM training is to guess these original token for the masked tokens.

Masked-based learning focuses on predicting words using surrounding contexts to create effective word-level representations. However, an essential set of applications involves assessing relationships between sentence pairs. These tasks include determining paraphrase similarity, entailment, or discourse coherence.

To address these application needs, BERT introduces a second learning objective known as **Next Sentence Prediction (NSP)**. NSP involves presenting pairs of sentences to the model and tasking it with discerning whether each pair consists of consecutive sentences from the training data or unrelated sentences. BERT's training involved 50% actual adjacent sentence pairs and 50% where the second sentence was randomly selected from elsewhere in the corpus. The NSP loss measures the model's ability to distinguish true pairs from random ones.

To facilitate NSP training, BERT incorporates two new tokens into the input representation (which also prove useful for fine-tuning). These tokens, [CLS] prepended to the sentence pair and [SEP] inserted between and after the sentences, aid in creating embeddings for the model to differentiate between the input sentences, enhancing its comprehension of the relationships between them. The [CLS] token is often called the sentence embedding because it represents the whole information in a sentence. It means that, besides an embedding vector for every word in the sequence, BERT also produces an embedding vector for representing the whole sentence. The [CLS] vector is usually the only vector that is fed into a classification and regression algorithm that comes after the NLP encoder.

It is also worth noting that BERT is only an encoder that maps the input tokens to a vector representation. Its output can be used for subsequent tasks such as classification, regression, named entity recognition, translation, text generation, etc.

As discussed in Sect. 11.4.5.1, Transformer-based language models can be pretrained and shared to be used by the end users. BERT, also, has versions that are trained for different tasks. The most widely used version is the general purpose BERT that is pretrained on a huge corpus of text. The pretrained model then can be downloaded by the end user and directly used for the subsequent tasks,,can be further trained using a domain-specific corpus to get more familiar with the texts in that domain, or can be fine-tuned by a set of labelled task-specific text documents to become task-specific BERT.

Fine-tuning involves taking a pretrained LLM, typically using added neural net classifiers or regressors that utilize the top layer's embeddings, and further training the model for specific tasks such as sentiment classification, named entity recognition, or regression. The idea is that the initial pretraining phase teaches the model a language understanding that encompasses nuanced representations of word meanings, facilitating easier adaptation ('fine-tuning') to new language tasks.

This pretrain-finetune approach aligns with transfer learning in machine learning, where knowledge gained from one task or domain gets applied (transferred) to solve a different task. In pretraining, the LLM's weights optimize to capture intricate word representations, while during fine-tuning, these weights undergo minor adjustments to better suit the requirements of the final task. Essentially, pretraining builds a

foundation of understanding, and fine-tuning tailors this knowledge for specific applications.

The original BERT model comprised the following components:

- A subword vocabulary containing 30,000 tokens, created using the WordPiece algorithm [24].
- Hidden layers with a size of 768 (this is the embedding vector length).
- 12 layers of transformer blocks, each containing 12 multihead attention layers.

This configuration resulted in a model with more than 100 million parameters. BERT and its related models rely on subword tokens generated through algorithms like WordPiece, instead of processing individual words. This means that every input sentence undergoes tokenization, and subsequent processing occurs using these subword tokens rather than full words.

11.4.6 Text Generation

Using a language model to create text is among the most significant applications of neural language models in NLP. Text generation, alongside image and code generation, forms a fresh domain in AI often referred to as **generative AI**.

A decoder (a.k.a. generator) is similar to a classifier that, once provided with the information about the current state (word), it calculates the probability of all the candidates for the next state (word). A trained decoder knows the joint probability distribution over the whole vocabulary. Therefore, if it is provided with the current and the previous word(s), it can guess the next word.

A trained decoder usually follows autoregressive text generation approach to generate a novel text. The process starts by feeding the starting token to the decoder, which is a token defined as the start of every sentence during the training, e.g. <s>. The decoder generates the next word by calculating the probability over the entire vocabulary conditioned on the previous entry and returns the word with the highest probability. Then, it receives the previous generated word(s) and generates the next word. The process continues until a predefined length is reached or the ending token is generated. It is a token at the end of every sentence as a signal that the sentence has ended, e.g. </s>. The result is a sequence of words or a sentence.

In a more complex model like GPTs, the starting token can be a sentence form the user. For example, user can ask a question and the generator generates the answer. The procedure is similar to abovementioned process, i.e. the algorithm finds a sequence of tokens with the highest probability conditioned on the previous tokens.

11.4.7 Evaluation Metrics

The best evaluation metric for NLP models is to try them on tasks, i.e. to see if an NLP model improves performance, competing with rival models. In self-supervised methods like RNN and Transformers, the loss of predicting the next token can be used as the measure of performance of model during its training phase. However, the out-of-sample or test set evaluation is still a challenge.

Nevertheless, it is still important to have an idea about the models intrinsic performance. The most widely used intrinsic metric is similarity measure. It is comparing the similarity scores of the embeddings produced by an algorithm with that of assigned by humans, a.k.a. gold standards. For example, WordSim-353 [38] is one of the most widely used gold standards of 353 word pairs with human labeled similarity scores on the scale of 0 to 10. In WordSim-353, (plane, car) pair, for example, have a similarity score of 0.577. SimLex-999 [39], TOEFL dataset [40], Stanford Contextual Word Similarity (SCWS) dataset [41], the Word-in-Context (WiC) dataset [42], and The semantic textual similarity task [43] are other useful gold standard datasets.

The most common metric to measure the similarity between two multidimensional vector representations is the cosine of the angle between the two vectors. It is called **Cosine similarity measure** and is computed by normalizing the dot product between the two vectors. Thus, the cosine similarity between two vectors w and v with lengths $|w| = \ \mid v|$ and dimensions N is:

$$cosine\,(v, w) = \frac{v.w}{|v| \mid w \mid} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}, \tag{11.15}$$

The cosine similarity metric ranges from $+1$ for the perfectly similar vectors pointing to the same direction, to -1 for the vectors the point to the opposite directions, with cosine $= 0$ showing the orthogonal vectors.

In finance and econometrics, researchers usually use classification and regression performance as the measure of NLP model performance. In classification tasks, sentences labeled with sentiment are used as the gold label to train a classifier.

11.4.8 Visualization

The tools and techniques for visualizing the information in text data is very limited, partly because of high dimensionality of the embedding vectors. Before extracting the embedding vectors, however, there are ways of representing some information in the text. Word clouds are among the most widely used methods of illustrating the frequency of words in a body of text. Shown in Fig. 11.3, a word cloud depicts the words separately on a canvas in which the size of the words are associated with

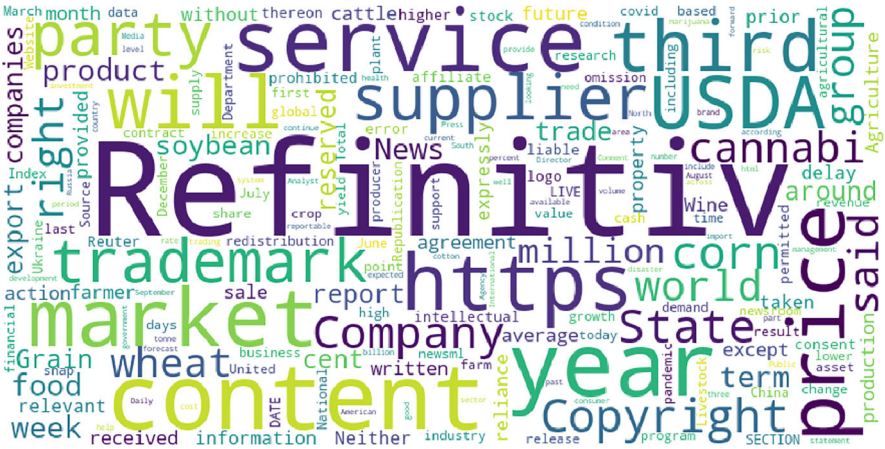


Fig. 11.3 Word cloud of the 2022 media news dataset before applying cleansing procedure

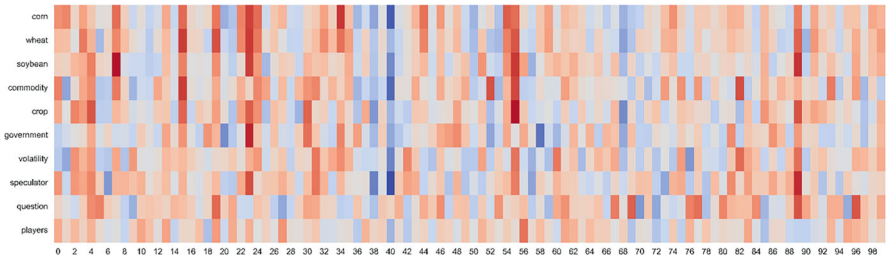


Fig. 11.4 The first 100 dimension (out of 768) of 10 embeddings. The embeddings are extracted using the pretrained model of base version of BERT algorithm, accessible using "bert-base-uncased" from Huggingface.com

their frequency in the corpus. In Fig. 11.3, for example, the most frequent words are "*Refinitive*", "*consent*", "*service*", and "*https*" as they are appeared with the biggest font sizes, while words like "*rate*", "*import*", and "*cost*" are examples of less frequent words. It is noteworthy that the list of words sorted from highest to lowest frequent words might be truncated at some threshold below which cannot be fitted in a word cloud illustration.

Visualizing a text using the word embedding vectors or sentence embedding is more challenging. One technique to represent an embeddings is to use color coding vectors. An example of color coded vectors is depicted in Fig. 11.4. In the figure, the words "*corn*", "*wheat*", "*soybean*", "*commodity*", and "*crop*" show similar patterns, while words like "*question*" and "*players*", from Shakespear quotes, are completely different. Note that this is only the first 100 dimensions out of 768 dimensions of the embeddings. The cosine similarity of the same words is presented in Fig. 11.5. The similarity of the words is inline with the patterns in Fig. 11.4.

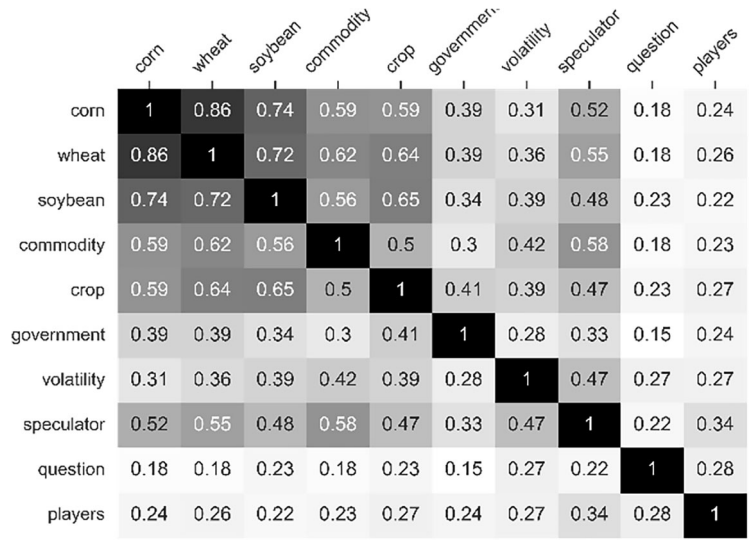


Fig. 11.5 Cosine similarity of a sample set of words in an agriculture-related media news text. The embeddings are extracted using the pretrained model of base version of BERT algorithm, accessible using "bert-base-uncased" from Huggingface.com

Color coding the embeddings is a useful method, yet comes with limitations. For example, the number of dimensions that can be fit into an illustration and the number of embeddings that can be shown are very limited. Therefore, it can be useful to compare only a sample of embeddings belonging to most interested words\sentences on a color coded diagram. Another technique that helps having an idea about related and opposite words is clustering. As a very simple method of clustering, dendrogram helps categorizing words based on their distance to the other words. Thus, more related words appear on the same branch, while words with different meanings or belonging to different domains tend to be on a separate branch. Figure 11.6 shows a dendrogram in which each major branch is color coded. In the dendrogram, words like "*wheat*", "*bread*", "*flour*", and "*dough*" are in a same branch. While "*news*", "*quote*", and "*correspondent*" appear to be more closely related.

