A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models

Galina Andreeva, Raffaella Calabrese, Silvia Angela Osmetti

PII: S0377-2217(15)00718-3
DOI: 10.1016/j.ejor.2015.07.062
Reference: EOR 13152

To appear in: *European Journal of Operational Research*

Received date: 23 January 2014
Revised date: 2 April 2015
Accepted date: 28 July 2015

Please cite this article as: Galina Andreeva, Raffaella Calabrese, Silvia Angela Osmetti, A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models, *European Journal of Operational Research* (2015), doi: 10.1016/j.ejor.2015.07.062

Highlights

- Comparison of insolvency risk predictors between Italy and the UK .
- Application of GEV model to account for low proportions of insolvent companies.
- Application of BGEVA to account for non-linearity between response and predictors.
- Comparison of two methods for treating the missing values.
- BGEVA on WoE method for missing data showed the best predictive accuracy.

A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models

Galina Andreeva[1], Raffaella Calabrese[2] and Silvia Angela Osmetti[3]

[1]Business School, University of Edinburgh

29 Buccleuch Place, Edinburgh EH8 9J, UK

Galina.Andreeva@ed.ac.uk

[2]Essex Business School, University of Essex

Wivenhoe Park, Colchester CO4 3SQ, UK

rcalab@essex.ac.uk

[3]Department of Statistical sciences, Università Cattolica del Sacro Cuore di Milano

Largo Gemelli 1, 20123 Milano, Italy

silvia.osmetti@unicatt.it

Corresponding author: Dr Galina Andreeva

Tel:+44(0)65123293

Fax:+44 (0)131 651 3197

E-mail: Galina.Andreeva@ed.ac.uk

Abstract

This paper presents a cross-country comparison of significant predictors of small business failure between Italy and the UK. Financial measures of profitability, leverage, coverage, liquidity, scale and non-financial information are explored, some commonalities and differences are highlighted. Several models are considered, starting with the logistic regression which is a standard approach in credit risk modelling. Some important improvements are investigated. Generalised Extreme Value (GEV) regression is applied in contrast to the logistic regression in order to produce more conservative estimates of default probability. The assumption of non-linearity is relaxed through application of BGEVA, non-parametric additive model based on the GEV link function. Two methods of handling missing values are compared: multiple imputation and Weights of Evidence (WoE) transformation. The results suggest that the best predictive performance is obtained by BGEVA, thus implying the necessity of taking into account the low volume of defaults and non-linear patterns when modelling SME performance. WoE for the majority of models considered show better prediction as compared to multiple imputation, suggesting that missing values could be informative.

Keywords: Decision support systems, Risk analysis, Credit Scoring, Small and Medium Sized Enterprises, Default prediction.

1 Introduction

Small and Medium Enterprises (SMEs) play a central role in the European Union (EU) economy, as recognised by the Small Business Act of the European Commission in 2008 (http://ec.europa.eu/enterprise/entrepreneurship/docs/sba/SBA_IA). In 2011 SMEs represented 99% of enterprises in Europe, employing more than two thirds of the workforce and contributing 58% of total EU added value. The importance of SMEs varies across the EU. In some countries, e.g. Italy, Spain and Portugal, SMEs have larger shares in employment and added value and higher presence than the EU average. On the contrary, these figures are lower than the EU average in other countries, e.g. the UK, Germany and France.

In this work we compare Italy and the UK since the economies of these countries are different, and it is of interest to explore the differences in predictors of SMEs failures, especially in the aftermath of the "credit crunch". The literature on SME default prediction is limited, in particular in cross-country comparisons, and the main objective of this paper is to fill in this gap. This paper contributes to the existing cross-country research by an initial exploratory investigation of risk predictors using accounting and some non-financial information that are available from public sources.

Several models are considered, starting with the logistic regression which is a standard modelling approach in credit risk research (Thomas et al., 2002). Yet in situations with low numbers of events (defaults), alternative approaches producing more conservative estimates of default probabilities might be of importance. In this paper we concentrate on asymmetric link function and non-linearity between the response and predictors. In real applications the number of defaults is small, therefore, suggesting the asymmetric link function might be beneficial. At the same time the assumption of linearity is not always supported by patterns in the real data. An additional contribution of this paper consists in extending the application of Generalised Extreme Value (GEV) regression that has been proposed for low default portfolios by Calabrese & Osmetti (2013) to two countries. Furthermore, the problem of non-linearity is explored through the application of non-parametric additive model (BGEVA).

The public sources often have incomplete data and this problem is particularly relevant for SMEs. Another objective and contribution of this paper consists in the exploration of two approaches to handling the missing values: multiple imputation and Weights of Evidence transformation which is credit industry's preferred approach.

The rest of the paper is structured as follows: Section 2 provides some background information on the importance of SMEs to the economy and some differences across the two countries. It also summarises previous research on SMEs failure prediction. Section 3 explains the methodology, and Section 4 presents the empirical results, including data description, comparison of predictive accuracy and comparison of statistically significant risk predictors. The final section concludes.

2 Background and literature review

There are some notable differences in characteristics of SMEs in the UK and Italy. In Italy, SMEs form 99.9% of the firms. In 2011 they employed around 81% of the workforce and contributed 68.3% of the Italian added value (EC, 2012a). In terms of the number of SMEs, Italy has the largest SME sector in the EU. With 3.813 million SMEs Italy has almost twice as many as UK (1.649 million). However, the vast majority of Italian SMEs are micro-firms with less than 10 employees. In fact, Italy's share of micro-firms, at 94.6%, exceeds the EU-average (92.2%). Hence, the micro-firms' contribution to employment (46.6% against the EU-average of 29.6%) and added value (29.4% against the EU-average of 21.2%) is high.

On the contrary, the UK economy is characterised by larger companies. In 2011 more than half of the UK added value was produced by large companies that employed less than half (45.7%) of the workforce and constituted only 0.4% of the UK companies. The percentage of micro-firms in the UK (89.5%) is lower than the EU-average (92.2%), and those employ only 20.3% of the workforce and create only 18.5% of the UK added value (EC, 2012b).

Financial crisis has substantially affected SMEs sectors in both countries and recovery has been weaker than in the EU on the whole. The Italian SME sector has reversed to the levels of 2005 (i.e. before the crisis) in terms of the number of firms, employment and value-added creation. In the UK, SMEs have been hit mostly in terms of employment and value-added creation, but the numbers of SMEs are higher than in 2005 and stable. In both countries larger firms suffered less as compared to the smaller ones.

Despite an important role that SMEs play in any economy, academic research into SMEs failure prediction is not very extensive. There are some (albeit not numerous) papers investigating success factors or default risk of SMEs in a specific country, e.g. Altman & Sabato

(2007) for the US, Fantazzini & Figini (2009) for Germany, Sohn & Kim (2013) for South Korea, Martens et al. (2011) for Flanders - to give some examples, yet literature on international comparisons of failure prediction is exceptionally limited.

The survey by Altman & Narayanan (1997) summarised previous research on the performance of companies (not only SMEs) in 22 countries that included both developed and developing economies. Most studies surveyed found measures of profitability, leverage, liquidity, cash flow management, growth, efficiency to be important for bankruptcy prediction, although specific measures used would vary from country to country. A more recent study by Lussier & Halabi (2010) compared performance of SMEs in the USA, Croatia and Chile. Among the variables that were found important for business performance were characteristics of managers (education, experience) and the quality of business functions (record keeping, financial control, planning, staffing).

The most comprehensive study of European SMEs to date is by Michala et al. (2013) where a simple hazard model (Shumway, 2001) has been applied to small businesses from eight European countries, namely Czech Republic, France, Germany, Italy, Poland, Portugal, Spain and the United Kingdom for the period of 2000-2009. The paper has confirmed the significance of indicators of profitability, coverage, leverage and cash flow for bankruptcy prediction in cross-country setting. In addition, some non-financial company characteristics have been investigated and the effect of macroeconomic variables. Pederzoli et al. (2013) modelled credit risk of EU innovative SMEs, but the authors did not make cross-country comparisons.

There were some comparisons between two countries. Ihua (2009) compared the key factors influencing SMEs failure between the UK and Nigeria, and found that economic conditions and infrastructure were more significant in Nigeria, whilst in the UK the key factors were due to internal company characteristics, including management efficiency.

Dietsch & Petey (2004) analysed default probabilities and asset correlations for French and German SMEs. Yet the focus of their analysis was more on comparison of correlations of SMEs as opposed to large corporations, the paper did not look at financial ratios or other predictors of default.

