Refining Understanding of Corporate Failure through a Topological Data Analysis Mapping of Altman's Z-Score Model

Wanling Qiu*[1], Simon Rudkin †[2], and Paweł Dłotko‡[3]

[1]School of Management, University of Liverpool, United Kingdom
[2]Economics Department, Swansea University, United Kingdom
[3]Mathematics Department, Swansea University, United Kingdom

April 16, 2020

Abstract

Corporate failure resonates widely, leaving practitioners searching for understanding of default risk. Managers seek to steer away from trouble, credit providers to avoid risky loans and investors to mitigate losses. Applying Topological Data Analysis tools, this paper explores whether failing firms from the United States organise neatly along the five predictors of default proposed by the Z-score models. Each firm is represented as a point in a five-dimensional point cloud, each dimension being one of the five predictors. Visualising that cloud using Ball Mapper reveals failing firms are not always located in similar regions of the point cloud, that is they are not concentrated in an easily split out area of the space. As new modelling approaches vie to better predict firm failure, often using black boxes to deliver potentially over-fitting models, a timely reminder is sounded on the importance of evidencing the identification process. Value is added to the understanding of where in the parameter space failure occurs, and how firms might act to move away from financial distress. Further, lenders may find opportunity amongst subsets of firms that are traditionally considered to be in danger of bankruptcy, but which the Ball Mapper plots developed herein clarify actually sit in characteristic spaces where failure has not occurred.

Keywords: Credit Scoring; Topological Data Analysis; Data Visualization; Bankruptcy Prediction

1 Introduction

Credit default prediction models intuitively find direction from the financial fundamentals of the corporation and identify how such can be used to indicate likely future failures. In Beaver (1966) and Beaver (1968) individual financial ratios are tested for their ability to discriminate between firms that go bankrupt the following year and those who do not. Altman (1968) advances this to select the five ratios that best isolate potential failures and employs linear discriminate analysis to assign a coefficient to each. Subsequent developments charted for the 50th anniversary of the original Altman (1968) model in Altman et al. (2017a) have included extensions of the considered ratio set, considerations of non-linearity, removal of the normal distribution assumption implicit in the original multiple discriminate analysis, and the introduction of machine learning (ML) models. Such works break down into those that extend the information set and those which seek to extract more from the information already provided. Each extension has merit but with both comes the danger of over-fitting to particular values. More advanced techniques also bring questions of being a "black

*Corresponding Author. Full Address: Accounting and Finance Subject Group, School of Management, University of Liverpool, 20 Chatham Street, Liverpool, L69 7ZH, United Kingdom. Tel: +44 (0)7955 109334 Email:wanling.qiu@liverpool.ac.uk

†Full Address: Economics Department, School of Management, Swansea University, Bay Campus, Swansea, SA1 8EN, United Kingdom. Email:s.t.rudkin@swansea.ac.uk

‡Full Address: Mathematics Department, College of Science, Swansea University, Bay Campus, Swansea, SA1 8EN, United Kingdom. Email:p.t.dlotko@swansea.ac.uk.

box" through which the link from input to output cannot be traced. It is then unsurprising that the original approaches continue to have resonance in the credit rating sector (Altman et al., 2017a).

This paper returns to the fundamental models of Altman (1968) and its predictions of default against the true cases of corporate failure. A Topological Data Analysis (TDA) Ball Mapper (BM) approach after (Dłotko, 2019) is used to produce an abstract two-dimensional representation of the financial ratio space to visualise where failures occur amongst the combinations of firm characteristics. Henceforth this approach is referred to as BM. Mapping the space in this way demonstrates how contemporary approaches in data science can break open the black box and illuminate how precisely the future of credit default prediction modelling should develop. Major advantages of the approach include a robustness to high levels of correlation between variables, robustness to noise within the dataset and critically respect the underlying data rather than imposing distributional assumptions.