11.4.9 Text as a Time Series

One of the main reasons we can study the effect of textual data on time series is the fact that the textual data are usually time series. This is particularly true about the textual data that are often used in financial and economic studies, i.e. news text, social media text, corporate reports, etc. These texts are issued regularly through time and as it is discussed, the texts can be represented as a multidimensional vector. Therefore, text can be treated as a multidimension time series because at

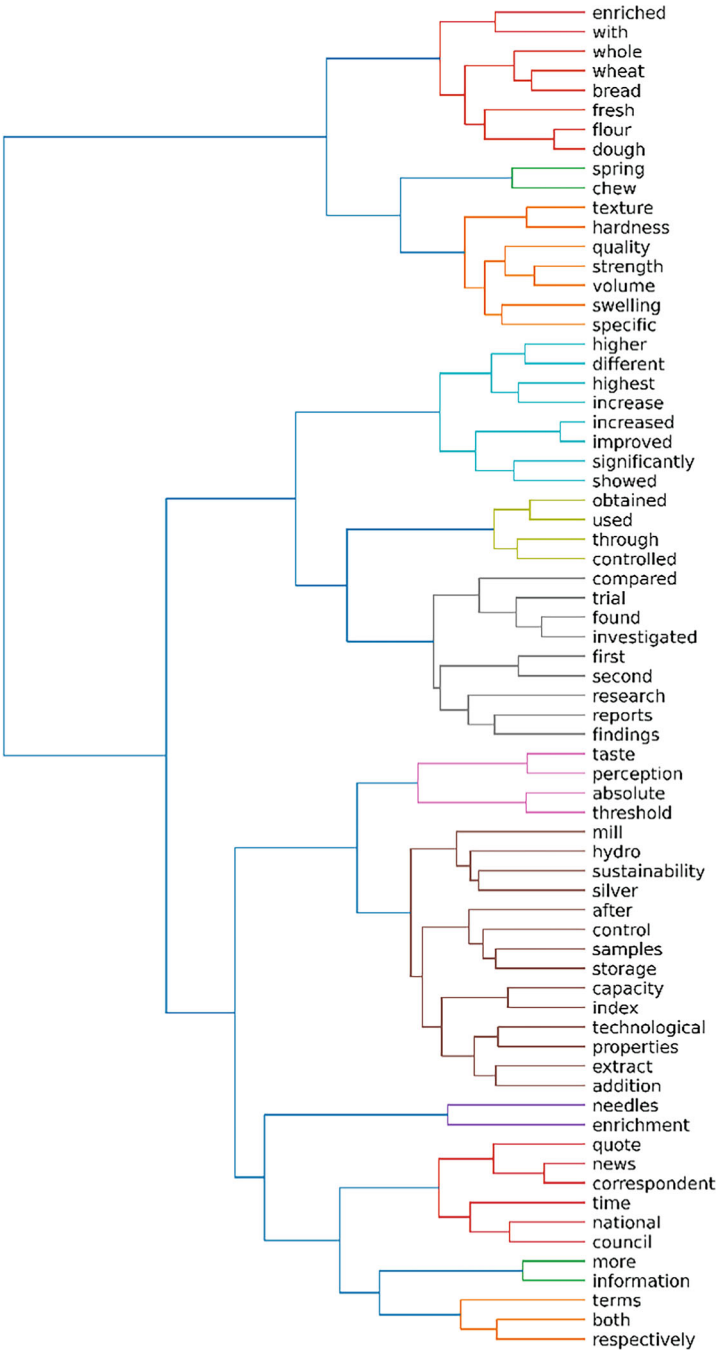


Fig. 11.6 An example of dendrogram. The embeddings are extracted using the pretrained model of base version of BERT algorithm, accessible using "bert-base-uncased" from Huggingface.com. The clustering is performed with regard to cosine distance between words of a sample of agriculture-related media news text

each point in time we might have one or more vectors. These time dependent vectors can be represented as sentiment using sentiment classification methods, and the effect of these time dependent sentiment on financial markets and economic factors can be investigated. Otherwise, the relationship between variability in the embedding vectors and the financial time series can be assessed directly using regression methods.

11.5 Experiment

In this section, examples of how an NLP model can be implemented for financial purposes is presented. The examples are comprised of several textual regression and sentiment regression approaches as well as a topic modelling task. In this spirit, first, the agricultural news data is collected from Refinitive's (formerly Thomson Reuters) Eikon platform for the time span between the 1st January, 2016 and 31st December, 2022. The raw text data is comprised of 411,883 news, of which approximately 75%, from beginning of 2016 to end of 2021, is used as the train data, and close to 25%, the whole 2022, as the out-of-sample test data.

Although the agricultural tags and keywords are used for filtering the news when collecting the texts, the news data is contaminated with unrelated news and texts. Therefore, the cleaning process is performed with obsession. This step consists of a combination of regex pattern matching, and dataset's specific metadata to eliminate the noise as much as possible, while retaining the information. The word cloud of the most frequent words before cleansing process is presented in Fig. 11.3. As shown on the figure, the most frequent words "*Refinitive*", "*consent*", "*service*", and "*https*" are not related to the market. The most important words for this study, i.e. "*wheat*", "*corn*", and "*soybean*" are approximately as frequent in the figure as words like "*said*" and "*copyright*", and less frequent than words like "*will*", "*party*", "*trademark*", and "*third*". This indicates the level of noise present in the raw text samples.

Comparing Fig. 11.3 with the word cloud after cleansing in Fig. 11.7 reveals the effectiveness of the cleansing process in retaining the text containing relevant words like "*wheat*", "*corn*", "*soybean*", "*price*", and "*cent*", while filtering unrelated and noise words. Interestingly, the word "*Ukraine*" can be seen as a comparatively frequent word which highlights the importance of the Russian invasion of Ukraine for the agriculture commodity.

The next step is to normalize the text for the NLP algorithms. It is worth mentioning that this study applies word normalization techniques, e.g. case folding, stop word removing, lemmatization, to the text only for the topic modelling task. Because the word normalization is not a necessary step for LLMs, i.e. BERT and FinBERT. Therefore, the normalization task starts by case folding, stop word removing, and lemmatization for topic modeling, and continues with segmenting the text into sentences. While it starts by segmenting the text into sentences for LLMs. These tasks are carried out using the tokenize module of NLTK library of Python.

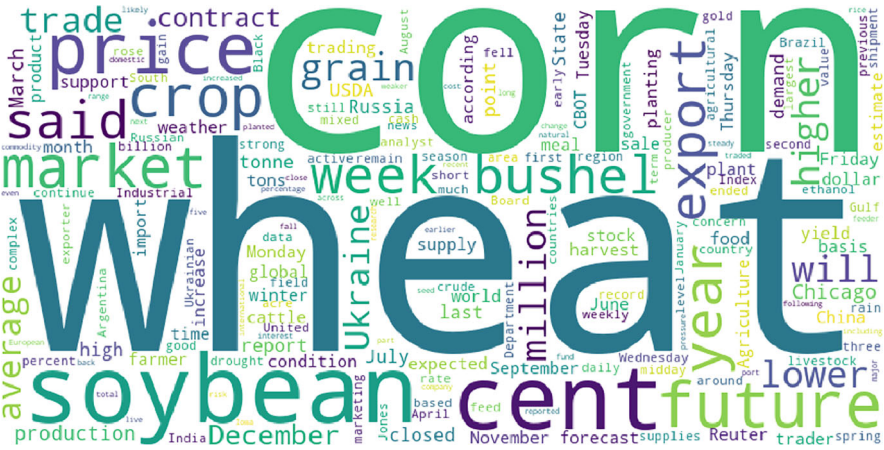


Fig. 11.7 Word cloud of the 2022 media news dataset after applying cleansing procedure

For the text and sentiment regression tasks, text normalization is continued by tokenizing the words. For this purpose, the tokenizer module of BERT is utilized. According to the documentation of BERT, it uses the WordPiece algorithm [24] to tokenize the words into subwords. The result of tokenization is approximately 7.5 billion tokens for the whole text dataset.

Continuing the analysis of the text, before implementing regression analysis, it would be beneficial to explore the general topics the news articles can be assigned to. In this regard, the text normalization steps are applied to the headlines of the news to prepare it for the algorithm. This task falls under the **topic modelling** task and has many methods, among which we opt for **Latent Dirichlet allocation (LDA)**. As one of the most popular and most widely used topic modelling algorithms, LDA is an unsupervised text clustering method that relies on the frequency matrices mentioned in Sect. 11.4.3 to convert the documents into vectors and assign each document a topic with regard to words in the document. Therefore, LDA is a bag-of-words method that disregards the relationship between the words in a document and assumes documents in each topic have a lot of words in common. Similar to common general clustering algorithms, the number of topics parameter is defined by the user (for more information see [44]).

To implement LDA, sklearn package of python is used to both constructing frequency matrices and applying LDA algorithm. The results are presented in Table 11.4 where each topic is represented with a number of terms. It depends on the user to assign a label to each topic, a task that may not be as easy because sometimes the words representing each topic can belong to various domains. However, there are topics whose labels can be easily assigned. For example, Topic 1 in Table 11.4 is more close to Russian invasion of Ukraine than others, the war which has affected the global agriculture production and trade. Topic 2 can be associated with the

Table 11.4 LDA topic modelling results for the full dataset

	Representing keywords	# of topics
Topic 1	'report, 'ukraine', 'food', 'russia', 'swap', 'crude', 'oil', 'price', 'inc', 'war'	873
Topic 2	'farmer', 'news', 'late', 'release', 'recent', 'press', 'trade', 'international', 'grain', 'stock'	919
Topic 3	'grain', 'wheat', 'soybean', 'future', 'rise', 'chicago', 'corn', 'fall', 'price', 'ukraine'	2428
Topic 4	'corn', 'grain', 'soy', 'bid', 'firm', 'soybean', 'cash', 'gulf', 'fob', 'cif'	1353
Topic 5	'report', 'oil', 'new', 'agricultural', 'ethanol', 'fas', 'daily', 'food', 'increase', 'service'	591
Topic 6	'corn', 'cbot', 'wheat', 'soybean', 'usda', 'fund', 'export', 'crop', 'net', 'table'	1569
Topic 7	'grain', 'farmer', 'inc', 'dtn', 'terminal', 'price', 'wheat', 'comment', 'cash', 'table'	1108

press news about agriculture, while Topic 6 might represent news related to USDA reports.

In the regression part, the base model of BERT, i.e. BERT-base, as well as its fine-tuned version for financial news sentiment classification FinBERT [45] are used as the LLMs. These LLMs, first, take a piece of text, transform it into tokens, add two special tokens [CLS] and [SEP] to the beginning and end of the text, respectively, and create a vector representation of length 768 for each token, including the special tokens. As the input sequence length is limited to 512 tokens for BERT, the input sequences are the sentences obtained from sentence segmentation step.

The embedding vectors are then used as the inputs for the regression and classification algorithms to either forecast corn's future price volatilities directly or classify the sentiment of the text. Figure 11.4 presents an example of embedding vectors for a sample of similar and non-relevant words extracted using above mentioned procedure. The figure presents only the first 100 dimensions for the sake of simplicity.

Using the embedding vector and Ridge regression method to estimate one-month-ahead (22 daily steps) of the realized volatility (RV) of corn constructs the **textual regression** models. Competing with textual regression models are the sentiment regression models where a classifier estimates the sentiment of each news using its embedding vectors and produces a sentiment index in the range [0, 1] for each sentiment class. The sentiment classes are negative, neutral, and positive represented by S_{neg}, S_{neu}, and S_{pos}, respectively.

The in-sample analysis for both textual regression and sentiment regression models are presented in Table 11.5. Following Manela and Moreira [46], the in-sample analysis of the textual regression models are performed using the estimated volatility ($\widehat{RV}$) after fitting the regression model on the textual embedding vectors produced by BERT and FinBERT because the embeddings, the independent variables of the regression, have 768 dimensions and analyzing such a high-dimensional regression model would not be feasible. While the independent variables of the sentiment

Table 11.5 Regression analysis of in-sample estimation for corn

	BERT	FinBERT
Panel A: Textual regression		
β_1	1.044^{***}	1.040^{***}
R^2	4.7	4.5
Panel B: Sentiment regression		
β_{neg}	$-20 \times 10^{-4***}$	$-1.67 \times 10^{-5***}$
β_{neu}	$-20 \times 10^{-4***}$	$-5.19 \times 10^{-5***}$
β_{pos}	$-5.27 \times 10^{-5***}$	$-2.09 \times 10^{-5***}$
R^2	0.00	0.2

Note: S_{neg}, S_{neu}, and $S_{pos} \in [0.1]$ are sentiment scores extracted by BERT and FinBERT algorithms. $\widehat{RV}$ is the estimated realized volatility extracted by fitting the regression model on the textual embedding vectors. The 1%, 5%, and 10% significance levels are represented by *, **, and *** respectively

Panel A represents textual regression models where the estimated $\widehat{RV}$, produced by fitting the regression model on the embedding vectors, are used as predictor, and panel B is the realized volatility regression using sentiment scores

regression models are the negative, neutral and positive sentiment scores extracted using the abovementioned models.

In general, the textual regression models provide significantly better fitting reflected by their higher R^2 values. The coefficients in panel A of Table 11.5 shows a slightly better fitting of BERT textual regression model compared to that of FinBERT's. However, the sentiments regression in panel B testifies the better performance of FinBERT's sentiment scores in explaining the realized volatility compared to BERT's sentiment scores. The better performance of FinBERT in sentiment regression is: expectable because this is the task-specific model optimized for financial news headlines sentiment classification. The regression details reveal a negative association between all the sentiments and the realized volatility, with a higher impact from neutral sentiment in the case of FinBERT and equal impact of negative and neutral, higher than positive, in BERT's case. However, it is important to note that R^2 values around 0 indicates the poor ability of sentiment regression models in explaining the volatility. Therefore, although their regression coefficients are statistically significant, the predictability power of sentiment-based models are so poor that might undermine the above conclusions about the impact of different sentiments.