As for SME research in the UK, Lin et al. (2012) compared different definitions of financial distress on a sample from 2001 to 2004 and concluded that although each definition changed the model composition substantially, the most useful variables in distinguishing between distressed and healthy companies, were profit related measures, growth and efficiency ratios. Altman et al. (2010) developed a default prediction model using financial indicators of leverage, profitability, working capital and non-financial information (e.g. age, default events in the past) using the data from 2000 to 2007. They found the non-financial variables provided a notable improvement in predictive performance. Orton et al. (2011) explored the behaviour of the UK SMEs from 2007 to 2010 - through the "credit crunch". They demonstrated that there was a significant degree of stability and accuracy of credit risk models, despite increases in the numbers of SMEs defaults. Similar to Altman et al. (2010) they found company demographics, derogatory events and information about directors to be of significant value.

Regarding the modelling approaches, the overwhelming majority of studies reviewed above used logistic regression. Other models included proportional odds or simple hazard model (Michala et al., 2013; Fantazzini et al., 2009), Bayesian and classic panel models (Fantazzini et al., 2009), random survival forests (Fantazzini & Figini, 2009), Support Vector Machines (Martens et al., 2011).

In Italy Vallini et al. (2009) attempted to model SME defaults on a sample of small firms from 2001- 2005 using profitability, liquidity and leverage ratios. Multiple discriminant analysis was compared to logistic regression, and the latter was found to produce better predictions. Later study by Ciampi & Gordini (2013) applied neural networks to the same dataset and reported their superior performance as compared to algorithms used in the earlier work. Both studies noted that credit scoring models could be built on accounting information, yet predicting default for SMEs was much more difficult as compared to large enterprises, with predictive accuracy decreasing in smaller firms segments.

Calabrese & Osmetti (2013) and Calabrese et al. (2013) applied GEV and BGEVA models to the sample of Italian SMEs from 2006 to 2011 and found superior performance of both models as compared to logistic regression. Variables found significant in predicting default were again measures of profitability, leverage and liquidity.

The current paper extends the existing literature by looking at two countries in comparison (Italy and the UK), by exploring SMEs failure in a more recent time period and by using more comprehensive list of financial measures.

3 Methodology

When constructing a credit scoring model, three common problems are often mentioned: first, low numbers of defaults, second, non-linear relationship between the response and predictors, and third, missing values in predictor variables.

Logistic regression is the most commonly used model for credit scoring applications (e.g., Altman & Sabato, 2007; Becchetti & Sierra, 2002; Lin et al., 2012; Zavgren, 1998). As noted above, the number of defaults in a sample is often very small (e.g., Kiefer, 2010; Lin et al., 2012). King & Zeng (2001) commented on difficulties of obtaining unbiased probability estimates of event occurring in rare events situations. This is due to the fact that the characteristics of defaults (events) are more informative than those of non-defaults. When there is a small number of defaults, there might be insufficient information to produce appropriate estimates of the default probability for values close to 1 (Calabrese & Osmetti, 2013). Wang & Dey (2010) showed that using an asymmetric link function improved the model fit.

In cases of low default portfolios, the conservative (higher) estimates of default probabilities might be preferred, and a flexible asymmetric link function can achieve such higher estimates for defaulters in comparison with logistic regression, as shown in Calabrese & Osmetti (2015).

In order to choose the link function, we consider that defaulters' features are represented by the tail of the response curve for values close to one. Furthermore, the Generalised Extreme Value (GEV) distribution is used in literature (Kotz & Nadarajah, 2000; Falk et al., 2010) to model the tail of a distribution. Therefore, to focus the attention on defaulters' characteristics, Calabrese & Osmetti (2013) propose the quantile function of a GEV random variable as a new link function

$$\frac{[-\ln(PD_i)]^{-\tau}-1}{\tau}=\eta_i=\alpha+\sum_{j=1}^{p}\beta_j x_{ji}, \tag{1}$$

where $\tau \in \Re$ is the tail parameter. As discussed, for instance, in Calabrese & Osmetti (2013), depending on the value of τ, several special cases can be recovered; e.g., when $\tau \to 0$ the GEV random variable follows a Gumbel distribution and its cumulative distribution is the log-log function (Agresti, 2002). In this way, Calabrese & Osmetti (2013) propose the *GEV regression model*.

Second, the logistic and the GEV (1) models assume a linear relationship between the explanatory variables and the response η_i. These models can mask possibly interesting non-linear patterns which can help improve our understanding of the underlying covariate-response relationships and perhaps improve the prediction accuracy of the scoring model as well (Berg, 2007; Calabrese et al., 2013; Chuang & Lin, 2009; Gestel et al., 2005; Huang et al., 2006; Lee & Chen, 2005; Lin et al., 2012; Ong et al., 2005). Therefore Calabrese et al. (2013) propose the BGEVA model, an extension of the GEV model based on penalized regression splines to flexibly determine covariate effects from the data.

In the GEV model, the right part of equation (1) is changed to obtain an additive model given by

$$\frac{[-\ln(PD_i)]^{-\tau}-1}{\tau}=\alpha+\sum_{j=1}^{p}\beta_j s(x_{ji}), \tag{2}$$

where the $s_j(x_{ij})$ are unknown one-dimensional smooth functions of the continuous covariates x_{ji}.

The smooth functions $s(x_{ij})$ in the model are approximated by a linear combination of K_j known (e.g., cubic or thin plate regression) spline bases $b_k(x_{ji})$ and unknown regression parameters, γ_{jk} (Wood, 2006; Calabrese et al., 2013):

$$s_j(x_{ji}) = \sum_{k=1}^{K_j} \gamma_{jk} b_k(x_{ji}).$$

Calculating $b_k(x_{ji})$ for k and each observation point gives K_j curves with different degrees of complexity which multiplied by some real valued parameters γ_{jk} and then summed to give an estimated curve for the smooth component (Ruppert et al., 2003). Replacing in model (2) the smooth terms with their regression spline expressions yields essentially a classic parametric model. Estimating the β_j parameters and the smooth functions $s(x_{ij})$ we can predict the default probabilities by using the inverse of the equation (2). The smooth functions show the existence of possible non-linear relationships between the response variable and the predictors and allow us to improve on the prediction results obtained using classic alternatives. The model is implemented in the `R` package *bgeva* (Marra et al., 2013) available for download from CRAN.

Third, SMEs may not provide full details of their financial statements (Sohn & Kim, 2013; Ciampi & Gordini, 2013), for this reason missing values could be a problem for scoring models for SMEs (Lin et al., 2012; Ciampi & Gordini, 2013). In the literature, missing values are classified into three types: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR). The missing values are MCAR if the probability that any variable is missing cannot depend on any other variable in the model of interest or on the potentially missing values themselves. If we have a single variable Z with missing data and a set of variables which is always observed X, the MCAR assumption can then be expressed by $P(I_z = 1|X, Z) = P(I_z = 1)$ where I is a dummy variable having a value of 1 if Z is missing and 0 if Z is observed. Therefore, the probability that Z is missing depends neither on the observed variables X nor on the possibly missing values of Z itself.

If the probability that Z is missing may depend on X, but it does depend on Z itself $P(I_z = 1|X, Z) = P(I_z = 1|X)$, the MAR assumption is satisfied. This means that MCAR is a special case of MAR. If the data are MAR, it is possible to get optimal estimates of parameters without directly modeling the missing data mechanism since the missing-data mechanism is ignorable. Unfortunately the MAR assumption is not testable. Finally, if the MAR assumption is violated, missing data are said to be NMAR.

There are several methods for handling missing values. The first is to delete cases with any missing data on the variables of interest. This method often deletes a large fraction of the sample and it is particularly suitable if the data are MCAR. When the data are MAR, this procedure may introduce bias into parameter estimates, so the use of a different method is preferable. The second method is to impute values for the missing covariates and carry out the analysis as if the imputed values were observed data. A wide variety of methods falls under the general heading of imputation, for example imputations based on the mean, on the linear regression or on the maximum likelihood and EM algorithm (see Rubin (1976, 1977, 1987)).

One of the widely used approach in the latter method is multiple imputation, which was proposed by Rubin (1987) and described in detail by Graham (2012). Multiple imputation can be described as a three-step process. First, in order to capture the uncertainty in the estimates of the missing values, more sets of plausible values for missing observations are created. Each of these sets of plausible values can be used to 'fill-in' the missing values and create a 'completed' dataset. Second, each of these datasets can be analysed using complete-data methods. Finally, the results are combined, which allows the uncertainty regarding the imputation to be taken into account. The multiple imputation requires that the missing values are MAR. The advantage of these methods is that it can be applied to any type of data and it is implemented in the conventional software. Moreover, it has optimal statistical properties (see Rubin (1987); Graham (2012)).