Contributions of our work are thus threefold. Firstly, a demonstration of the application of a contemporary data science technique to financial analysis inspires a new understanding of the space upon which we conduct our evaluation of credit models. It is immediately seen that the areas of the characteristic space classified as likely to contain failures cover a lot of volume in which there are no failures. Secondly, it is demonstrated how non-linearity and interaction terms can all be accounted for in the evaluation of credit default risk; insights gained therefrom become invaluable for assessing firms. Finally a research agenda is signposted which can aid understanding of credit risk and open the "black box" of ML. BM thus offers much to the discussion of bankruptcy prediction.

The remainder of the paper is organised as follows. Section 2 offers a brief overview of the literature defining the problem of balancing fit against the risk of over-fitting. Section 3 details the data used for the empirical work, with Section 4 highlighting the BM approach and the theoretical expectations formed prior to the analysis. Section 5 illustrates the Altman (1968) model through the BM lens. Extensions emerging from the consideration are discussed in Section 6, with Section 7 offering a review of the lessons learned. Finally, Section 8 concludes.

2 Literature Review

Corporate failure has obvious repercussions for investors and those to whom the failed firm has liabilities. Consequently it is banks and financial institutions who are the greatest users of these models. For such users decisions on creditworthiness of potential borrowers must be traceable, clearly motivated by evidence and free from any allegations of black boxes. Whilst ML methods may generate more accurate predictions, their data driven nature places too many unknowns in the process of getting from input information to output decision. Hence whilst there is a growing literature praising the virtues of a ML approach, this paper stands as a note of caution there against.

2.1 Development of Credit Default Models

Altman (1968) and Altman (1983) models were constructed by multiple discriminate analysis (MDA) from a set of candidate factors. Each accounting ratio put forward as a potential explanatory variable is assessed for its ability to explain firm failure, with only those making a significant contribution being considered for the final model. The five factors that made the final Altman (1968) Z-score were chosen from a set of twenty-one. Later models from Altman (1983) were designed to reflect non-listed firms who did not have a market value of equity to use in their evaluation. Such MDA approaches to identifying the drivers of firm failure dominated the literature to the turn of the century (Altman et al., 2017a). As well as MDA there was a growing thread after Ohlson (1980) application of logistic regression which employed probabilistic models. Extensions were made into other countries[1] but the fundamental ideas fell into one of these two categories. Critically, as demonstrated by the removal of asset turnover in the second of the Altman (1983) frameworks because of differences between industries on this characteristic, the appraisal of the researcher was always maintained as a final check on the model produced.

21st century work has been dominated by the growth of ML, with models seeking information from within the firm characteristic dataset through a variety of techniques. Many review studies chart this development,

[1]See Altman et al. (2017b) for a review.

a good example being Barboza et al. (2017). In the early works two main families of model found favour. Support vector machines (SVMs) after Cortes and Vapenik (1995) have found credibility because they sit on the bridge between the MDA and ML, generating functions for credit scoring without the restrictions on functional form that regression necessitates (Altman et al., 2017a). However, the empirical work on European countries in Altman et al. (2017a) does not offer support to linear SVMs, suggesting significant non-linearity in the data. Petropoulos et al. (2016) contends that although SVMs have intuition on their side they remain "black boxes" that do not offer simple visualisations of the process through which they arrive at their predictions. Likewise the neural network ML models that also gained early prominence are openly critiqued for the inability of the practitioner to see the process through which the outcome was derived. Such a lack of transparency has consistently led back to the MDA models.

Ensemble learning offers scope to combine the individual models to gain maximal prediction accuracy. In this way the algorithm is detecting what each approach is saying about the risk of the firm defaulting and then using that information through a weighted average to construct an overall expectation. Within the class of ensemble learning models lie bagging and boosting, which run the same algorithm multiple times and form a linear combination from the outcomes. Son et al. (2019) is a recent example of work on bankruptcy prediction using a boosting algorithm to extract more information from a neural networks model. Use of multiple algorithms is considered as "hybrid" ensemble learning. Early works in this area combined MDA and logistic models (Li et al., 2012), while more recent work uses a fuller suite of ML models as well (Choi et al., 2018). De Bock (2017) exposits how spline-rule ensembles can move learning beyond the linear combinations and, in so doing, address many of the non-linearities present within financial data. Unlike standard ensemble models, these rule based ensemble learners can include different candidate algorithms at each node of the decision tree, different combination rules for each node, and offer greater options for model interpretation.