The RV forecasting performance for out-of-sample 2022 dataset is presented in Table 11.6 where text regressors outperformed not only sentiment regression models, as was expected from in-sample analysis, but also powerful rivals like Heterogeneous Autoregressive model of realized volatility (HAR) [48]. In the table, RW and RM are the 60 days simple rolling window, and risk metrics approach that are widely used as an RV forecasting method in practice (see [49] for the formulae).

Table 11.6 Forecasting performance for out-of-sample RV of corn in 2022

	MSE ($\times 10^{-8}$)	% improvement of BERT-Ridge
BERT-Ridge	9.7550	–
FinBERT-Ridge	9.7542	−0.0082
$\text{BERT}_{\text{sent}}$-Ridge	10.348	5.7306^{**}
$\text{FinBERT}_{\text{sent}}$-Ridge	10.289	5.1900^{*}
HAR	11.090	12.038^{***}
RW	12.547	22.252^{***}
RM	11.313	13.772^{***}
ARIMA(1, 1, 1)	15.626	37.572^{***}

Note: The 1%, 5%, and 10% significance levels are represented by *, **, and *** respectively Diebold and Mariano test [47] significance level is reported by * annotation in the second column to investigate if the improvement of BERT-Ridge compared to its rivals is statistically significant

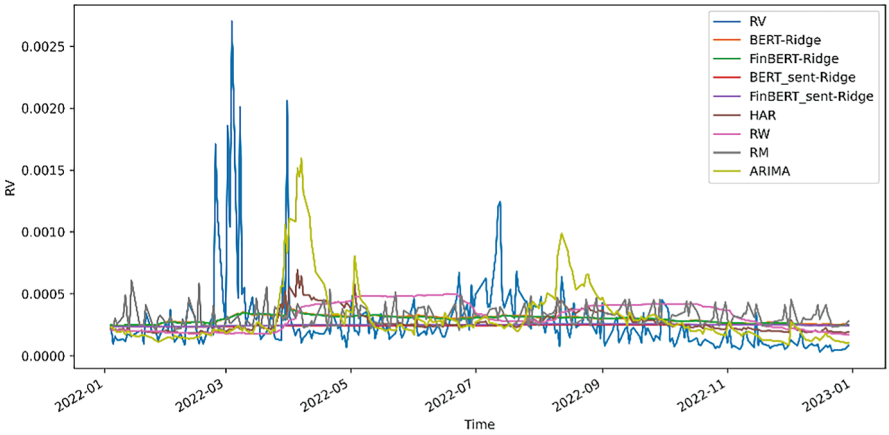


Fig. 11.8 Out-of-sample RV vs. forecasted values for out-of-sample corn dataset in 2022

Moreover, Diebold and Mariano test [47] shows that the forecasting improvement of textual regression models, proxied by BERT-Ridge, depicted in "**% improvement of BERT-Ridge"** column of the table is statistically significant. It is also noteworthy that there is approximately no difference between BERT and FinBERT in textual regression models, which is not surprising because the fine-tuning procedure only amends the classification layer of the BERT model to make it FinBERT. While textual regression models use the hidden state of the last transformer layer of FinBERT, a step before the classification layer, that is identical to that of BERT's. The forecasted vs observed values for all models is illustrated in Fig. 11.8.

11.6 Conclusion

This chapter aimed at introducing textual and sentiment analysis in agriculture market. It first reviewed research works carried out on sentiment analysis in agriculture market in general. Current textual and sentiment analysis works in agriculture market can be roughly divided into three main categories: sentiment proxy analysis; event analysis; news volume as sentiment proxy. This chapter introduced textual and sentiment analysis as a forth method. This method is implemented in other financial markets with a very promising outcomes. Therefore, as an introduction to textual and sentiment analysis for agriculture market, this chapter provided methods that are widely used in the finance. Finally, it presents examples of implementing textual and sentiment regression for forecasting one-month-ahead realized volatility of corn future volatility. Among text-based models, textual regression models that use text as input and forecasts corn's RV were the best model with a statistically significant superiority over their rivals including sentiment regression models and HAR. Additionally, sentiment regression models that use the LLM extracted sentiment to regress the RV were better than powerful traditional volatility forecasting models like HAR.

References

1. Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys, 34*(3), 512–547. https://doi.org/10.1111/joes.12370
2. Sinha, A., Kedas, S., Kumar, R., & Malo, P. (2022). SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology, 73*(9), 1314–1335. https://doi.org/10.1002/asi.24634
3. von Beschwitz, B., Keim, D. B., & Massa, M. (2018). First to 'read' the news: New analytics and algorithmic trading. *International Finance Discussion Paper, 2018*(1233). https://doi.org/10.17016/ifdp.2018.1233
4. Seo, Y.-W., Giampapa, J., & Sycara, K. (2004). Financial news analysis for intelligent portfolio management. *Defense Technical Information Center*. https://doi.org/10.21236/ada599073
5. Foucault, T., Hombert, J., & Roşu, I. (2016). News trading and speed. *The Journal of Finance, 71*(1), 335–382. https://doi.org/10.1111/jofi.12302
6. Zheng, Y. (2014). The linkage between aggregate stock market investor sentiment and commodity futures returns. *Applied Financial Economics, 24*(23), 1491–1513. https://doi.org/10.1080/09603107.2014.925073
7. Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x
8. Hamadi, H., Bassil, C., & Nehme, T. (2017). News surprises and volatility spillover among agricultural commodities: The case of corn, wheat, soybean and soybean oil. *Research in International Business and Finance, 41*, 148–157. https://doi.org/10.1016/j.ribaf.2017.04.006
9. Akyildirim, E., Cepni, O., Pham, L., & Uddin, G. S. (2022). How connected is the agricultural commodity market to the news-based investor sentiment? *Energy Economics, 113*, 106174. https://doi.org/10.1016/j.eneco.2022.106174

10. Borgards, O., & Czudaj, R. L. (2022). Long-short speculator sentiment in agricultural commodity markets. *International Journal of Finance & Economics., 28*, 3511–3528. https://doi.org/10.1002/ijfe.2605
11. Balcilar, M., Sertoglu, K., & Agan, B. (2022). The COVID-19 effects on agricultural commodity markets. *Agrekon, 61*(3), 239–265. https://doi.org/10.1080/03031853.2022.2078381
12. Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics, 228*(2), 221–243. https://doi.org/10.1016/j.jeconom.2020.07.053
13. Karali, B., Isengildina-Massa, O., Irwin, S. H., Adjemian, M. K., & Johansson, R. (2019). Are USDA reports still news to changing crop markets? *Food Policy, 84*, 66–76. https://doi.org/10.1016/j.foodpol.2019.02.005
14. Adjemian, M. K., & Irwin, S. H. (2020). The market response to government crop news under different release regimes. *Journal of Commodity Markets, 19*, 100110. https://doi.org/10.1016/j.jcomm.2019.100110
15. Indriawan, I., Martinez, V., & Tse, Y. (2021). The impact of the change in USDA announcement release procedures on agricultural commodity futures. *Journal of Commodity Markets, 23*, 100149. https://doi.org/10.1016/j.jcomm.2020.100149
16. Bian, S., Serra, T., Garcia, P., & Irwin, S. (2022). New evidence on market response to public announcements in the presence of microstructure noise. *European Journal of Operational Research, 298*(2), 785–800. https://doi.org/10.1016/j.ejor.2021.07.030
17. Cao, A. N. Q., & Robe, M. A. (2021). Market uncertainty and sentiment around USDA announcements. *Journal of Futures Markets, 42*(2), 250–275. https://doi.org/10.1002/fut.22283
18. Caporale, G. M., Spagnolo, F., & Spagnolo, N. (2016). Macro news and commodity returns. *International Journal of Finance & Economics, 22*(1), 68–80. https://doi.org/10.1002/ijfe.1568
19. Phan, H.-L., & Zurbruegg, R. (2019). The time-to-maturity pattern of futures price sensitivity to news. *Journal of Futures Markets, 40*(1), 126–144. https://doi.org/10.1002/fut.22046
20. Klomp, J. (2020). The impact of Russian sanctions on the return of agricultural commodity futures in the EU. *Research in International Business and Finance, 51*, 101073. https://doi.org/10.1016/j.ribaf.2019.101073
21. Sun, T.-T., Su, C.-W., Mirza, N., & Umar, M. (2021). How does trade policy uncertainty affect agriculture commodity prices? *Pacific-Basin Finance Journal, 66*, 101514. https://doi.org/10.1016/j.pacfin.2021.101514
22. Davis, S. J., Liu, D., & Sheng, X. S. (2019). Economic policy uncertainty in China since 1949: The view from mainland newspapers. *Fourth Annual IMF-Atlanta Fed Research Workshop on China's Economy Atlanta, 19*, 1–37.
23. Jurafsky, D., & Martin, J. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 3). https://web.stanford.edu/~jurafsky/slp3/
24. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*.
25. Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071
26. Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137. https://doi.org/10.1108/eb046814
27. Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.
28. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001* (p. 71). Lawrence Erlbaum Associates.
29. Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*. https://doi.org/10.3115/1220575.1220619

30. Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication, 45*(4), 363–407. https://doi.org/10.1177/0021943608319388
31. Loughran, T., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x
32. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*.
33. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051
34. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/d14-1162
35. Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010, September). *Recurrent neural network based language model*. Interspeech 2010. https://doi.org/10.21437/interspeech.2010-343
36. Olah, C. (2015). Understanding LSTM Networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems (Vol. 30)*. Curran Associates. https://proceedings.neurips.cc/paper_files/989paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
38. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems, 20*(1), 116–131. https://doi.org/10.1145/503104.503110
39. Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics, 41*(4), 665–695. https://doi.org/10.1162/coli_a_00237
40. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. https://doi.org/10.1037/0033-295x.104.2.211
41. Huang, E., Socher, R., Manning, C., & Ng, A. (2012). Improving word representations via global context and multiple word prototypes. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 873–882). Association for Computational Linguistics. https://aclanthology.org/P12-1092
42. Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1267–1273). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1128
43. Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). SemEval-2012 task 6: A Pilot on semantic textual similarity. In E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, & D. Yuret (Eds.), **SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 385–393). Association for Computational Linguistics. https://aclanthology.org/S12-1051
44. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet allocation. In *Advances in neural information processing systems 14* (pp. 601–608). The MIT Press. https://doi.org/10.7551/mitpress/1120.003.0082.
45. Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models*.

46. Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics, 123*(1), 137–162. https://doi.org/10.1016/j.jfineco.2016.01.032
47. Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*(1), 134–144. https://doi.org/10.1198/073500102753410444
48. Corsi, F. (2008). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics, 7*(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001
49. Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics, 160*(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Chapter 12
Applications of Singular Spectrum Analysis in Agricultural Financial Time Series

Rahim Mahmoudvand

Abstract In this chapter, we delve into the application of Singular Spectrum Analysis (SSA) for the examination and prediction of agricultural financial time series data. The erratic nature of agricultural markets is shaped by various factors, including seasonal trends, climatic conditions, and economic directives, posing a significant challenge for analysis. SSA stands out with its capacity to break down a time series into discernible components like trend, oscillatory elements, and noise, providing a sophisticated lens to interpret market dynamics.

The study utilizes SSA on a diverse array of agricultural financial time series data, including Fruit Planted Area, Fruit Home Production, Boxed Beef Prices for Choice and Select cuts, and CO2 Emission Intensity for rice commodities in European countries. We aim to achieve two primary goals: first, to unearth the intrinsic patterns and tendencies that dictate the movements of agricultural financial time series; and second, to project future trends, concentrating on enhancing strategies for investment and policymaking. Our findings highlight the prowess of SSA in sifting through the noise to uncover periodic behaviors and anomalies that conventional analysis might miss. The predictive model, founded on the reassembled components, exhibits notable precision in forecasting imminent price fluctuations, offering crucial insights to participants in the agricultural finance arena.

This research not only reaffirms the value of SSA in the realm of financial time series analysis but also sets the stage for its broader adoption in sectors where decoding intricate, non-linear patterns is of essence.

This study was funded by the European Union—NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018—CUP F53C22000760007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

R. Mahmoudvand (✉)
Department of Statistics, Bu-Ali Sina University, Hamedan, Iran
e-mail: r.mahmoudvand@basu.ac.ir; rahim.mahmoudvand@unica.it

Department of Economics and Business Sciences, University of Cagliari, Cagliari, Italy

H. Assa et al. (eds.), *Quantitative Risk Management in Agricultural Business*,
Springer Actuarial, https://doi.org/10.1007/978-3-031-80574-5_12

12.1 Introduction

The world we live in is bound by time, which I believe is a unique feature that influences everything we understand. Moreover, each element in the world can be described from different perspectives. Let's consider an example: how can we describe a fruit tree? We might examine the tree's height, circumference at different heights, the number of stems and leaves, the volume and quality of fruits, and the tree's overall health. These properties are called data and can be recorded over a period of time. Studying these data over time helps us extract useful information. For instance, the health of a tree can be assessed using its growth, height, circumferences, volume of fruits, and other related properties. Many questions might be addressed for such applications. For example, what level of growth per year reveals that a fruit tree is healthy? What is the normal pattern of the increment of height of a healthy fruit tree? How much does a healthy tree produce good quality fruit? Discussion about such questions can be considered from scientific perspectives and administrative points of view. Some countries have departments for Horticultural Affairs that help gardeners produce enough and high-quality fruits. Such departments might use data analysis to be aware of the future of fruit production and make appropriate decisions using scientific models.