In this paper we apply a multiple imputation based on an MCMC algorithm known as fully conditional specification (Graham, 2012). The basic idea is to impute incomplete variables one at time by linear regression, using the filled-in variable from one step as a predictor in all subsequent steps. We have chosen this particular method since Florez-Lopez (2010) showed in application to credit scoring that it is superior to other methods of handling missing values.

Another approach to cope with missing values is based on so-called coarse-classification (Thomas et al., 2002). This procedure consists in dividing the values of a numeric predictor into categories or classes. Normally there are 10-20 fine classes initially produced for the range of ordered values from minimum to maximum. In this paper we divide the numeric predictors into 10 classes of approximately the same size (maintaining exactly the same size is not possible because of the varying numbers of missing values for different variables).

For each fine class a proportion of defaults (or bad accounts or simply Bads) is calculated, and adjacent categories can be further grouped together into coarse classes, if the default rates are sufficiently close. Missing values are entered as a separate category. Categories can be entered into the model as binary dummies or alternatively are transformed into Weights of Evidence (WoE):

$$WoE_i = ln\left[\frac{b_i/g_i}{B/G}\right] = ln\left(\frac{b_i\,G}{g_i\,B}\right), \tag{3}$$

where b_i is the number of bads (defaults) in category i of a variable, g_i is the number of goods (non-defaults) in category i, B is the total number of Bads, G is the total number of Goods in the sample.

The term (WoE) goes back to early days of computer science and information theory and is defined by Good (1950) as the weight of evidence (or degree of corroboration) in favour of a hypothesis H given by evidence (or information or an experiment outcome) E:

$$WoE = ln\left[\frac{P(H|E)/(1-P(H|E))}{P(H)/(1-P(H))}\right]. \tag{4}$$

Equation (3) above is a generalisation of Equation (4). It has extensively been used in early classification algorithms and specifically in Naive Bayes classifier, please see Good (1985); Greiff (1999); Hand & Adams (2000); Hand et al. (2001). WoE approach can be criticized on the grounds of imposing the ordering of categories observed for each predictor taken separately and not allowing for interactions between predictors (Thomas, 2009). There may also be a concern about using the dependent variable in transforming a predictor. Despite its limitations this transformation is widely used in practice (Anderson, 2007; Baesens, 2014; Siddiqi, 2006; Thomas, 2000, 2009). An alternative to WoE approach consists in partitioning the variables and then turning k partitions into k-1 dummy (0/1) variables. This approach does not impose any ordering or dependency, but has a disadvantage of producing a lot of variables (Thomas, 2000, 2009). Banasik et al. (2003) compared WoE and dummy variable approaches and found them similar. Following the latter paper and also Banasik & Crook (2007); Bijak & Thomas (2012); Lin et al. (2012); Malik & Thomas (2010); Orton et al. (2015), and the wide-spread industry practice, we use WoE in this paper.

Given the fact that logistic regression is the most commonly used approach in credit scoring (Thomas et al., 2002), WoE is appealing since this transformation produces log odds measures (same scale as logistic regression). Furthermore, log-odds of each category are compared to that of the whole sample: positive values would indicate riskier classes and negative values - more creditworthy customers.

We use this approach as the benchmark to compare the performance of alternative methods to cope with missing values (multiple imputation) and non-linearity (BGEVA model).

4 Empirical Analysis

4.1 Data description

The empirical analysis is based on explanatory variables from 2010 to predict the default in 2011 for $39,785$ UK SMEs and $154,934$ Italian SMEs. The data are from AMADEUS-Bureau van Dijk (BvD), a database of comparable financial and business information on Europe's public and private companies. The time horizon considered here is of extreme interest as it includes the European sovereign debt crisis of 2011. In summer 2011 interest rates on Italian national debt went out of control.

The definition of SME by the European Commission is adopted. That is, a business must have an annual turnover of less than 50 million of Euro, a balance sheet total less than 43 million of Euro and the number of employees should not exceed 250 (http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index.htm). Furthermore, the number of subsisdiaries is capped at 6, in accordance with Lu & Beamish (2001), and the number of directors is 10 maximum, consistent with Gabrielsson (2007); Michala et al. (2013).

In this work, we consider a default to have occurred when a specific SME enters a bankruptcy or a liquidation procedure. Moreover, a SME is classified as default also if it is active and it has not paid a debt (classified as default of payment by BvD) or it is in administration or receivership or under a scheme of arrangement (defined as insolvency proceedings by BvD). On the contrary, non-defaulters include active and dormant SMEs (only 29 for both samples). A dormant company is still registered, but has no significant activity (and no significant accounting transactions during the accounting period). Consistent with previous studies (Altman & Sabato, 2007; Altman et al., 2010; Pederzoli et al., 2013) we exclude dissolved firms that no longer exist as a legal entity, but the reason for *dissolution* is not specified. This is in line with the objective of this paper that models the probability of going bankrupt using publicly available information. Dissolved category comprises SMEs that may not necessarily experience financial difficulties, they may stop trading because the owner retires or for similar reasons. *The descriptive statistcs for dissolved category is shown in Table 8 and Table 9 in the Appendix.* Future research can investigate dissolved as a separate category.

The use of the common database has ensured the availability of the common set of variables measured in the same way for both countries. We used financial ratios that have been found important in previous research on SMEs (Altman & Sabato, 2007; Lin et al., 2012; Michala et al., 2013). Adopting the classification of variables suggested in Altman & Sabato (2007) the variables in this research covered all five major groups usually used:

- Leverage (e.g. Gearing, Solvency ratio);
- Liquidity (e.g. Current ratio, Liquidity ratio, Shareholder liquidity ratio);

- Profitability (e.g. EBITDA margin, Profit margin, ROCE, ROE);
- Coverage (e.g. Interest cover);
- Activity /Scale/Size (e.g. Total assets, Shareholder funds, No of employees, No of directors, No of subsidiaries).

Following Michala et al. (2013) who found cash flow management significant in predicting default, we also include cash flow based measures (e.g. Cash flow, Cash flow / Operating revenue). The variables have been checked for linear dependence, and highly collinear ones have not been used in the analysis. Table 1 presents short and full names of the variables initially considered and some descriptive statistics on the training sample.

Table 1 around here.

The SMEs in the UK sample are larger as compared to Italian SMEs in terms of Total assets, Operating revenue, No of employees, No of directors. This is consistent with the EU statistics reported in Sections 1-2. The summary statistics for Age and No of subsidiaries are similar for the two countries. The UK businesses have higher liabilities, but profitability is also higher. The Italian companies show better Cash flow and lower debt. Despite using the common source of the data, the percentages of missing values are different across the countries. For Italy, the variable with the highest number of missing is Cash flow / Operating revenue, with 19.5% missing. For the UK, the problem is much more acute, the highest percentage of missing is 59.2% for ROCE. This has an effect on the results, depending on how missing values have been treated, as can be seen from Tables 2 and Table 3 that show the variables that are significant at 10% level or lower across the models.

Table 2 around here

Table 3 around here

4.2 Predictive accuracy

To avoid sample dependency, the predictive accuracy for the models was tested on control samples, i.e. we used out-of-sample tests. For each country the whole dataset was split into training (70%) and control (30%) samples using a stratified random sampling with stratification on default indicator. Measures of predictive accuracy used include mean absolute error (MAE), mean square error (MSE) and Area under the ROC curve (AUC). MAE and MSE are standard measures of predictive accuracy in forecasting studies. Obviously, scoring models with lower MSE and MAE should forecast defaults and non-defaults more accurately. For a bank it is much more costly to classify an SME as a non-defaulter when it is a defaulter than the opposite. If a defaulter is classified as a non-defaulter, then it will be accepted for credit, which will subsequently be lost (in part or as a whole). Yet when a non-defaulter is classified as a defaulter, it is only a lost opportunity. Therefore, in this study MSE and MAE are reported for defaults only and they are denoted by MSE^+ and MAE^+. AUC is the most popular measure of model performance in credit scoring (Thomas et al., 2002) that summarises the ability of the model to rank-order the risk correctly over the whole range of predicted PDs. Higher value indicate better performance.

Table 4 around here

Table 5 around here

Tables 4 and 5 summarise the results[1] for the UK and Italian models for imputed and Weights of Evidence (WoE) data.