Financial practitioners are keen to know which variables are important and what probability of default is associated with each level of the explanatory factors. De Bock (2017) exposition of spline-rule ensemble learning shows how low cash ratios are associated with higher failure probabilities but increasing percentages of late payments being made by the company are unsurprisingly positively linked to failure. Some factors such as the solvency ratio and return on investment are non-linear. Such outputs, combined with the inherent increased accuracy that comes from combining measures, make ensemble learning models appeal to practice. However, the "black box" nature of the inputs to the ensemble set mean that the issue of traceability does not go away.

Ziba et al. (2019) showcases another phase in the use of ensemble models whereby synthetic elements are introduced to understand outcomes. Such creation of artificial factors is common elsewhere in statistics and hence represents a natural extension to credit default risk. Such constructions help improve fit, and have intuition as the measurements that were not able to be captured in the real data, but are inherently consequences of the observed data points and hence potentially over-fit.

Financial default prediction models are thus becoming more accurate as data science offers ever improving means of extracting information from data. Whether through pattern recognition, probability of default function fitting, or simple classifications, the increased ability to use data improves fit relative to the Altman (1968) approach. However, it is imperative that understanding is driven from fundamentals and is clear to all users; to this end the ML model set have much to make up. It is to this requirement for transparency we speak.

2.2 Topological Data Analysis and Ball Mapper

In its most intuitive form a topological analysis seeks to create a map of a dataset, helping the analyst view what is going on in each part of the space. When constructing maps we intuitively turn to the longitude and latitude as point co-ordinates and then overlay information as contours, as colour or as points. This paper demonstrates a contemporary approach that generates a "map" of the financial ratios of firms. Rather than simply the two dimensions of the page, the algorithm is constructing a two-dimensional visualisation of the multiple dimensional data that can then be placed onto the page. Other characteristics of maps, like colour and labelling, can readily be overlaid. Because of the loss of dimensionality required to create the two dimensional map there is no longer any scale that allows measurement between points but the full information remains to allow the computer so to do. From a practice perspective, being able to maintain

the complete information set, whilst providing a visual representation thereof, has obvious value. That the resulting plot may be understood by all users enhances this value further.

BM representations have the immediate advantage of being constructed solely from the data collected. Once the variables to be collected are decided, measurements are taken and the data is plotted onto the space. By looking at the shape of the data, BM is constructing a map of what is there and not seeking to impose any relationships upon the data. Intuitively a linear model is assuming that there can be a straight line drawn through the data such that the outcome is a linear function of the input. Familiarity with adding regression lines to a scatter plot means the effect of linear regression is understood. By using increasingly complex non-linear functions it is possible to fit the data better and reduce the residuals, but this comes at the cost of the model estimated being particular to the data in the sample. This is also known as over-fitting. ML is helping to make that non-linear relationship more accurate to the data, but the over-fitting criticism still stands. Employing TDA enables understanding of the linearity of the relationship and, if there is no linearity, allows the researcher to know more about the shape of the data.

From a modelling perspective there are further obvious advantages created because models can be run and their fit to the respective parts of the point cloud assessed. Where poor fit is detected, improvements can be made which allow the model to fit such firms better, for example through the introduction of an interaction term or joint threshold effect. In the Altman (1968) sense a low z-score is achieved where sales are low or profit is low but this would cause a firm with low sales to get a low z-score even if they had higher profit. Adding a joint requirement that says sales and profits must be low, or an interaction term to say that the product of sales and profit must be low, would be a way around this. BM graphs would be able to show quickly which helped to fit the data better. Critically the non-linearity these terms induced would be visible and explainable unlike the automated processes through which non-linear ML models introduce such terms. For a neat discussion of the econometric and ML issues raised here see Mullainathan and Spiess (2017).