Another related example in this area includes soft commodities, like corn, wheat, coffee, sugar, soybeans, and pork. These commodities are crucial for human survival as they directly relate to food and, to a lesser extent, clothing. Agricultural commodities have unique characteristics such as seasonality, perishability, and dependency on weather conditions, which can greatly affect their supply and prices. Modeling agricultural commodities is essential for maintaining the balance between supply and demand, ensuring economic stability, and securing food supplies on a global scale.

Simply, quantified data that is recorded and ordered at subsequent time points constitutes time series data. This ordering is usually through time, but other dimensions, such as spatial ordering, are sometimes encountered [25, 35]. These data can be found in many areas, including economics, environmental sciences, medical sciences, social sciences, engineering and this list is by no means complete. Let's continue with a theoretical definition for time series. Assume T and E are two arbitrary sets and define a time series as a set $\{y_t;\ t \in T, y_t \in E\}$ where y_t's are random variables. We call sets T and E as the index set and state space, respectively. Table 12.1 shows a classification of several cases.

Table 12.1 A classification of time series in terms of state space and index set

		Index set	
		Countable	Uncountable
State space	Countable	Case 1	Case 3
	Uncountable	Case 2	Case 4

This classification is not exhaustive, but it covers some of the main cases and concepts related to time series data. Let's provide examples for each case.

1. Let y_t show the number of fruit trees that have not any fruit in year t. Then, $\{y_t; t = 2010, \ldots, 2022\}$ is an example of the cases 1.
2. If we consider y_t as the total amount of fruits that have been produced by all trees at a city on years $t = 2010, \ldots, 2022$, then we provided an example for case 2. Additional examples of agricultural time series include the daily or weekly prices of commodities such as grains, fruits, and vegetables at wholesale markets.
3. Assume y_t shows the total number of broken fruit trees on time t. Then, y_t can be considered as a realization for case 3.
4. Denote by y_t the temperature of a location at time t, then it can be considered as an example of case 4. The volume of soft commodities at any time point is an example of a time series with an uncountable index set and a countable state space.

In practice, case 2 is more common than other cases. It is also possible to approximate other cases by case 2. A simple form of the time series is a set of observations like as $y_1, \ldots, y_n$ where in a more general form it can be considered as $y_{t_1}, \ldots, y_{t_n}$. We can also generalize the time series in terms of the dimension of the state space. Table 12.2 shows a classification of several cases along with an example in parenthesis for each cases. Each cases in Table 12.1 can be belong to all cases in Table 12.2 depending the dimension of state space.

Let provide several real examples that could be considered as real time series for each cases in Table 12.2.

1. Farm labor statistics: These time series represent the number of workers employed in agriculture during different seasons or years.
2. Weather data: Temperature readings, precipitation data, and wind speeds are examples of weather data. Assuming $\mathbf{y}_t = (T_t, P_t, W_t)$ shows the vector of temperature, precipitation and wind speed for hours $t = 1, \ldots, 24$, we are dealing with vector time series.

Table 12.2 A classification of time series in terms of the form and dimension of state space

		Form	
		Real-valued data	Functional form
Dimension	Zero dimension	Scalar (e.g: number of accidents)	Single function (e.g: financial transaction)
	Vector	Multiple dimension (e.g: weather data)	Vector of functions (e.g: stock prices)
	Matrix	Multiple dimension (e.g: water table depth)	Matrix of functions (e.g: satellite photos)

3. Coffee Prices: The data might be organized into a matrix time series dataset, where each row represents a specific time point or observation, such as a month, and each column represents a different region or market. The cells in the matrix contain the recorded prices of coffee at each time point and location.
4. Financial transaction: These data are treated as a sequence of functions observed over time. For instance, the daily curves of financial transaction data can be considered.
5. Stock prices: This data considers stock prices as a collection of curves observed sequentially over time.
6. Satellite photos: Each image can be considered a function, and a collection of these images observed over time can be organized into a matrix of functions.

Assigning the above examples into the categories in Table 12.2 is not strict; means that for instance the weather data could be recorded functionally and instead the financial transactions could be also recorded as scalar.

This short introduction is presented in my language, but there are plenty of books and useful sources that can be used for further details regarding the types of time series with various examples. For instance, Box, Jenkins, Reinsel, and Ljung's seminal work on time series analysis provides comprehensive coverage of various methods and applications [3]. Additionally, Brockwell and Davis offer an excellent introduction to time series and forecasting, with practical examples and applications [4]. For those interested in specific applications such as financial time series, Tsay's book on the analysis of financial time series is a valuable resource [32]. Moreover, Hyndman and Athanasopoulos' book on forecasting principles and practice offers practical insights and techniques for time series analysis [21]. These references, among others, provide a solid foundation for delving deeper into the complexities of time series analysis.

The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of the many conventional statistical methods that traditionally depend on the assumption of independent and identically distributed observations [30]. This problem is not new, having a theoretical background of more than a century. Therefore, there are strong reasons for developing new methods, which necessitates an overview of the attempts made so far

The rest of the current chapter is structured as follows: in Sect. 12.2, a brief review of the methods that can be applied for time series analysis is presented. Section 12.3 introduces the method that we are investigating as a powerful technique for time series analysis. Section 12.4 then generalizes the methods presented in Sect. 12.3, and finally, several examples are explained in detail.

12.2 Time Series Analysis

Analyzing time series data can be done for various purposes, such as smoothing, pattern recognition, change point and structural breaks detection, missing analysis, and forecasting. There are a variety of methods for time series analysis. Each method

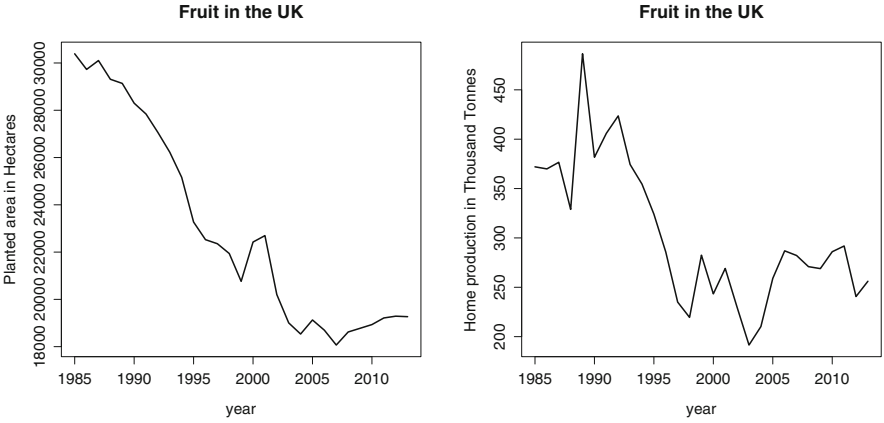


Fig. 12.1 Fruit Planted Area (left) and Fruit home production (right) in the UK 1985/1986–2013/2014 (prov)

has some advantages, disadvantages, or restrictions for the practice. There are two main approaches to analyzing time series data. The first approach involves using descriptive tools such as graphs, while the second approach requires the use of models and theoretical tools. When using descriptive tools, it is necessary to specify the scope of analysis. In the classical view of time series data, four components are typically explored: trend, cycle, seasonal, and irregular. However, exploring these components can be challenging and requires the expertise of the analyst. For example, Fig. 12.1 shows the time series of fruit planted area and fruit home production in the UK over a 29-year period, which exhibits a declining trend with no apparent seasonal component.

The results with descriptive tools are useful for a preliminary analysis and it is impossible to extract exact conclusions by them. Inferential tools comes into the field to produce quantitative results about the changes in time series and then provide useful conclusions. This approach work by model. There are a plenty of models that could be compared with together from different perspectives. Let mention some of the comparisions

- Linear versus Non-linear models: The difference between linear and non-linear time series models lies in the nature of the relationship between the variables. Linear time series models assume that the relationship between the variables can be described by a linear equation, such as a straight line. This means that the change in the dependent variable is proportional to the change in the independent variable. Examples of linear time series models include autoregressive (AR) and moving average (MA) models. Non-linear time series models, on the other hand, do not assume a linear relationship between the variables. Instead, they allow for more complex and non-linear relationships, such as exponential or quadratic relationships. Examples of non-linear time series models include the GARCH model and neural network models. In general, non-linear time series models are

more flexible and can capture more complex relationships between variables, but they can also be more difficult to estimate and interpret (see details in De Gooijer [5] and Tsay and Chen [33]). Linear time series models, on the other hand, are simpler and more interpretable, but may not capture the full complexity of the data. However, Linear models serve as a stepping stone for building a solid understanding of time series analysis, making them a common starting point in time series textbooks (see, for example, Wei [35] and Box et al. [3]).

- Parametric versus Non-Parametric models: The difference between parametric and non-parametric time series models lies in the assumptions they make about the underlying data. Parametric models assume a specific functional form for the data, such as a linear trend or a polynomial function, and estimate the parameters of this form from the data. Non-parametric models, on the other hand, do not make any assumptions about the functional form of the data and instead use the data itself to estimate the underlying relationship. This can make non-parametric models more flexible and robust, but they may require more data to estimate accurately.
- Time domain versus frequency domain methods: Time-domain techniques stem from classical correlation theory, focusing primarily on autocovariance and cross-covariance functions to develop autoregressive moving-average models for individual series and transfer-function models for causally related series. The parameter estimation methods used in these models often resemble advanced forms of linear regression. On the contrary, frequency-domain methods in spectral analysis extend Fourier analysis concepts, suggesting that any analytic function over a finite interval can be accurately approximated by a weighted sum of sine and cosine functions with increasing harmonic frequencies.

Pollock [28] has provided a good review of the time series analysis methods to that date. According to his review, the key pre-1900 time series analysis techniques included trend analysis, decomposition, harmonic/Fourier analysis, autocorrelation, and moving averages. These early approaches laid the groundwork for the more sophisticated time series models that emerged in the twentieth century. The autoregressive (AR) model is considered one of the oldest and most fundamental time series models, with origins dating back to the 1920s and 1930s. The combination of AR and MA models into the ARMA model in the 1970s was a significant milestone in the history of time series analysis. Extensive works for introducing new methods for analysing time series needs some strong reason such as uncertainty with respect to the future, model misspecification and data availability.

Time series analysis, like any analytical technique, faces its own set of challenges and limitations. Outliers and anomalies, which are observations that deviate significantly from the expected pattern, can distort the statistical properties of the data and affect the accuracy of forecasting models. Identifying and handling outliers appropriately is crucial to ensure reliable forecasts (for a review see Blázquez-García et al. [2]). One of the primary challenges of time series analysis is dealing with missing data, which can significantly impact the accuracy and reliability of time series models. An et al. [1] has conducted a review about this topic and

particularly evaluated the effects of imputation methods for replacing missing values with estimated values. Ribeiro [29] has also reviewed the missing values imputation methods and provided a case study in finanicial data. The choice of an appropriate model is another limitation in time series analysis, as each model has its own assumptions and limitations, and selecting the right model for a given dataset can be a complex task. Furthermore, time series analysis assumes that the underlying data is linear and follows a specific pattern. However, in many real-world scenarios, the data may exhibit non-linear patterns or dependencies, which can pose challenges in model selection and interpretation.

The non-linearity of time series data is particularly difficult to handle, as there are an infinite number of non-linear models to choose from. Finding the best model in such a vast space, be honestly, can be impossible. These challenges have led researchers and data analysts to look for simpler approaches for analyzing time series data. In my opinion, the theoretical restrictions on current parametric models, both linear and non-linear models, limit their applications. Therefore, approaches that don't start with a fixed and predefined model come into play. Singular Spectrum Analysis (SSA) is one such time series analysis method that relaxes most of the restrictive assumptions and tries to deny the time series at first, adjusting for the effects of shocks and outliers to some extent. SSA is employed for analyzing various time series in different fields, demonstrating its ability to provide appropriate results. It has been compared with numerous other time series analysis methods, showing its superiority in certain cases. For instance, Hassani [16], Hassani et al. [17], and Hassani et al. [19] compared SSA with Box-Jenkins SARIMA models, the ARAR algorithm, the Holt-Winter algorithm, exponential smoothing (ETS), and NN using well-known time series data sets, such as monthly accidental deaths in the USA, industrial production sectors in Germany, France, and the UK, and tourist arrivals into the US. They concluded that the SSA technique provides a much more accurate forecast than the other methods mentioned above. SSA is also used as a complementary method, mixed with other methods to eliminate the side effects of outliers and improve results (see, for example, Arteche and García-Enríquez [22]; Wang and Li [34]; Plazzi et al. [27]).