Considering WoE approach on the UK data, the GEV model shows better performance on

[1]To obtain these results we use SPSS for imputed missing values and the package "bgeva" of R-program.

MAE and AUC than the logistic model, although the latter has lower MSE (Table 4). Moreover, by applying the non-parametric model (BGEVA) the performance and MAE$^+$ improves further, and on MSE$^+$ it becomes the same as for additive logistic. This fact justifies the use of a non-parametric credit scoring model that can capture non-linear relationships between the accounting characteristics of SMEs and response.

As for imputed values on the UK data, the best MAE$^+$ and MSE$^+$ are for BGEVA, whilst the best AUC is shared between BGEVA and additive logistic model. It can be argued that the improvement provided by additive models over GEV and logistic on WoE is modest. This is not surprising, since one of the objectives of WoE and coarse-classification is to cope with non-linearities (see, e.g. Thomas (2009)). Still it appears there is some benefit from applying the semi-parametric approach, albeit it is less pronounced as compared to improvement of BGEVA over GEV on imputed data. This further emphasises the advantage of BGEVA in forecasting defaults in low default portfolios that performs well on both methods of treating the missing values.

Considering WoE approach on Italian data (Table 5), we observe results similar to the UK models. BGEVA has the best MAE$^+$ and MSE$^+$, whilst additive logistic produces slightly higher AUC, but the difference is negligible. For Italian imputed values the results are mixed. The additive logistic model shows the lowest values of the MAE$^+$ and MSE$^+$, whilst the GEV and logistic models show higher values of the AUC.

The comparison of the predictive accuracy between the countries should be interpreted with caution due to the different sample sizes, different proportions of missing values and different number of significant variables (as discussed in the next section). Since the UK sample size is smaller than the Italian one and the percentage of UK missing values is higher than for Italy (see Table 1), one can expect a decrease in the predictive accuracy. However, for completeness it could be stated that all models for Italy have better performance than the UK models. Moreover, the Italian best model (BGEVA) has also a lowest MAE$^+$.

It should also be noted that WoE coding provides better performance as compared to

Imputation with the only exception of MAE$^+$ of BGEVA for the UK.

In conclusion, the empirical results confirm that the BGEVA model performs well for SMEs default forecasting for both countries. This can be attributed to the fact that the linearity assumption is not supported by the data of both countries, as will be discussed in the next section.

4.3 Comparison of risk predictors between Italian and UK SMEs

There are differences between the countries in terms of significant variables and their number depending on the model/approach used. Whilst logistic regression for both countries and GEV model for Italy show the same number of variables irrespective of imputation or WoE, there are differences in model composition even in these cases. For example, in logistic regression for the UK - Cash flow, Interest cover and Operating revenue are significant with WoE coding, but not with Imputation; yet with Imputation the following variables become significant: Profit margin, Shareholder funds and Total assets. For the rest of models the numbers of significant variables differ with the extreme cases of GEV and BGEVA for the UK, where WoE coding increases the number of significant variables from 11 to 20. This may be interpreted as suggesting that at least for some variables values cannot be assumed to be missing at random, therefore WoE increase the number of significant variables.

Only two variables consistently appear across all 16 models for the two countries: No of directors and Solvency ratio (Tables 2 and 3). No of subsidiaries appear in all models, but one. Profit margin and Shareholder funds enter 14 models. Other frequent variables that are significant at 10 per cent level or lower across all 16 models for the two countries are Liquidity ratio (13), Age (12), EBITDA margin (12), No of employees (12), Operating revenue (12), Cash flow / Operating revenue (10), Total assets (10), ROE (10). When looking at most frequent significant variables for each country separately (e.g. common variables that are in more than half of the models for each country) these include No of directors, Solvency ratio, No of subsidiaries, Profit margin, Shareholder funds and Liquidity ratio. This confirms the results from previous research that suggests measures of profitability, leverage and liquidity are important (Altman & Narayanan, 1997; Altman et al., 2010; Michala et al., 2013). Shareholder funds can be interpreted as the interest the shareholders have in the company, and also the ability of the company to raise funds for growth/expansion. Solvency ratio emphasis the importance of the proportion of Shareholder funds in the assets of the company. No of directors and No of subsidiaries may be interpreted as proxies for company size and the scale of the activity, with No of directors also acting as a crude proxy for quality of management (assuming more directors would mean better management).

Table 6 around here

Table 7 around here

Despite the commonality reported above, there are some interesting differences between the countries. The most notable one is the fact that Gearing is significant in all UK models, whilst not being significant in Italy at all. This suggests the importance of the firm's ability to pay both long-term debt and short-term one in the UK. For Italy measures of profitability are relatively more prominent: EBITDA margin and ROE appear in almost all Italian models, in addition to Profit margin which is common to both countries. Age and No of employees are twice more frequent in the UK models. Age has been previously found important in Altman & Sabato (2007). No of employees indicates the size of the company or its scale. Financial scale for Italy is most frequently represented by Operating revenue, which appears in all Italian models, but only in half of the UK ones. Cash flow/ Operating revenue is also present in all Italian models.

As an example of more detailed cross-country differences, consider the estimates of BGEVA model on imputed values presented in Tables 6. The interpretation of WoE is less straightforward since it requires the information on category boundaries and WoE values. This information and details of other models are available on request. Financial measures common to both countries include ratios of profitability (Profit margin), leverage (Solvency ratio), liquidity (Liquidity ratio) and scale (Shareholder funds, Total assets). In addition, there are common non-financial variables across the two countries: Age, No of directors, No of employees, No of subsidiaries. This fact emphasises the value of non-financial information in modelling SMEs and confirms some previous research (Altman et al., 2010).

Tables 6 and 7 report the estimation results of the parametric and non-parametric components of the BGEVA model for the two countries and for multiple imputation. Some of the covariate effects are reported in the parametric part of the BGEVA model since their smooth function estimates were linear. Explanatory variables significant deviations from the linearity assumption are reported in the smooth terms part. The variables show different degrees of non-linearity (Edf). The parameter Edf (degrees of freedom) in Tables 6 and 7 controls the smoothness of the curve. The variables with Edfs equal to 1 show linear smooth function so they are reported in the parametric part. The estimated smooth parameters that exhibit Edfs considerably greater than 1 are reported in smooth terms part. Larger Edf allows a very flexible curve, e.g., a curve that can have multiple local maxima and minima. The values of degrees of freedom are estimated from the data. The most interesting smooth terms are displayed in the Figures 1 and 2. In line with the interpretation for the parametric components, if the estimated smooth function of a covariate is decreasing then the estimated PD decreases when the explanatory variable increases, and vice versa.

Figure 1 around here

Figure 2 around here

There is some commonality between the countries with Liquidity ratio and Age being non-linear for both countries. No of directors and Total assets exhibit non-linear relationship with the response for the UK, but not for Italy, on the contrary, Cash Flow and No of employees show non-linear patterns for Italy only.

Consider Liquidity ratio that shows a non-linear relationship for both countries (Figures 1 and 2). For Italy when this variables increases, the PD decreases (although in a non-linear way), in accordance with the expectations and prior research by Pederzoli et al. (2013). Yet for the UK the relationship is more complex. Up to 30 and from 75 the relationship of this covariate to PD is negative (as expected). However, in the middle section it is the opposite: increasing values of Liquidity ratio signal increasing chances of default. This may be related to difficulties in getting credit for SMEs, if Current Liabilities in denominator are decreasing.

Previous research summarised in Section 2 did not use exactly the same ratio, yet Altman et al. (2010) report a negative relationship between a similar variable (Current ratio) and the PD. It should be noted though, that the authors did not comment on potential non-linearity. For German SMEs Fantazzini et al. (2009) and Figini & Giudici (2011) observed a counter-intuitive sign for Liabilities ratio and explained it by the fact that many small business owners cover their debts from external sources.

Examples of variables that show non-linear relationships and are not common for the two countries are Total Assets for the UK and No of Employees for Italy, both can be interpreted as proxies for SME size. From Figure 2 looking at Total assets we can deduce that the UK small and micro enterprises show higher default risk, in line with Fantazzini et al. (2009) for German SMEs. Then for companies with Total assets higher than 20 million euros, when this variable increases the PD decreases. Altman et al. (2010) also noted the non-linear nature of Total assets. Finally, from the plot for Number of Employees (Figure 1) Italian small and micro enterprises have higher PD when the number of employees increases. For medium enterprises this relationship becomes negative, although the confidence intervals are wide. These results highlight some interesting patterns observed from the data, yet further research would be beneficial in order to fully understand the implied relations.