12.3 Singular Spectrum Analysis (SSA)

SSA operates within the realm of classical time series analysis and draws upon tools from multivariate statistics, multivariate geometry, dynamical systems, and signal processing to analyze time series data. With SSA, it is possible to decompose a time series into a small number of independent and interpretable components, such as a slowly varying trend, oscillatory components, and structureless noise. The literature on SSA includes over a hundred papers showcasing its superiority over other time series analysis techniques in various applications. Recent advancements in the theory and methodology of SSA can be found in Golyandina and Zhigljavsky [13]. The fundamental SSA method consists of two complementary stages: decomposition

and reconstruction. The first stage involves decomposing the time series, while the second stage focuses on reconstructing the noise-free time series. The reconstructed time series can then be used for forecasting new data points. A brief description of the SSA technique is provided below. More information can be found in Golyandina et al. [9], Golyandina and Zhigljavsky [11, 13] and Hassani and Mahmoudvand [20].

12.3.1 Basic SSA for Univariate Time Series (Smoothing)

Let start with a common cases of univariate time series data as $\mathbf{y}_N = \{y_1, \ldots, y_N\}$ where y_t is a real valued scalar. In summary, SSA change the vector of $\mathbf{y}_N$ to a Hankel matrix and then decompose the matrix into the sum of several marines where they are corresponding to the different components of the time series. Further details can be found in the following algorithm.

1. Embedding: Denote a parameter called window length by L, where is an integer between 2 and $N-1$ and $K = N - L + 1$, we define the trajectory matrix $\mathbf{Y}$ as below:

$$\mathbf{Y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_k \\ y_2 & y_3 & \cdots & y_k \\ \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_N \end{bmatrix} \tag{12.1}$$

2. Singular Value Decomposition (SVD): In this step, matrix $\mathbf{Y}$ will be decomposed using SVD as $\mathbf{Y} = \mathbf{Y}_1 + \cdots + \mathbf{Y}_d$, where $\mathbf{Y}_i = \sqrt{\lambda_i} U_i V_i^\top$ and $V_i = \mathbf{Y}^\top U_i / \sqrt{\lambda_i}$ with $\lambda_1 \geq \ldots \geq \lambda_L$, the eigenvalues of $\mathbf{S} = \mathbf{Y}\mathbf{Y}^\top$ and $U_1, \ldots, U_L$, the corresponding eigenvectors.
3. Grouping: The grouping step corresponds to splitting the elementary matrices into m disjunct subsets $I_1, \ldots, I_m$, and summing the matrices within each group. In the simplest case, we have $m = 2$, i.e. only two groups. $I_1 = \{1, \ldots, r\}$ and $I_2 = \{r+1, \ldots, L\}$ are related to the signal and noise components, respectively.
4. Diagonal averaging: The purpose of diagonal averaging is to transform each matrix $\mathbf{Y}_{I_j}$ into a new series of length N. Using diagonal averaging we have that $\mathbf{Y} = \widetilde{\mathbf{Y}}_{I_1} + \cdots + \widetilde{\mathbf{Y}}_{I_m}$, where $\widetilde{\mathbf{Y}}_{I_j}$ is the hankelized form of $\mathbf{Y}_{I_j}$, $j = 1, \ldots, m$. Considering $\tilde{y}_{m,n}^{(I_j)}$ the $(m,n)^{th}$ entry of the estimated matrix $\widetilde{\mathbf{Y}}_{I_j}$ and denoting by $\left\{\tilde{y}_{j_1}, \ldots, \tilde{y}_{j_T}\right\}$ the reconstructed components in the matrix $\widetilde{\mathbf{Y}}_{I_j}$, $j = 1, \ldots, m$, applying diagonal averaging follows that

$$\tilde{y}_{j_l} = \begin{cases} \frac{1}{s-1} \sum_{n=1}^{s-1} \tilde{y}_{n,s-n}^{(I_j)} & 2 \leq s \leq L-1, \\ \frac{1}{L} \sum_{n=1}^{L} \tilde{y}_{n,s-n}^{(I_j)} & L \leq s \leq K+1, \\ \frac{1}{K+L-s+1} \sum_{n=n-K}^{L} \tilde{y}_{n,s-n}^{(I_j)} & K+2 \leq s \leq K+L. \end{cases}$$

The vectors $\widetilde{\mathbf{y}}_{I_j} = \left\{\tilde{y}_{j_1}, \ldots, \tilde{y}_{j_T}\right\}$ denotes denoised form of the components of the time series and could be used for further analysis. Producing forecasts for future observations is one of the main purposes for time series analysis. Next section address this task by SSA.

12.3.2 Forecasting by SSA

The basic requirement to make SSA forecasting is that the time series satisfies a linear recurrent formula (LRF). A time series $\mathbf{y}_N = \{y_1, \ldots, y_N\}$ satisfies LRF of order d if:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_d y_{t-d}, \qquad t = d+1, \ldots, N. \tag{12.2}$$

Although there are several versions of univariate SSA forecasting algorithms we consider here two of the mostly widely used: Recurrent SSA (RSSA) [6, 7] and Vector SSA (VSSA) [26]. In what follows, we give a brief description of these algorithms. Further details can be found in Golyandina et al. [9].

We should mention at first that the following approach assumes that a single component selected for forecasting. In the common and simplest case, we assume that the time series is simply as the sum of signal and noise components and the signal component extracted by using the first r einentriples. We explain forecasting algorithm for this case; however it can be easily considered for each components. Let us assume that $U_j^{\triangledown}$ is the vector of the first $L-1$ components of the eigenvector U_j and π_j is the last component of U_j $(j = 1, \ldots, r)$. Denoting $\upsilon^2 = \sum_{j=1}^{r} \pi_j^2$ we define the coefficient vector $\mathfrak{R}$ as:

$$\mathfrak{R} = \frac{1}{1-\upsilon^2} \sum_{j=1}^{r} \pi_j U_j^{\triangledown}.$$

12.3.2.1 Recurrent SSA

Considering the above notation, the RSSA forecasts $\left(\hat{y}_{N+1}, \ldots, \hat{y}_{N+M}\right)$ can be obtained by

$$\hat{y}_i = \begin{cases} \tilde{y}_i, & i = 1, \ldots, N \\ \mathfrak{R}^{\top} Z_i, & i = N+1, \ldots, N+M \end{cases}, \tag{12.3}$$

where, $Z_i = \left[\hat{y}_{i-L+1}, \ldots, \hat{y}_{i-1}\right]^{\top}$ and $\tilde{y}_1, \ldots, \tilde{y}_N$, are the values for the reconstructed time series and can be obtained from 4th Step in above mentioned algorithm.

12.3.2.2 Vector SSA

Define linear operator:

$$\mathcal{P}^{(v)}Y = \begin{pmatrix} \mathbf{\Pi} Y_{\triangle} \\ \mathfrak{R}^T Y_{\triangle} \end{pmatrix}, \; Y \in \text{span}\{U_1, \ldots, U_r\}, \tag{12.4}$$

where $\mathbf{\Pi} = \mathbf{U}^{\triangledown}\mathbf{U}^{\triangledown T} + (1 - v^2)\mathfrak{R}\mathfrak{R}^T$ and $Y_{\triangle}$ denotes the last $L - 1$ elements of Y. Suppose the vector Z_j is defined as follows

$$Z_j = \begin{cases} \widetilde{Y}_j & \text{for } j = 1, \ldots, K \\ \mathcal{P}^{(v)} Z_{j-1} & \text{for } j = K+1, \ldots, K+M+L-1 \end{cases}, \tag{12.5}$$

where $\widetilde{Y}_j$ are the j^{th} reconstructed columns of the trajectory matrix of the time series after grouping and discarding noise components. Now, by constructing the matrix $\mathbf{Z} = [Z_1, \ldots, Z_{K+M+L-1}]$ and performing diagonal averaging, we obtain a new time series $\hat{y}_1, \ldots, \hat{y}_{N+M+L-1}$, where $\hat{y}_{N+1}, \ldots, \hat{y}_{N+M}$ form the M terms of the VSSA forecast.

12.3.3 SSA Parameter Selection

The SSA calibration depends upon two basic, but very important, parameters: the window length L, and the number of eigentriples used for reconstruction r. The choice of improper values for the parameters L or r yield incomplete reconstruction and the forecasting results might be misleading. Despite the importance in choosing proper values for these parameters, no theoretical solution has been proposed to solve this problem. Some of the techniques to choose the appropriate value of L can be found in Golyanidina [8], Hassani et al. [18], Mahmoudvand and Zokaei [24] and Mahmoudvand et al. [23]. An overall agreeable suggestion to choose the window length is to have it close to the middle of the series and proportional to the number of observations per period (e.g. to 12 for monthly time series, to four for quarterly time series, etc.). However, this choice does not guarantee the best predictions (e.g. Mahmoudvand, et al. [23]). For better results, the parameter choice should be made accordingly to available data and intended analysis.

In practice it is relatively rare that the number of singular values r, needed to be selected to reconstruct noise free series from a noisy time series, is known a priori. Among several ways to determine r described in the literature, the easiest way is done by checking breaks in the eigenvalues spectra. As a rule of thumb, a pure noise series produces a slowly decreasing sequences of singular values. Another useful insight is provided by considering separability between signal and noise components, which is a fundamental concept in studying SSA properties, by using w-correlations [9] between two vectors $\mathbf{y}^{(1)} = [y_1^{(1)}, \ldots, y_N^{(1)}]^{\top}$ and

$\mathbf{y}^{(2)} = [y_1^{(2)}, \ldots, y_N^{(2)}]^\top$:

$$\rho_w = \frac{\langle \mathbf{y}^{(1)}, \mathbf{y}^{(2)} \rangle_w}{\sqrt{\langle \mathbf{y}^{(1)}, \mathbf{y}^{(1)} \rangle_w \langle \mathbf{y}^{(2)}, \mathbf{y}^{(2)} \rangle_w}}, \tag{12.6}$$

where, $w_j^{L,N} = \min\{j, L, N - j + 1\}$ and $\langle \mathbf{y}^{(m)}, \mathbf{y}^{(n)} \rangle_w = \sum_{j=1}^{N} w_j^{L,N} y_j^{(m)} y_j^{(n)}$ for $m, n = 1, 2$. According to this measure, two series are separable if the absolute value of their w-correlation is small. Therefore, we determine the groups in such a way that the reconstructed components in the same group have a high w-correlation and a small w-correlation with the components in other groups. Plotting pair of eigenvectors help us also to see which components may belong to the same group. Another way to determine r is by examining the forecast accuracy, i.e. r is determined in such a way that the minimum error in forecasting will be obtained.

12.3.4 Examples

Let see how SSA is working in practice. Consider a simulated time series with a deterministic signal as below:

$$y_t = 3 \sin\left(\frac{2\pi t}{12}\right) + \epsilon_t \quad , \quad t = 1, \ldots, 50 \tag{12.7}$$

Assuming ϵ_t come from a normal distribution with mean zero and standard deviation 0.5, Fig. 12.2 shows a realisation of such series.

Figures 12.3 and 12.4 shows singular values, w-correlation and paired eigenvectors that helped us to specify the right number of components for reconstruction. It should be mentioned that, here we have considered $L = 25$ and using the results of the aforementioned figures confirm that the first two einentriples are in the same group and this group is enough to extract the signal from noise.

Let see the results for the fruit time series. The measures and recommendations confirm that the first five components are enough for signal extraction by SSA (Figs. 12.5 and 12.6).

To assess the quality of forecasting using SSA for this time series, we utilized 5 components to reconstruct the time series and generate forecasts for various forecast horizons. To validate the results, an expanding window approach was employed for each forecast horizon. This method involves using an expanding dataset for training and testing the forecasting model. For example, the process involved using y_1 to y_{29} to forecast observation y_{30}, then expanding the training set to include y_1 to y_{30} to forecast y_{31}, and continuing this process up to the final forecast, which in this case is y_{38}. By following this approach, we obtained 9 one-step-ahead forecast errors, providing insight into the accuracy of the forecasting method.

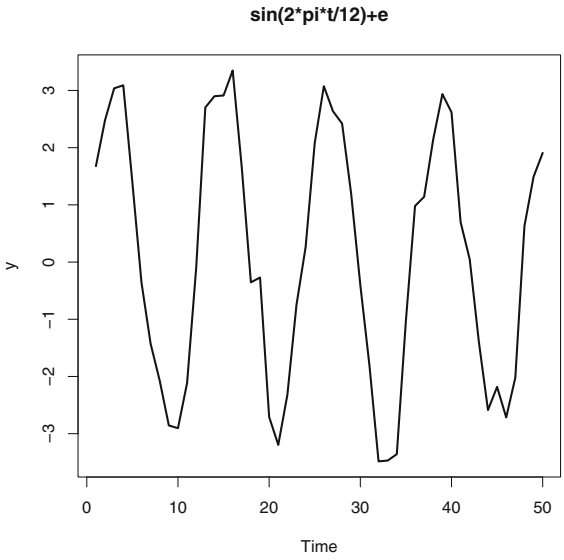


Fig. 12.2 A realisation of a synthetic data

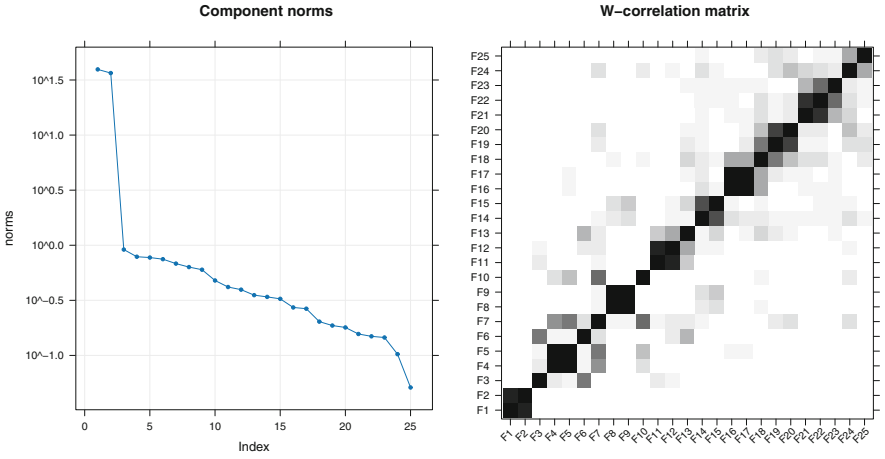


Fig. 12.3 Singular values and w-correlation for the synthetic data

The results are presented in Table 12.3. The table compares the forecasting quality of different SSA algorithms with ARIMA and ETS models across various forecast horizons. The evaluation is based on two key metrics: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). For a one-year forecast horizon, the ARIMA algorithm outperforms others with the lowest RMSE (77.59) and MAPE (0.10), followed by ETS, RSSA, VSSA, and Bootstrap. In the case of a two-year forecast horizon, ARIMA continues to exhibit superior performance with the lowest RMSE (116.22) and MAPE (0.16). However, the SSA algorithms demonstrate better performance for longer-term forecasts (3–6 years).

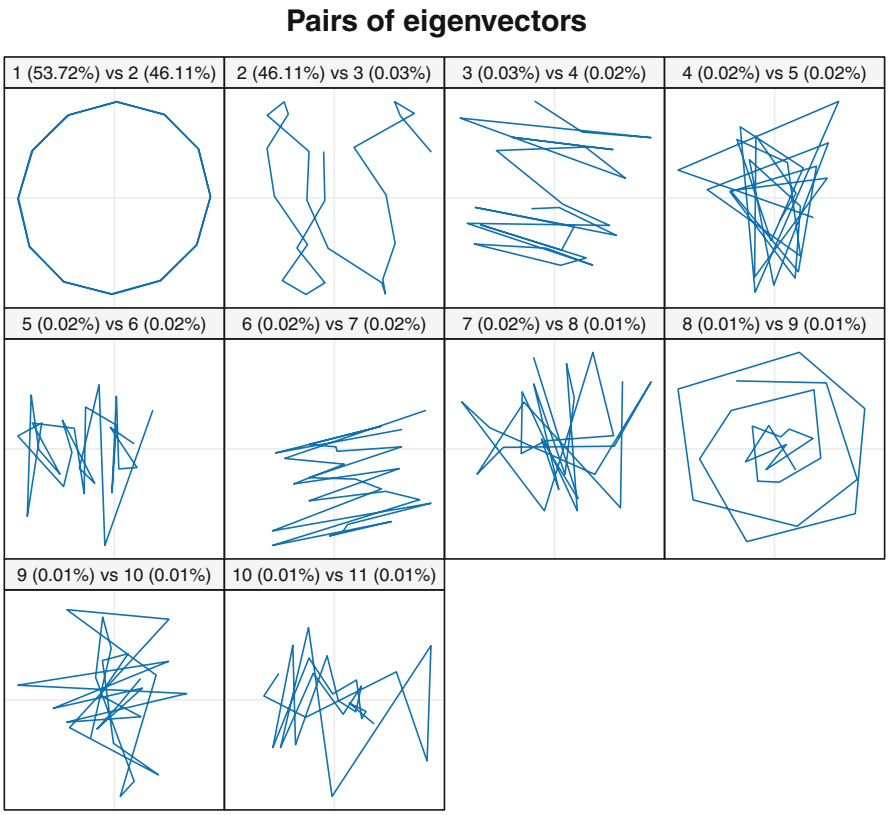


Fig. 12.4 Paired eigenvectors of a synthetic data

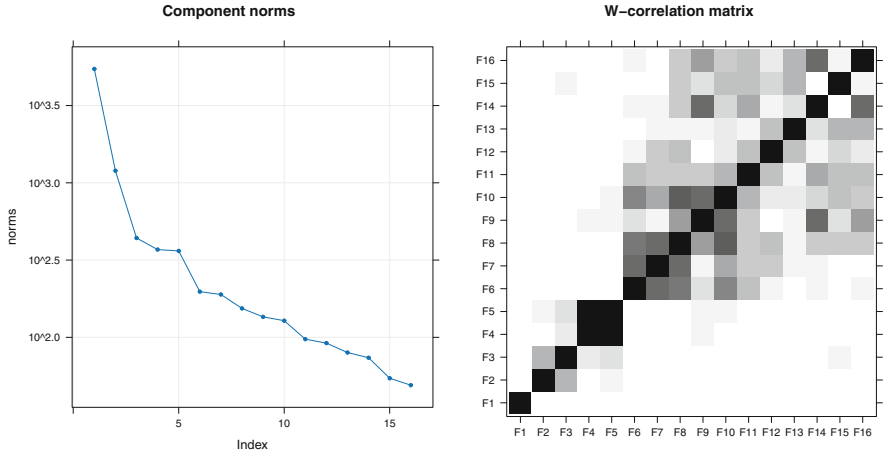


Fig. 12.5 Scree plot of the Singular values of the fruit time series data

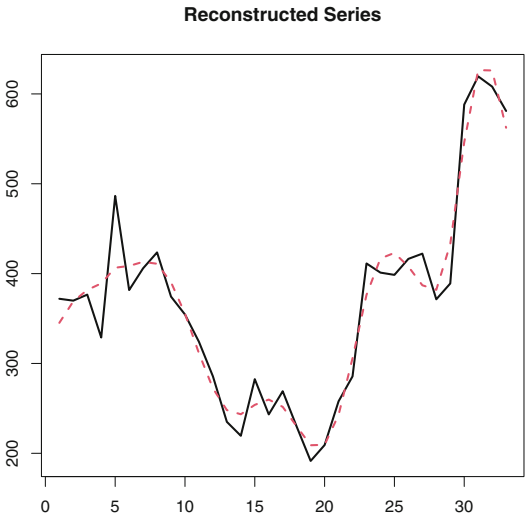


Fig. 12.6 Reconstruction of the time series using the first five components

Table 12.3 Comparing the quality of forecasts using different algorithms of SSA

Horizon	Measure	ARIMA	ETS	RSSA	VSSA	Bootstrap
1	RMSE	**77.59**	81.86	80.95	85.72	82.78
	MAPE	**0.10**	0.11	0.13	0.14	0.14
2	RMSE	**116.22**	116.80	125.58	124.96	118.94
	MAPE	**0.16**	0.16	0.21	0.22	0.20
3	RMSE	140.51	139.52	152.38	**128.42**	138.60
	MAPE	**0.21**	0.22	0.27	0.22	0.24
4	RMSE	155.99	153.90	152.98	**116.63**	131.83
	MAPE	0.26	0.25	0.25	**0.20**	0.21
5	RMSE	165.92	164.04	142.45	197.26	**137.03**
	MAPE	0.30	0.28	**0.21**	0.29	0.22
6	RMSE	167.36	167.76	**156.19**	363.61	272.98
	MAPE	0.30	0.29	**0.23**	0.46	0.36

We have highlighted the best results with the minimum values in each row using bold font.

12.4 Extension to the Basic SSA

We have mentioned in the introduction that many cases could be considered as time series. One cases assumes a multivariate quantity for each time points. It is easily possible to extend SSA to consider the multivariate time series. There are several version for doing such; however we follow the approach of Golyandina. Multivariate

SSA, or MSSA, is a natural extension of SSA for analysing multivariate time series. Let $Y_t = \left[y_t^{(1)}, \ldots, y_t^{(M)}\right]$, $t = 1, \ldots, N$, denote a sample of a M-variate time series with length N. Note that it is possible to consider different number of observations for the individual time series in the multivariate framework, but we first assume equal number of observations. Let us assume that Y_t can be written in terms of a signal plus noise model as:

$$\mathbf{Y}_N = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} y_1^{(1)} & \ldots & y_1^{(M)} \\ y_2^{(1)} & \ldots & y_2^{(M)} \\ \vdots & & \vdots \\ y_N^{(1)} & \ldots & y_N^{(M)} \end{bmatrix} = \begin{bmatrix} s_1^{(1)} & \ldots & s_1^{(M)} \\ s_2^{(1)} & \ldots & s_2^{(M)} \\ \vdots & & \vdots \\ s_N^{(1)} & \ldots & s_N^{(M)} \end{bmatrix} + \begin{bmatrix} n_1^{(1)} & \ldots & n_1^{(M)} \\ n_2^{(1)} & \ldots & n_2^{(M)} \\ \vdots & & \vdots \\ n_N^{(1)} & \ldots & n_N^{(M)} \end{bmatrix}.$$

The MSSA algorithm denoises (smooths) the multivariate time series $\mathbf{Y}_N = [Y_1, \ldots, Y_N]^\top$ using the same steps as the univariate SSA, i.e. embedding, SVD, grouping and reconstruction. The only difference is related to the definition of trajectory matrix. Although there are several forms to define the trajectory matrix in MSSA, here we use the stacked form of the univariate trajectory matrices. The simplest case include the horizontal form defined as:

$$\mathbf{Y} = \left[\mathbf{Y}^{(1)} \ldots \mathbf{Y}^{(M)}\right] =$$

$$\begin{bmatrix} y_1^{(1)} & y_2^{(1)} & \ldots & y_K^{(1)} & \ldots & y_1^{(M)} & y_2^{(M)} & \ldots & y_K^{(M)} \\ y_2^{(1)} & y_3^{(1)} & \ldots & y_{K+1}^{(1)} & \ldots & y_2^{(M)} & y_3^{(M)} & \ldots & y_{K+1}^{(M)} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ y_L^{(1)} & y_{L+1}^{(1)} & \ldots & y_T^{(1)} & \ldots & y_L^{(M)} & y_{L+1}^{(M)} & \ldots & y_N^{(M)} \end{bmatrix}.$$

where L and K are chosen similarly as before and $\mathbf{Y}^{(j)}$ is an Hankel matrix for the column j of the $\mathbf{Y}_N$. This means that the trajectory matrix for the MSSA algorithm is a block Hankel matrix and this property is considered for the reconstruction step.

Denote by $\widetilde{Y}_N^{(j)} = (\widetilde{y}_1^{(j)}, \ldots, \widetilde{y}_N^{(j)})^\top$ the reconstructed values of the time series j. The $h-$steps ahead forecasts obtained by the MSSA algorithm can be obtained with the following recursive formula:

$$\widehat{y}_{h|N}^{(j)} = \begin{cases} \widetilde{y}_h^{(j)} & h = 1, \ldots, N \\ \sum_{t=1}^{L-1} a_t \widehat{y}_{(h-t)|N}^{(j)} & h = N+1, \ldots \end{cases} \tag{12.8}$$

where $R = (a_{L-1}, \ldots, a_1)^\top$ will be obtained from trajectory matrix $\mathbf{Y}$ similarly as in the univariate SSA-R. The methodology of the multivariate SSA shows that it also needs two choices for its application in practice: the window length L, and the cutting point r. We can determine these values using the same approaches mentioned for the univariate SSA. It is worth mentioning that the complexity of

the multivariate SSA model is smaller than the univariate SSA, when we apply both models for analysing multivariate time series. This happens because multivariate SSA needs two choices (L, r) whereas the univariate SSA needs $2M$ choices $\{(L_1, r_1), \ldots, (L_M, r_M)\}$. In addition, the window length in univariate time series is recommended to be less than half of the series length. However, here for the HMSSA, it is possible to consider larger values for the window length. Because the number of columns for the trajectory matrix increases by the dimension of the time series and if we consider L larger than half of series length it is possible to obtain more information than the cases with L chose to the half of series length. Another point that might be mentioned is related to the time series lengths. Although we have considered the same series lengths for all dimensions of the time series, in practice it is possible to have time series with different lengths. However, it is obvious from the definition of the trajectory matrix that there is no restriction as we can consider the same window length for all components and stack the univariate Hankel matrices horizontally.

12.4.1 Further Extension to MSSA

There are several methods for extending MSSA. The first idea is to define trajectory matrix by stacking univariate trajectory matrices vertically. In this case, we have to define window length in such a way that $L_i - N + 1$ are the same for all time series. The second idea is related to the forecasting engine. As we mentioned in the univariate case, we have two different approaches for producing forecasts for future observations: recurrent and vector approaches. Combining these approaches with two methods for defining trajectory matrix produce four cases. These cases is presented in Table 12.4.

In what follow, we explain each algorithm briefly. For simplicity in writing the equations, denote by $\mathbf{Z}[, j]$ and $\mathbf{Z}[i,]$, the $j-$th column, and the $i-$th row of the matrix $\mathbf{Z}$, respectively.