5 Conclusions and extensions

This paper has compared predictors of SMEs insolvencies across the UK and Italy, using publicly available information from 2010 to model the company status in 2011. The choice of the time period after the credit crisis makes this comparison particularly relevant, due to different economic situations in the two countries. Whilst Italy was experiencing high interest rates for its national debt, that was not the case in the UK despite the latter showing low economic growth. There are also differences across the two countries in the relative importance that SMEs play in the two economies, as discussed in Section 2. Despite these differences, there were some financial measures significant in predicting insolvency. These included measures of profitability, leverage, liquidity and scale. In addition, there was some commonality in non-financial measures, thus highlighting the importance of soft information for analysis of SME performance. As for the differences, profitability measures are significant more frequently for Italy, whilst for the UK Gearing is a significant predictor, not featuring in Italian models.

A number of different modelling approaches have been explored in order to improve predictive accuracy. Generalised Extreme Value (GEV) regression with asymmetric link function was applied in comparison to the logistic regression, which is a standard approach in credit risk modelling. The assumption of non-linearity was relaxed through application of BGEVA, non-parametric additive model based on the GEV link function. In addition, two methods of handling missing values were compared: multiple imputation and Weights of Evidence (WoE) transformation. The results suggest that the best predictive performance is obtained by BGEVA, thus implying the necessity of taking into account the volume of defaults and non-linear patterns when modelling SME insolvencies. WoE generally showed better prediction as compared to imputation, suggesting that missing values could be informative.

This study presents an initial attempt to understand the cross-country drivers of SMEs insolvencies, and is exploratory in the general approach adopted. Further extensions could include exploration of additional countries and additional variables, in particular, of non-financial nature, but this depends on the data availability. Causal relations through structural equation models can be investigated. On the practical side, it would be of interest to consider predictors significant to both countries and construct a generic model with the objective of comparing it to a country-specific model. Finally, different groups of SMEs that go out of business can be explored, e.g. dissolved.

In this paper we perform a cross-sectional analysis. As a possible direction for future research, we are planning to extend the BGEVA model to a panel data setting, and compare the performance of SMEs for the two countries across time.

Acknowledgements

The Authors gratefully acknowledge the help of Armando Benincasa and Carlotta Prete at Bureau Van Dijk in providing the data. They also acknowledge comments and suggestions by the anonymous referees, that have helped improving the paper. Dr Andreeva is grateful to The Royal Society of Edinburgh for financial support.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York.

Altman, E. & Narayanan, P. (1997). An international survey of business failure classification models. *Financial Markets, Institutions and Instruments*, 6, 1–57.

Altman, E. & Sabato, G. (2007). Modeling credit risk for smes: Evidence from the us market. *ABACUS*, 43(3), 332–357.

Altman, E., Sabato, G., & Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk*, 6, 1–33.

Anderson, R. (2007). *The Credit Scoring Toolkit*. Oxford University Press: Oxford.

Baesens, B. (2014). *Analytics in a Big Data World*. Wiley, Wiley and SAS Business Series.

Banasik, J. L. & Crook, J. N. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183(3), 1582–1594.

Banasik, J. L., kin, J. N., & Thomas, L. C. (2003). Sample selection bias in credit scoring models. *Journal of Operational Research Society*, 54(8), 822–832.

Becchetti, L. & Sierra, J. (2002). Bankruptcy risk and productive efficiency in manufacturing firms. *Journal of Banking and Finance*, 27, 2099–2120.

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, 23, 129–143.

Bijak, K. & Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, 39(3), 2433–2442.

Calabrese, R., Marra, G., & Osmetti, S. A. (2013). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *ArXiv*, 1307.6081, 1–28.

Calabrese, R. & Osmetti, S. (2015). Improving forecast of binary rare events data: A gam-based approach. *Journal of Forecasting*, Forthcoming.

Calabrese, R. & Osmetti, S. A. (2013). Modelling sme loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics*, 40(6), 1172–1188.

Chuang, C. L. & Lin, R. H. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, 36, 1685–1694.

Ciampi, F. & Gordini, N. (2013). Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of italian small enterprises. *Journal of Small Business Management*, 51(1), 23–45.

Dietsch, M. & Petey, J. (2004). Should sme exposure be treated as retail or as corporate exposures? a comparative analysis of default probabilities and asset correlation in french and german smes. *Journal of Banking and Finance*, 28, 773–788.

EC (2012a). Sba fact sheet 2012 for italy. *Enterprise and Industry working paper.*

EC (2012b). Sba fact sheet 2012 for uk. *Enterprise and Industry working paper.*

Falk, M., Haler, J., & Reiss, R. (2010). *Laws of Small Numbers: Extremes and Rare Events.* Springer.

Fantazzini, D. & Figini, S. (2009). Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, 11, 29–45.

Fantazzini, D., Figini, S., Giuli, E. D., & Giudici, P. (2009). Enhanced credit default models for heterogeneous sme segments. *Journal of Financial Transformation*, 25(1), 31–39.

Figini, S. & Giudici, P. (2011). Statistical merging of rating models. *Journal of the Operational Research Society*, 62(6), 1067–1074.

Gabrielsson, J. (2007). Boards of directors and entrepreneurial posture in medium-size companies. putting the board demography approach to a test. *International Small Business Journal*, 25, 511–537.

Gestel, T. V., Baesens, B., Dijcke, P. V., Suykens, J. A. K., Garcia, J., & Alderweireld, T. (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines. *Journal of Credit Risk*, 1(4), 31–60.

Good, I. J. (1950). *Probabllity and the Weighing of Evidence.* London: Charles Griftin; New York: Hafners.

Good, I. J. (1985). Weight of evidence: a brief survey. *Bayesian Statistics*, 27, 249–270.

Graham, J. W. (2012). *Missing Data: Analysis and Design.* Springer, New York.

Greiff, W. R. (1999). *Maximum Entropy, Weight of Evidence and Information Retrieval.* PhD Thesis: University of Massachusetts Amherst.

Hand, D. J. & Adams, N. M. (2000). Defining attributes for scorecard construction in credit scoring. *Journal of Applied Statistics*, 27(5), 527–540.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Huang, J. J., Tzeng, J. H., & Ong, C. S. (2006). Two-stage genetic programming (2sgp) for the credit scoring model. *Applied Mathematics and Computation*, 174, 1039–1053.

Ihua, U. B. (2009). Smes key failure-factors: a comparison between the united kingdom and nigeria. *Journal of Social Science*, 18(3), 199–207.

Kiefer, N. M. (2010). Journal of business and economic statistics. *Journal of Business Finance & Accounting*, 28(2), 320–328.

King, G. & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9, 321–354.

Kotz, S. & Nadarajah, S. (2000). *Extreme Value Distributions. Theory and Applications*. Imperial College Press, London.

Lee, T. S. & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 28, 743–752.

Lin, S. M., Ansell, J., & Andreeva, G. (2012). Predicting default of a small business using different definitions of financial distress. *Journal of the Operational Research Society*, 63, 539–548.

Lu, J. W. & Beamish, P. W. (2001). The internationalization and performance of smes. *Strategic Management Journal*, 22, 565–586.

Lussier, R. & Halabi, C. (2010). A three-country comparison of the business success versus failure prediction model. *Journal of Small Business Management*, 48(3), 360–377.

Malik, M. & Thomas, L. C. (2010). Modelling credit risk of portfolios of consumer loans. *Journal of the Operational Research Society*, 61(3), 411–420.

Marra, G., Calabrese, R., & Osmetti, S. A. (2013). *bgeva: Binary Generalized Extreme Value Additive Models*. R package version 0.2.

Martens, D., Vanhoutte, C., Winne, S. D., Baesens, B., Sels, L., & Mues, C. (2011). Identifying financially successful start-up profiles with data mining. *Expert Systems with Applications*, 38, 5794–5800.

Michala, D., Grammatikosa, T., & Filipea, S. F. (2013). Forecasting distress in european sme portfolios. *LSF Research Working Paper Series*, (13-02).

Ong, C. S., Huanga, J. J., & Tzeng, G. H. (2005). Building credit scoring models using genetic programming. expert systems with applications. *Expert Systems with Applications*, 29, 41–47.

Orton, P., Ansell, J., & Andreeva, G. (2011). Recent developments in commercial scoring. *Credit Scoring & Credit Control XII Conference*, Edinburgh, UK.

Orton, P., Ansell, J., & Andreeva, G. (2015). Exploring the performance of small- and medium-sized enterprises through the credit crunch. *Journal of the Operational Research Society*, 66, 657–663.