Table 12.4 Possible forecasting algorithms for multivariate SSA

Trajectory form	Forecasting method	Abbreviation
Horizontal	Recurrent	HMSSA-R
	Vector	HMSSA-V
Vertical	Recurrent	VMSSA-R
	Vector	VMSSA-V

12.4.2 HMSSA-R(V)

Denote by $\mathbf{U}_r$ the matrix of the first r eigenvectors of $\mathbf{Y}\mathbf{Y}^\top$ corresponding to the r largest singular values of $\mathbf{Y}$ and assume that $\mathbf{U}_r^\nabla$ and $\mathbf{U}_{\nabla r}$ are the first $L-1$ rows of $\mathbf{U}_r$ and last row of $\mathbf{U}_r$, respectively. In addition, define:

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ 0 & \hat{\mathcal{A}} \end{bmatrix} \quad , \quad \hat{\mathcal{A}} = \left(1 - \mathbf{U}_{\nabla r}\mathbf{U}_{\nabla r}^\top\right)^{-1} \mathbf{U}_{\nabla r}\mathbf{U}_r^{\nabla\top}, \tag{12.9}$$

where $\mathbf{I}$ is the $(L-1)\times(L-1)$ identity matrix and $\mathbf{0}$ is a column vector with $L-1$ zeros. Then the h-steps ahead forecasts by HMSSA-R can be obtained by:

$$\hat{y}_{N+h}^{(m)} = \mathbf{W}^h[L,]\tilde{\mathbf{S}}^{(m)}[,mK], \quad m = 1,\ldots,M, \quad h = 1,2,\ldots, \tag{12.10}$$

where $\mathbf{W}^h$ represents the h power of the matrix $\mathbf{W}$. If we made a change in matrix $\mathbf{W}$ as below:

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{\Pi} \\ 0 & \hat{\mathcal{A}} \end{bmatrix} \quad , \quad \mathbf{\Pi} = \mathbf{U}_r^\nabla \mathbf{U}_r^{\nabla\top} + \hat{\mathcal{A}}^\top(1 - U_{\nabla r}U_{\nabla r}^\top)\hat{\mathcal{A}}, \tag{12.11}$$

Then the h-steps ahead forecasts by HMSSA-V can be obtained by:

$$\hat{y}_{N+h}^{(m)} = \frac{1}{L}\sum_{\ell=h}^{h+L-1} W^\ell[L-\ell+h,]\tilde{S}^{(m)}[,mK], \quad m = 1,\ldots,M, \quad h = 1,2,\ldots, \tag{12.12}$$

where $W^h[\ell,]$ denotes the $l-$th row of $\mathbf{W}^h$.

12.4.3 VMSSA-R(V)

Denote by $\mathbf{U}_r$ the matrix of the first r eigenvectors of $\mathbf{Y}\mathbf{Y}^\top$ corresponding to the r largest singular values of $\mathbf{Y}$ and assume that $\mathbf{U}_r^\nabla$ is constructed by removing the rows $L, 2L, \ldots, ML$ from $\mathbf{U}_r$, and $\mathbf{U}_{\nabla r}$ is the matrix that is constructed by stacking the rows $L, 2L, \ldots, ML$ of $\mathbf{U}_r$. In addition, define:

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ 0 & \hat{\mathcal{A}}_0[1,] \\ \mathbf{0} & \mathbf{I} \\ 0 & \hat{\mathcal{A}}_0[2,] \\ \vdots & \vdots \\ \mathbf{0} & \mathbf{I} \\ 0 & \hat{\mathcal{A}}_0[M,] \end{bmatrix}, \quad \hat{\mathcal{A}} = \left(I_{M\times M} - \mathbf{U}_{\triangledown r}\mathbf{U}_{\triangledown r}^{\top}\right)^{-1}\mathbf{U}_{\triangledown r}\mathbf{U}_{r}^{\triangledown\top}, \tag{12.13}$$

where $\mathbf{I}$ is the $(L-1)\times(L-1)$ identity matrix, $\mathbf{0}$ is a column vector with $L-1$ zeros and $[0, \hat{\mathcal{A}}_0[i,]]$ is a vector of size LM where before each $L-1$ elements of $\hat{\mathcal{A}}[i,]$ a zero is added ($i = 1, \ldots, M$). Then the h-steps ahead forecasts by VMSSA-R can be obtained by:

$$\hat{y}_{N+h}^{(m)} = \mathbf{W}^h[mL,]\tilde{\mathbf{S}}^{(m)}[,K], \quad m = 1, \ldots, M, \quad h = 1, 2, \ldots, \tag{12.14}$$

where $\mathbf{W}^h$ represents the h power of the matrix $\mathbf{W}$. If we made a change in $\mathbf{W}$ as below:

$$\mathbf{W} = \begin{bmatrix} \mathbf{0} & \mathbf{\Pi}_1 \\ 0 & \hat{\mathcal{A}}_0[1,] \\ \mathbf{0} & \mathbf{\Pi}_2 \\ 0 & \hat{\mathcal{A}}_0[2,] \\ \vdots & \vdots \\ \mathbf{0} & \mathbf{\Pi}_M \\ 0 & \hat{\mathcal{A}}_0[M,] \end{bmatrix}, \quad \mathbf{\Pi} = \mathbf{U}_r^{\triangledown}\mathbf{U}_r^{\triangledown\top} + \hat{\mathcal{A}}^{\top}(I_{M\times M} - U_{\triangledown r}U_{\triangledown r}^{\top})\hat{\mathcal{A}}, \tag{12.15}$$

where $\hat{\mathcal{A}}$ is defined and divided as VMSSA-R in Appendix A.3, $\mathbf{0}$ is a column vector with $L-1$ zeros and $\mathbf{\Pi}_j$ represents the rows number $(j-1)(L-1)+1, \ldots, j(L-1)$ of $\mathbf{\Pi}$, $j = 2, \ldots, M$. Then the h-steps ahead forecasts by VMSSA-V can be obtained by:

$$\hat{y}_{N+h}^{(m)} = \frac{1}{L}\sum_{\ell=h}^{h+L-1} W^{\ell}[L-\ell+h,]\tilde{S}^{(m)}[,K], \quad m = 1, \ldots, M, \quad h = 1, 2, \ldots, \tag{12.16}$$

where $W^h[\ell,]$ denote the $l-$th row of $\mathbf{W}^h$.

Let provide examples that can be used to see how MSSA works in practice. Using simulation could be better as we are aware from the true relationships. One challenges here is related to the interpretation of the components.

12.4.4 Comparing MSSA Methods

One of the common questions often asked pertains to the methodology for selecting the best method. First and foremost, it is important to acknowledge that there is no definitive answer to this question. This implies that both approaches, HMSSA and VMSSA, could exhibit similar performance in certain problems, while one may outperform the other in different scenarios. However, we can compare them theoretically. Considering the methodology of each approach, we can conclude the following:

- HMSSA-R(V) employs the same Linear Recurrent Formula (LRF) for all time series, whereas VMSSA-R(V) utilizes specialized LRFs for each time series.
- HMSSA provides L singular values, whereas VMSSA provides $M \times L$ singular values, indicating that the methods have different complexities. However, it may be possible to consider different window lengths to observe the same dimensionality.

12.4.4.1 Empirical Comparison

Empirical comparison is crucial in evaluating the performance of SSA methods as it provides objective evidence of their effectiveness in real-world scenarios. By comparing the outcomes of different SSA and MSSA techniques through empirical data, researchers can validate the reliability and accuracy of these methods, ensuring their applicability in practical forecasting and analysis. We have examined two data sets.

12.4.4.2 Data Set 1: Boxed Beef Price

Monthly Boxed beef prices for Choice and Select cuts in USA over the period 2000 to 2022 are presented in Fig. 12.7, that is publicly available on the USDA webpage. This figure illustrates a positive correlation between both series. The Pearson correlation coefficient between the series is 0.951, confirming a strong positive relationship. Notably, one of the time series contains two missing observations in the middle, which may not be clearly visible in the graph. It is imperative to address this gap before proceeding with analysis. Various approaches can be employed for handling missing data, such as the missing analysis approach or simple imputation methods like replacing missing points with the average of nearest observations.

Let us provide the results using different methods. Firstly, note that observations 125 and 126 are missing. Hence:

1. Method 1: Observation 125 is replaced with observation 124, which is 167.09, and observation 126 is replaced by observation 127, which is 165.46.

2. Method 2: Observations 125 and 126 are replaced with the average of observations 124 and 127, resulting in a value of 166.275.
3. Method 3: For observations 1 to 124, SSA is employed to forecast observations 125 and 126, resulting in values of 146.87 and 147.58, respectively (using the first three eigen triplets). Similarly, using observations 288 to 127, forecasts of 164.12 and 158.26 are obtained for observations 125 and 126 (using the first 5 eigen triplets). Averaging these values yields estimates of 155.5 and 152.92, respectively.

Choosing the estimation method for missing values poses a challenge due to the differing values obtained through various approaches. However, I assume that SSA provides a better estimation.

We employed MSSA to analyze these datasets as a bivariate time series. Note also that the univariate SSA can be applied to each time series separately, albeit we anticipate more accurate results from MSSA compared to SSA. Therefore, we employ both approaches and compare the results.

Let's divide the dataset into two parts: (1) observations from 2000 to 2022 as the training set and (2) all 12 observations of 2023 for testing. In what follows, I compared the effect of window length on the singular value pattern. Figure 12.8 display the scree plot of the singular values for the MSSA using window lengths of 36, 48, 60 and 72; however, I checked it also for 84, 96, 108, and 120 but I have not reported the results here as the results and conclusion were the same. As observed in these plots, the first four eigentriples suffice to capture the most useful part of the time series. It is noteworthy that all plots draw the same conclusion.

We employed a window length of $L = 84$ and trained the model using data up to the last month of 2022. Subsequently, we generated forecasts for the 12 months

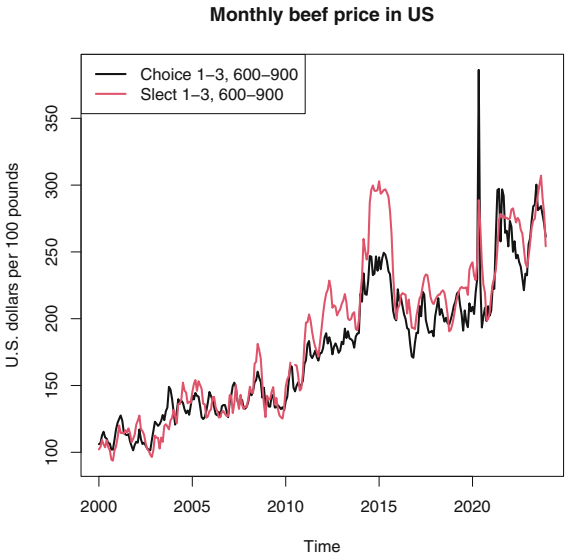


Fig. 12.7 Meat prices

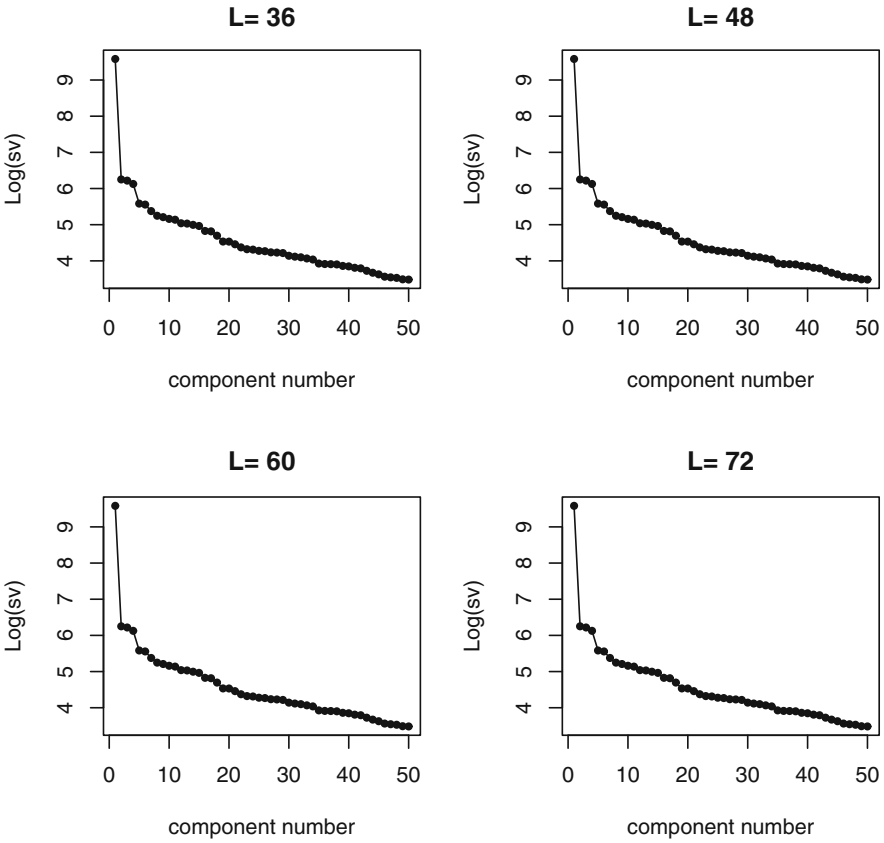


Fig. 12.8 Singular values of the trajectory matrix for meat prices time series for L=36 to 72

in 2023 using various MSSA forecasting algorithms. We also analyzed these time series using univariate SSA. Examining different values of the window lengths by the scree plot of the singular values showed us that $r = 7$ for both time series is appropriate. We used $L = 88, r = 7$ for the first time series and $L = 72, r = 7$ for the second time series. Using these values, we obtained forecasts for 12 months of the year 2023 and computed the RMSE. Table 12.5 presents the results. Notably, the VMSSA-R algorithm and SSA-V outperform other forecasting methods, as indicated by the bold styling, for forecasting the first and second time series, respectively. Comparing the RMSE values with the time series data, we note that errors are consistently below 5%, reflecting the good performance of the forecasting approach. This example confirms that univariate SSA may be superior to multivariate SSA in some cases. However, I must also discuss the complexity of models obtained here. The simplest model is provided by HMSSA-R, and the most complicated model is associated with VMSSA-V.

Table 12.5 RMSE for forecasting 2023 using different MSSA algorithms

Method	Boxed beef cutout	Boneless beef
HMSSA-R	15.69	21.18
HMSSA-V	10.71	18.74
VMSSA-R	10.20	**14.85**
VMSSA-V	18.14	34.46
SSA-R	10.06	25.88
SSA-V	**8.78**	21.08

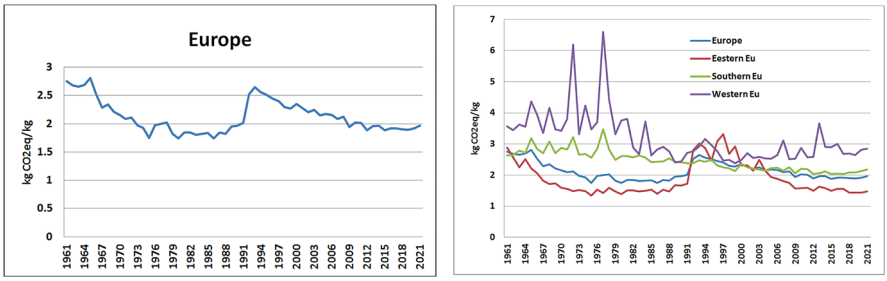


Fig. 12.9 CO2 emission intensity from rice by different segmentation in Europe over the period 1961 to 2021

12.4.4.3 Data Set 2: CO2 Emission Intensity for Rice Commodity in European Countries

The significance of the emissions intensity associated with rice in European nations is a key factor for gauging the ecological footprint of its cultivation, as per reports from the Food and Agriculture Organization (FAO). According to the FAO's findings, there's a noticeable variation in the emissions intensity from rice farming across different European countries, which is shaped by elements like farming methods, land utilization, and energy usage. There has been a deliberate push in recent times to curtail the emissions intensity linked to rice farming across Europe. The FAO's data underscores the strides made by some European nations in cutting down the emissions intensity from rice farming, thereby aiding the environmental health of the agrifood sector. Often, this dip in emissions intensity stems from embracing farming practices that are more eco-friendly, enhancing the efficiency of resource utilization, and deploying technologies designed to curb greenhouse gas emissions.

The left graph in Fig. 12.9 illustrates the emission intensity for rice cultivation in Europe spanning roughly six decades, expressed in kg CO2eq per kg. This graph indicates that the emission intensity typically ranged from 1.5 to 3 kg CO2eq per kg, featuring two periods of diminishing emissions, each approximately 30 years in duration. Notably, there was a discernible shift in 1992, marked by an uptick in emissions, followed by a subsequent reduction. Moving on to a comparative analysis, the right graph in Fig. 12.9 incorporates emission intensity data for rice during the identical timeframe but segmented into three regions: Eastern Europe,

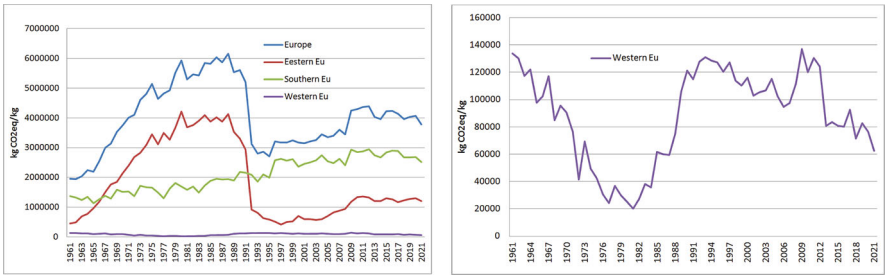


Fig. 12.10 CO2 emission production from rice by different segmentation in Europe over the period 1961 to 2021

Southern Europe, and Western Europe. This comparison reveals inconsistencies at specific intervals. Our computations determine that the aggregate index is derived from a weighted mean of these regional segmentations, with weights assigned in proportion to their respective CO2 emission volumes. Figure 12.10 charts the CO2 emissions from rice production according to these regional divisions and for the entire Eurozone. The graph demonstrates a significant decline in the CO2 emission contribution from Eastern Europe post-1990, which, after a steady rise, appears to have stabilized in recent years. For the Southern European region, the trend indicates a consistent escalation throughout the period. The emission pattern for the Western European zone was ambiguous in the left graph, prompting its exclusive depiction in the right panel for clarity. This isolated representation confirms the presence of fluctuating patterns over time.

Two analytical approaches, SSA and MSSA, were employed to analyze the time series of CO2 emission intensity. Their effectiveness was assessed by comparing the root mean squared errors (RMSE). To ensure the robustness of the findings, expanded window techniques were used, and the assessments were repeated across different forecasting periods. The outcomes are shown in Table 12.6. For a straightforward comparison, the RRMSE (Relative Root Mean Squared Error) was computed by dividing the RMSE value for MSSA by that of SSA, with the results detailed in the final column. The data in the table reveal that MSSA significantly outperforms SSA in terms of accuracy across all predicted time frames.

12.5 Concluding Remark

This chapter provided a concise introduction to Singular Spectrum Analysis (SSA), commencing with a broad overview of various time series types and the diverse methods of time series analysis. This background information emphasized the significance of the topic and the extensive research conducted in this field.

Subsequently, the chapter narrowed its focus to SSA, highlighting its applications and extensions to the multivariate case. The chapter demonstrated the effectiveness of SSA by applying it to different data sets and comparing the results with other

Table 12.6 Comparison of SSA and MSSA in CO2 emission intensity forecast

Horizon	Method	RMSE	RRMSE
1	SSA-R	0.20	0.32
	HMSSA-R	0.06	
2	SSA-R	0.28	0.27
	HMSSA-R	0.07	
3	SSA-R	0.38	0.22
	HMSSA-R	0.08	
5	SSA-R	0.56	0.17
	HMSSA-R	0.09	
10	SSA-R	1.26	0.04
	HMSSA-R	0.05	

methods. It was concluded that, while SSA may not be the optimal model, it offers several advantages that warrant its inclusion in the list of methods used for time series analysis.

In addition to the extensions discussed in this chapter, it is worth noting that researchers have introduced several other extensions of SSA. For instance, Golyandina and Usevich [10] proposed an extension for analyzing images, while Haghbin et al. [14] and Trinka et al. [31] developed extensions for functional time series. Furthermore, Golyandina et al. [12] introduced extensions for complex data and tensors.

For computational needs, the R packages Rssa [13] is a valuable resource, providing the necessary tools for various SSA cases, excluding functional time series. The R package Rfssa [15] has been specifically developed for functional time series analysis. These packages facilitate the application of SSA and its extensions in practical data analysis scenarios.

References

1. Ahn, H., Sun, K., & Kim, K. P. (2022). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials & Continua*, *70*(1), 767–779.
2. Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, *54*(3), 1–33.
3. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.
4. Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. Springer.
5. De Gooijer, J. G. (2017). *Elements of nonlinear time series analysis and forecasting* (Vol. 37). Springer.
6. Danilov, D. (1997). Principal components in time series forecast. *Journal of Computational and Graphical Statistics*, *6*, 112–121.
7. Danilov, D. (1997) The 'Caterpillar' method for time series forecasting. In D. Danilov, & A. Zhigljavsky (Eds.), *Principal components of time series: The 'caterpillar' method* (pp. 73–104). University of St. Petersburg (In Russian).
8. Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. arXiv preprint arXiv:1005.4374.

9. Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. (2001). *Analysis of time series structure: SSA and related techniques*. Chapman & Hall/CRC.
10. Golyandina, N., & K. Usevich (2010). 2D-extension of singular spectrum analysis: Algorithm and elements of theory. In V. Olshevsky, & E.Tyrtyshnikov (Eds.), *Matrix methods: Theory, algorithms and applications* (pp. 449–473). World Scientific Publishing.
11. Golyandina, N., & Zhigljavsky, A. (2013). *Singular spectrum analysis for time series*. Springer.
12. Golyandina, N., Korobeynikov, A., Shlemov, A., & Usevich, K. (2015): Multivariate and 2D extensions of singular spectrum analysis with the Rssa package. *Journal of Statistical Software, 67*(2).
13. Golyandina, N., & Zhigljavsky, A. (2018). *Singular spectrum analysis for time series*. Springer.
14. Haghbin, H., Najibi, S. M., Mahmoudvand, R., Trinka, J., & Maadooliat, M. (2021). Functional singular spectrum analysis. *Stat*, *10*(1), e330.
15. Haghbin, H., Trinka, J., Najibi, S. M., & Maadooliat, M. (2024). Package 'Rfssa'.
16. Hassani, H. (2007). Singular spectrum analysis: Methodology and comparison. *Journal of Data Science*, *5*, 239–257.
17. Hassani, H., Heravi, S., & Zhigljavsky, A. (2009). Forecasting European industrial production with singular spectrum analysis. *International Journal of Forecasting*, *25*(1), 103–118.
18. Hassani , H., Mahmoudvand, R., & Zokaei, M. (2011). Separability and window length in singular spectrum analysis. *Comptes Rendus Mathematique*, *349*, 987–990.
19. Hassani, H., Webster, A., Silva, E. S., & Heravi, S. (2015). Forecasting U.S. tourist arrivals using optimal singular spectrum analysis. *Tourism Management*, *46*, 322–335.
20. Hassani, H., & Mahmoudvand, R. (2018). *Singular spectrum analysis: Using R*. Springer.
21. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.
22. Josu, A., & Javier, G.-E. (2017). Singular Spectrum Analysis for signal extraction in Stochastic Volatility models. *Econometrics and Statistics*, *1*, 85–98.
23. Mahmoudvand, R., Najari, N., & Zokaei, M. (2013). On the parameters for reconstruction and forecasting in the singular spectrum analysis. *Communication in Statistics: Simulations and Computations*, *42*, 860–870.
24. Mahmoudvand, R., & Zokaei, M. (2012). On the singular values of the Hankel matrix with application in singular spectrum analysis. *Chilean Journal of Statistics*, *3*, 43–56.
25. Mastrangelo, C. M., Simpson, J. R., & Montgomery, D. C. (2013). Time series analysis. In S. I. Gass, & M.C. Fu (Eds.), *Encyclopedia of operations research and management science*. Springer.
26. Nekrutkin, V. V. (1999). Approximation spaces and continuation of time series. In S. M. Ermakovand, & Y. N. Kashtanov (Eds.), *Statistical models with applications in econometrics and neibouring fields* (pp. 3–32). University of St. Petersburg (In Russian).
27. Palazzi, R. B., Maçaira, P., Meira, E., & Klotzle, M. C. (2023). Forecasting commodity prices in Brazil through hybrid SSA-complex seasonality models. *Production*, *33*, e20220025.
28. Pollock, D. S. G. (1987). The methods of time-series analysis. *Interdisciplinary Science Reviews*, *12*(2), 128–135.
29. Ribeiro, S. M., & de Castro, C. L. (2021). Missing data in time series: A review of imputation methods and case study. Learning and Nonlinear Models-Revista Da Sociedade Brasileira De Redes Neurais-Special Issue: *Time Series Analysis and Forecasting Using Computational Intelligence*, *19*(2).
30. Shumway, R., & Stoffer, D. (2019). *Time series: A data analysis approach using R*. Chapman and Hall/CRC.
31. Trinka, J., Haghbin, H., Shang, H., & Maadooliat, M. (2023). Functional time series forecasting: Functional singular spectrum analysis approaches. *Stat*, *e621*.
32. Tsay, R. S. (2014). *Analysis of financial time series*. John Wiley & Sons.
33. Tsay, R. S., & Chen, R. (2018). *Nonlinear time series analysis* (Vol. 891). John Wiley & Sons.
34. Wang, J., & Li, X. (2018). A combined neural network model for commodity price forecasting with SSA. *Soft Computing*, *22*(16), 5323–5333.
35. Wei, W. W. S. (2006). *Time series analysis: Univariate and multivariate methods.* Pearson Addison Wesley.