Pederzoli, C., Thoma, G., & Torriccelli, C. (2013). Modelling credit risk for innovative smes: the role of innovation measures. *Journal of Financial Services Research*, 44, 111–129.

Rubin, D. (1976). Inference and missing data. *Biometrika*, (63), 581592.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538543.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, London.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, 74(1), 101124.

Siddiqi, N. (2006). *Credit risk scorecards : developing and implementing intelligent credit scoring*. Wiley: New Jersey.

Sohn, S. Y. & Kim, Y. S. (2013). Behavioral credit scoring model for technology-based firms that considers uncertain financial ratios obtained from relationship banking. *Small Business Economics*, 41(4), 931–943.

Thomas, L., Edelman, D., & Crook, J. C. (2002). *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149–172.

Thomas, L. C. (2009). *Consumer Credit Models*. Oxford University Press: Oxford.

Vallini, C. F., Ciampi, F., Gordini, N., & Benvenuti, M. (2009). Are credit scoring modeld able to predict small enterprise default? statistical evidence from italian small enterprises. *International Journal of Business & Economics*, 8(1), 3–18.

Wang, X. & Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *The Annals of Applied Statistics*, 4(4), 2000–2023.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall, Boca Raton.

Zavgren, C. (1998). The prediction of corporate failure: the state of the art. *Journal of Accounting Literature*, 2, 1–37.

6 Appendix

Table 8 around here

Table 9 around here

Variables		Italy, n=106967			UK, n=27132		
Short name	*Description*	*% Missing*	*Mean*	*Std Dev*	*% Missing*	*Mean*	*Std Dev*
Age	Age of the company, months	0.2	207.55	156.072	0.2	206.59	192.983
Capital	Capital, th EUR	0.2	203.46	816.490	2.0	331.96	2689.960
Cash_flow	Cash flow, th EUR	2.4	109.23	617.864	21.4	47.69	40224.523
Cash_flow_oprev	Cash flow / Operating revenue,%	19.5	7.00	8.250	39.4	11.98	14.668
Current_liab	Current liabilities, th EUR	0.2	1606.02	3045.892	1.6	2368.52	6464.733
Current_ratio	Current assets/Current liabilities,%	0.3	1.76	2.767	3.7	4.75	10.125
EBITDA_Margin	EBITDA/Operating revenue, %	3.5	6.79	14.564	23.2	8.64	20.992
Gearing	(Long term liab. + Short term loans)/ Shareholders funds, %	16.5	188.34	220.371	20.7	76.62	149.386
Interest_cover	P(L) before interest/ Interest paid, %	8.2	26.13	96.408	57.6	39.52	119.958
Liquidity_ratio	(Current assets - Stock)/Current liab.	0.3	1.36	2.392	5.0	4.53	10.119
Loans	Loans, th EUR	0.2	437.09	1300.625	3.3	1025.80	4005.364
Net_income	Net income, th EUR	0.2	11.46	635.865	3.2	109.074	4357.018
No_directors	Number of current directors/managers	0.0	1.33	1.810	0.0	4.74	2.555
No_employees	Number of employees	0.2	13.52	22.239	2.2	37.19	47.120
No_subsidiaries	No of recorded subsidiaries	0.0	0.41	0.872	0.0	0.47	0.977
Noncurrent_liab	Non-current liabilities, th EUR	0.2	613.73	1602.776	1.7	1001.87	4763.591
Op_rev	Operating revenue (Turnover), th EUR	0.2	2998.49	5457.546	1.7	6152.03	8545.211
PL_beforetax	Profit (Loss) before tax, th EUR	0.2	56.38	606.305	3.0	179.731	4391.610
Profit_employee	Profit per employee, th EUR	2.1	10.35	60.350	7.0	28.41	182.253
Profit_margin	P(L) before tax/ Operating revenue, %	1.9	1.03	14.395	7.8	7.66	25.868
ROCE	P(L) before tax/ (Total assets - Cur. liab.)	6.5	10.77	57.921	59.2	19.08	81.989
ROE	P(L) before tax/ Shareholder funds, %	7.7	13.36	97.496	18.0	25.85	109.174
Shareh_liquidity_ratio	Shareholders funds/ Long term liab., %	2.2	7.46	37.293	52.0	36.03	105.713
Sharehold_funds	Shareholders funds, th EUR	0.2	865.00	2381.846	1.6	1495.01	7183.142
Solvency_ratio	Shareholders funds/Total assets, %	1.1	23.35	24.527	6.3	45.43	38.256
Tot_assets	Total assets, th EUR	0.2	3084.70	5422.522	1.8	4860.58	6705'262

Table 1: Descriptive statistics for training samples

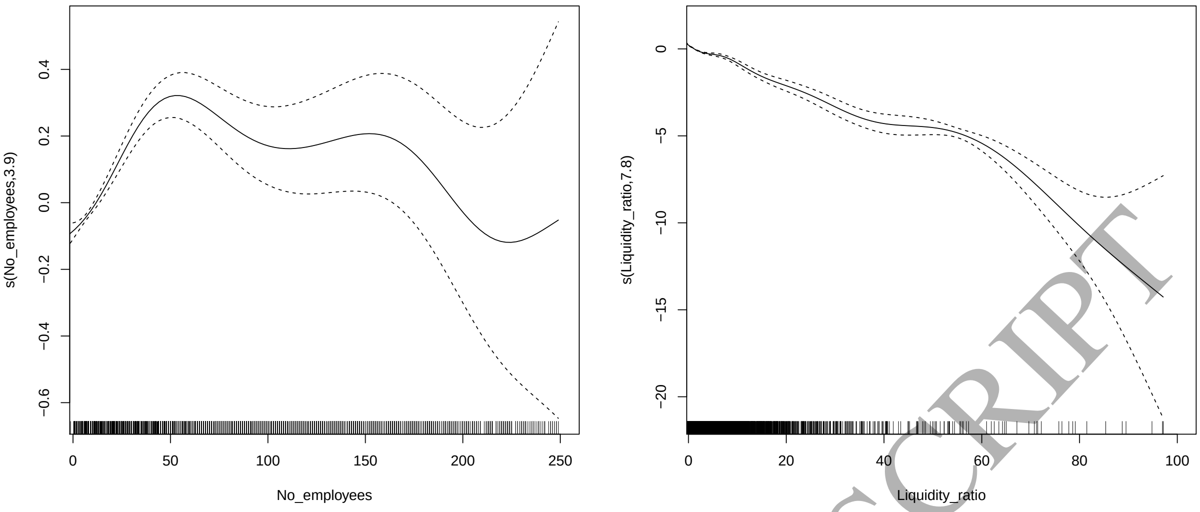

Figure 1: Smooth component estimates of the 2 (out of 4) continuous variables that exhibit a non-linear pattern. These were obtained from applying the BGEVA model on the Italian SME data. Results are on the scale of the predictor. The plot show the 95% confidence intervals. The numbers in brackets in the y-axis captions are the estimated degrees of freedom (Edf) of the smooth curves.

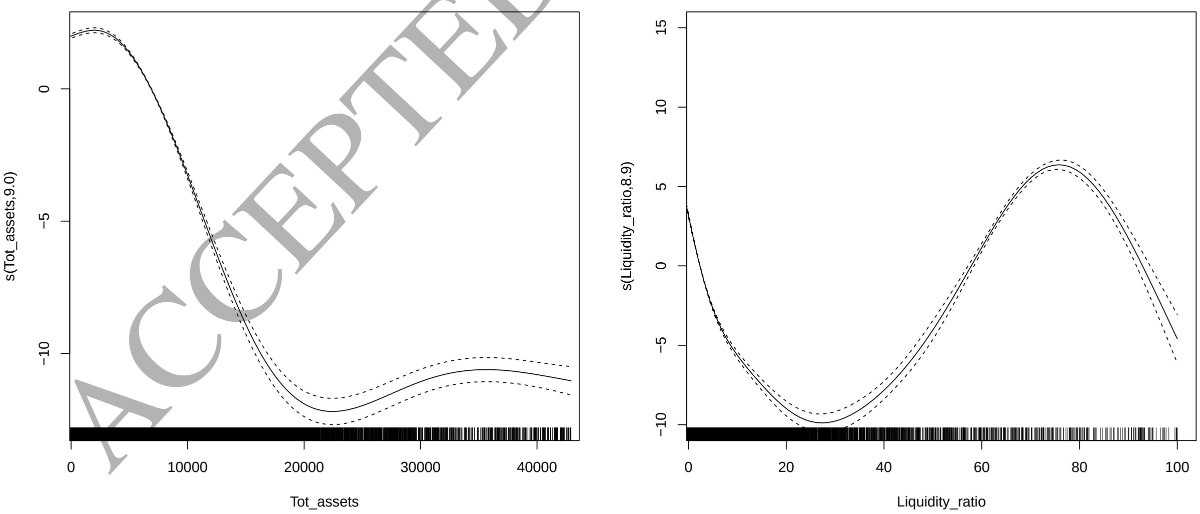

Figure 2: Smooth component estimates of the 2 (out of 4) continuous variables that exhibit a non-linear pattern. These were obtained from applying the BGEVA model on the UK SME data. Results are on the scale of the predictor. The plot show the 95% confidence intervals. The numbers in brackets in the y-axis captions are the estimated degrees of freedom (Edf) of the smooth curves.

Variables	*Logistic regression*				*Additive Logistic regression*				*times in*
	Italy		*UK*		*Italy*		*UK*		*all 16 models*
Short name	*Imp*	*Woe*	*Imp*	*Woe*	*Imp*	*Woe*	*Imp*	*Woe*	
Age	X	0	X	X	SX	0	SX	X	12
Cash_ flow	X	0	0	X	SX	0	0	SX	8
Cash_flow_oprev	X	X	0	0	X	X	0	0	10
Current_ratio	0	0	0	0	X	0	0	0	2
EBITDA_Margin	X	X	0	X	X	X	0	X	12
Gearing	0	0	X	X	0	0	X	X	8
Interest_cover	0	X	0	X	0	SX	0	X	8
Liquidity_ratio	X	X	X	X	SX	0	SX	X	13
Net_income	0	X	0	0	0	SX	0	0	6
No_directors	X	X	X	X	X	X	SX	X	16
No_employees	X	0	X	X	SX	0	X	X	12
No_subsidiaries	X	X	X	X	X	X	X	X	16
Op_rev	X	X	0	X	X	X	0	SX	12
PL_beforetax	0	X	0	0	0	SX	0	X	7
Profit_margin	X	X	X	0	X	X	X	0	14
ROCE	0	X	0	0	0	SX	0	0	6
ROE	X	X	0	0	X	0	0	SX	9
Sharehold_funds	X	X	X	0	X	X	X	0	14
Shareh_liquidity_ratio	0	0	0	0	0	0	X	0	5
Solvency_ratio	X	X	X	X	X	X	X	X	16
Tot_assets	X	0	X	0	X	0	SX	0	10

Table 2: Significant variables across the countries for logistic and additive logistic models. X - the variable is significant at 10% s.l. or lower; SX - the smooth term of the variable is significant at 10% s.l. or lower

Variables	*GEV model*				*BGEVA model*				*times in*
	Italy		*UK*		*Italy*		*UK*		*all 16 models*
Short name	*Imp*	*Woe*	*Imp*	*Woe*	*Imp*	*Woe*	*Imp*	*Woe*	
Age	X	0	X	X	SX	0	SX	X	12
Cash_flow	X	0	0	X	SX	0	0	SX	8
Cash flow_oprev	X	X	0	X	X	X	0	X	10
Current_ratio	0	0	0	0	X	0	0	0	2
EBITDA_Margin	X	X	0	X	X	X	0	X	12
Gearing	0	0	X	X	0	0	X	X	8
Interest_cover	0	X	0	X	0	SX	0	X	8
Liquidity_ratio	X	X	X	X	SX	0	SX	X	13
Net_income	0	X	0	X	0	SX	0	X	6
No_directors	X	X	X	X	X	X	SX	X	16
No_employees	X	0	X	X	SX	0	X	X	12
No_subsidiaries	X	X	X	X	X	X	X	X	16
Op_rev	X	X	0	X	X	X	0	SX	12
PL_beforetax	0	X	0	X	0	SX	0	X	7
Profit_margin	X	X	X	X	X	X	X	X	14
ROCE	0	X	0	X	0	SX	0	X	6
ROE	X	X	0	X	X	0	0	SX	9
Sharehold_funds	X	X	X	X	X	X	X	SX	14
Shareh_liquidity_ratio	0	0	X	X	0	0	X	SX	5
Solvency_ratio	X	X	X	X	X	X	X	X	16
Tot_assets	X	0	X	X	X	0	SX	X	10

Table 3: Significant variables across the countries for GEV and BGEVA models. X - the variable is significant at 10% s.l. or lower; SX - the smooth term of the variable is significant at 10% s.l. or lower

Methods for missing values	*measure*	*GEV model*	*logistic*	*BGEVA model*	*additive logistic*
Weight of Evidence	MAE^+	0.784	0.798	0.782	0.797
	MSE^+	0.722	0.705	0.702	0.702
	AUC	0.741	0.731	0.722	0.717
Imputation	MAE^+	0.862	0.909	0.761	0.969
	MSE^+	0.807	0.838	0.713	0.941
	AUC	0.632	0.632	0.677	0.677

Table 4: Forecasting accuracy measures for out-of-sample exercise obtained from applying the GEV and logistic model and BGEVA and logistic additive models to UK data.

Methods for missing values	*measure*	*GEV model*	*logistic*	*BGEVA model*	*additive logistic*
Weight of Evidence	MAE^+	0.803	0.804	0.781	0.782
	MSE^+	0.679	0.684	0.651	0.662
	AUC	0.813	0.812	0.824	0.825
Imputation	MAE^+	0.835	0.814	0.891	0.803
	MSE^+	0.730	0.711	0.835	0.701
	AUC	0.806	0.806	0.799	0.801

Table 5: Forecasting accuracy measures for out-of-sample exercise obtained from applying the GEV and logistic model and BGEVA and logistic additive models to Italian data.

Variables names	*Italy*			*UK*		
of parametric model	*Estimate*	*Std.Error*	*p-value*	*Estimate*	*Std.Error*	*p-value*
Intercept	-1.308e+00	1.526e-02	< 2e-16	3.888e+00	6.149e-02	< 2e-16
Cash_flow_oprev	4.011e-03	1.308e-03	0.002	-	-	-
Current_ratio	1.274e-01	1.473e-03	< 2e-16	-	-	-
EBITDA_Margin	-4.701e-03	1.153e-03	4.56e-05	-	-	-
Gearing	-	-	-	1.086e-02	2.316e-04	< 2e-16
No_directors	-2.935e-01	8.068e-03	<2e-16	-	-	-
No_employees	-	-	-	7.436e-02	1.512e-03	< 2e-16
No_subsidiaries	-1.145e-01	1.150e-02	< 2e-16	-9.365e-01	1.929e-02	< 2e-16
Op_rev	1.563e-05	2.519e-06	5.39e-10	-	-	-
Profit_margin	-3.655e-03	8.980e-04	4.70e-05	-7.075e-02	1.436e-03	< 2e-16
ROE	-2.986e-04	5.703e-05	1.64e-07	-	-	-
Shareh_liquidity_ratio	-	-	-	-1.416e-02	2.864e-04	< 2e-16
Sharehold_funds	-1.137e-04	8.391e-06	< 2e-16	8.769e-04	1.811e-05	< 2e-16
Solvency_ratio	-8.894e-03	3.854e-04	< 2e-16	-8.453e-02	1.724e-03	< 2e-16
Tot_assets	4.043e-05	2.595e-06	< 2e-16	-	-	-
of Smooth terms	*Edf*	*Est.rank*	*p-value*	*Edf*	*Est.rank*	*p-value*
age	2.987	3	0.021	9.000	9	< 2e-16
Cash_flow	8.950	9	<2e-16	-	-	-
Liquidity_ratio	8.084	9	<2e-16	8.914	9	< 2e-16
No_directors	-	-	-	9.000	9	< 2e-16
No_employees	3.898	4	<2e-16	-	-	-
Tot_assets	-	-	-	9.000	9	< 2e-16

Table 6: Parametric and smooth component summaries obtained from applying the semiparametric BGEVA model to the samples of Italian and UK SMEs. The missing values are analysed by imputation method. The values of τ parameters for Italian and UK models are -0.41 and -0.9, respectively.

Variables names	*Italy*			*UK*		
of parametric model	*Estimate*	*Std.Error*	*p-value*	*Estimate*	*Std.Error*	*p-value*
Intercept	-1.334	0.011	< 2e-16	-1.572	0.028	<2e-16
age_w	-	-	-	5.367	0.049	<2e-16
cash_flow_oprev_w	0.128	0.018	3.77e-12	0.398	0.021	<2e-16
EBITDA_Margin_w	0.108	0.020	3.92e-08	1.506	0.017	<2e-16
Gearing_w	-	-	-	-1.617	0.025	<2e-16
Interest_cover_w	-	-	-	0.947	0.009	<2e-16
Liquidity_ratio_w	-	-	-	0.879	0.013	<2e-16
Net_income_w	-	-	-	-0.536	0.040	<2e-16
No_directors_w	0.588	0.013	<2e-16	3.486	0.027	<2e-16
No_employees_w	-	-	-	4.004	0.039	<2e-16
No_subsidiaries_w	0.216	0.031	1.83e-12	6.242	0.058	<2e-16
Op_rev_w	-0.262	0.033	2.43e-15	-	-	-
PL_beforetax_w	-	-	-	0.947	0.033	<2e-16
Profit_margin_w	0.106	0.022	2.08e-06	0.261	0.017	<2e-16
ROCE_w	-	-	-	0.439	0.010	<2e-16
Sharehold_funds_w	-0.154	0.023	5.90e-11	-	-	-
Solvency_ratio_w	0.365	0.018	<2e-16	2.087	0.021	<2e-16
Tot_assets_w	-	-	-	0.750	0.024	<2e-16
of smooth terms	*Edf*	*Est.rank*	*p-value*	*Edf*	*Est.rank*	*p-value*
Cash_flow_w	-	-	-	8.808	9	< 2e-16
Interest_cover_w	8.488	9	<2e-16	-	-	-
Net_income_w	5.592	6	<2e-16	-	-	-
Op_rev_w	-	-	-	8.602	9	< 2e-16
PL_beforetax_w	8.908	9	<2e-16	-	-	-
ROCE_w	8.649	9	<2e-16	-	-	-
ROE_w	-	-	-	8.231	9	< 2e-16
Shareh_liquidity_ratio_w	-	-	-	7.303	8	< 2e-16
Sharehold_funds_w	-	-	-	3.961	4	3.49e-12

Table 7: Parametric and smooth component summaries obtained from applying the semiparametric BGEVA model to a sample of Italian and UK SMEs. The missing values are analysed by Weight of Evidence method. The values of τ parameter for Italian and UK models are -0.41 and -0.42, respectively

	Default =0 (healthy), n=151178			*Default =1 (default), n=11525*			*Default = 2 (dissolved), n= 2272*		
	% miss	Mean	St dev	% miss	Mean	St dev	% miss	Mean	St dev
Age	0.00	206.61	155.07	0.00	172.16	137.10	0.00	145.63	103.33
Capital	0.03	248.51	2241.88	0.35	335.40	3089.62	0.61	63.86	356.54
Cash_flow	0.05	137.44	607.13	0.40	-387.08	5249.43	0.6875	-35.01	394.77
Cash_flow_oprev	0.20	7.05	8.23	0.73	6.61	10.25	0.87	8.54	10.84
Current_liab	0.03	1686.67	4204.70	0.35	2606.82	6569.06	0.61	459.50	1910.91
Current_ratio	0.03	1.78	2.80	0.35	1.50	3.33	0.62	2.47	5.64
EBITDA_Margin	0.06	7.48	13.67	0.45	-6.27	24.47	0.72	-7.74	25.87
Gearing	0.17	187.20	218.93	0.66	229.94	248.69	0.78	116.42	193.03
Interest_cover	0.11	27.22	97.25	0.43	4.21	69.97	0.72	8.88	98.82
Liquidity_ratio	0.03	1.36	2.37	0.35	1.15	2.98	0.62	2.29	5.78
Loans	0.03	445.10	1453.38	0.35	888.28	3246.21	0.61	94.76	720.79
Net_income	0.03	33.62	560.09	0.35	-530.82	5254.84	0.61	-56.22	366.75
No_directors	0.00	1.37	1.82	0.00	0.13	0.69	0.00	0.00	0.00
No_employees	0.03	13.73	22.70	0.34	13.29	24.56	0.61	6.23	14.50
No_subsidiaries	0.00	0.39	0.80	0	0.12	0.45	0.00	0.00	0.00
Noncurrent_liab	0.03	720.19	3586.76	0.35	1121.89	10009.62	0.61	141.74	997.82
Op_rev	0.03	3100.14	5649.10	0.35	2378.32	5347.09	0.61	776.59	2811.41
PL_beforetax	0.03	79.92	635.17	0.34	-483.08	3715.36	0.61	-60.77	411.73
Profit_employee	0.05	11.77	66.88	0.44	-6.10	63.39	0.64	-6.80	29.68
Profit_margin	0.04	1.71	13.29	0.43	-12.53	25.13	0.66	-12.92	28.44
ROCE	0.08	12.17	52.98	0.53	-24.31	118.41	0.75	-33.93	156.63
ROE	0.09	14.87	93.09	0.60	-34.53	175.52	0.77	-33.83	173.69
Shareh_liquidity ratio	0.05	7.79	38.08	0.40	2.69	30.77	0.72	8.16	51.18
Sharehold_funds	0.03	1068.87	4489.58	0.35	-242.39	7134.65	0.61	113.77	1651.52
Solvency_ratio	0.04	24.26	23.68	0.41	4.75	32.58	0.67	16.87	41.34
Tot_assets	0.03	3475.74	9410.08	0.35	3486.66	9829.77	0.61	714.95	3542.29

Table 8: Some descriptive statistics for healthy, defaulted and dissolved Italian SMEs.

	Default =0 (healthy), n=39371			*Default = 1 (default), n= 960*			*Default = 2 (dissolved), n= 1192*		
	% miss	Mean	St dev	% miss	Mean	St dev	% miss	Mean	St dev
Age	0.00	206.53	191.72	0.00	234.26	199.60	0.00	152.93	132.51
Capital	0.03	611.88	12748.83	0.34	1081.25	8673.70	0.52	1619.43	33925.85
Cash_flow	0.22	358.28	3999.09	0.44	-243.55	6358.60	0.71	261.57	4742.55
Cash_flow oprev	0.40	12.18	14.79	0.68	7.94	10.75	0.84	20.27	21.60
Current_liab	0.03	3506.71	28913.74	0.34	5180.11	14967.72	0.52	1053.59	7899.34
Current_ratio	0.05	4.79	10.13	0.35	2.45	6.74	0.57	5.44	11.44
EBITDA_Margin	0.24	9.24	21.43	0.47	0.71	19.48	0.73	7.75	33.79
Gearing	0.21	78.05	151.61	0.56	132.03	182.34	0.71	51.63	136.52
Interest_cover	0.58	39.29	118.91	0.56	10.92	74.55	0.91	18.15	90.00
Liquidity_ratio	0.06	4.55	10.08	0.36	2.30	6.80	0.57	5.34	11.42
Loans	0.04	1825.15	15277.83	0.35	2824.16	11707.98	0.55	761.48	7943.12
Net_income	0.04	130.72	4114.47	0.34	-421.83	6078.29	0.51	106.89	3658.45
No_directors	0.00	4.85	2.47	0.00	1.68	2.53	0.00	0.07	0.50
No_employees	0.03	37.29	47.26	0.33	48.21	53.33	0.50	15.06	34.96
No_subsidiaries	0.00	0.45	0.89	0.00	0.36	0.85	0.00	0.00	0.05
Noncurrent_liab	0.03	2310.76	18034.88	0.34	1608.22	8490.13	0.52	420.97	4789.33
Op_rev	0.03	6310.97	8689.12	0.34	7848.69	9420.69	0.50	1814.56	5632.73
PL_beforetax	0.04	207.81	4179.76	0.34	-370.17	6133.68	0.51	138.49	3686.10
Profit_employee	0.08	36.89	233.55	0.4	31.52	278.23	0.54	25.97	177.92
Profit_margin	0.09	8.04	25.79	0.39	-3.16	26.71	0.57	10.37	38.96
ROCE	0.58	19.17	82.45	0.62	-1.89	101.30	0.93	12.87	162.04
ROE	0.18	25.83	106.85	0.54	5.55	121.55	0.71	13.12	173.76
Shareh_liquidity ratio	0.52	35.31	104.34	0.61	23.95	94.05	0.91	18.18	65.22
Sharehold_funds	0.03	2349.38	19766.74	0.34	2981.21	29629.11	0.52	1615.8	19086.83
Solvency_ratio	0.07	45.51	38.11	0.39	23.48	38.55	0.60	46.61	44.40
Tot_assets	0.03	8143.35	41542.95	0.34	9782.48	37483.74	0.53	3142.41	22499.54

Table 9: Some descriptive statistics for healthy, defaulted and dissolved UK SMEs.