Broadly, TDA has been limited to the physical sciences where it is valued for its robustness to noise and ability to capture relationships between data points irrespective of the way in which they are differentially perturbed. For example in looking for genetic mutations it is important to be able to distinguish between small differentials between individuals and genuine changes in the gene that might signal a need for treatment(Nicolau et al., 2011). Work to bring TDA into the finance field has typically focused on time series and considered the possibility of financial crashes (Gidea and Katz, 2018). Therein it is the use of TDA to monitor for potential crashes in dynamical systems, such as production lines, that provides the inspiration. This paper represents one of the first applications of the cross section approach outside the natural sciences[2].

From the perspective of the institution seeking to understand the riskiness of a particular business, the value of the model based approaches is obvious. However, to really understand where the risk is high, drawing from the data is more intuitive. Taking the known characteristics of a firm and placing it within the picture can guide on the risk for such a firm. BM does not offer regression coefficients, but by looking at firms in the same space inference can be gained. Though analysts may have a feeling about which parts of space are risky, the BM algorithm may either confirm, or sit at odds with, those initial thoughts. A process of learning what is really going on in the data is then the first step to getting the best impression of credit default risk.

3 Data

Data is constructed from Compustat and covers the period from 1961 through to 2015. Although there is more contemporary data available there are few recorded cases of failure since 2015 at the time of writing. This is due to the lag in cases entering the Compustat data. Formally a firm is regarded as failed if it either files for bankruptcy or liquidates in the financial year. For failed firms, data from the most recent financial statements is provided alongside a deletion reason[3].

Explanatory variables are taken from the respective works of Altman (1968) and Altman (1983). Each is constructed from Compustat data using the formulae defined in Table 1 and contains an allowance for size through a denominator of either total assets or, in the case of X_4, the total liabilities of the firm. These

[2]At the time of writing the only known example is Vejdemo-Johansson et al. (2012), which looks at voting behaviours in the United States of America House of Representatives.

[3]Specifically this paper considers either bankruptcy (code 02) or Liquidation (code 03)

Table 1: Variable Construction and Summary Statistics

	Description	Compustat	Mean	s.d.	Min	Max
X_1	Working Capital / Total Assets	$(act - lct)/at$	0.216	0.218	-0.483	0.748
X_2	Retained Earnings / Total Assets	re/at	-0.089	0.900	-7.699	0.694
X_3	EBIT / Total Assets	$(ni + xint + txt)/at$	0.041	0.170	-1.014	0.310
X_4	Market Value of Equity / Total Liabilities	$(csho \times prcc_f)/tl$	2.943	4.216	0.093	31.92
X_5	Sales / Total Assets	$sale/at$	1.119	0.704	0.001	3.542
	Firm Failure	$delrsn = 1, 2$	0.037	0.188	0	1

Notes: All data is sourced from Compustat. Description provides the formulae from Altman (1968) or Altman (1983) for the construction of the X variables. The column Compustat details the variable names used in the construction of the explanatory factors (X_1 to X_5). Compustat variable names are as follows *act* - current assets, *lct* - current liabilities, *at* - total assets, *re* - retained earnings, *ni* - net income, *xint* - interest payments made, *txt* - taxation on earnings paid, *chso* - current shares outstanding, $prcc_f$ - price of the share at the financial year end, *sale* - total sales of the firm and *delrsn* is the reason for deletion from the Compustat database. Firm failure is a dummy for deletion from the Compustat dataset in the subsequent year owing to either bankruptcy or liquidation. Sample from 1961 to 2015, $n = 110668$

Table 2: Full Sample Correlations

	X_1	X_2	X_3	X_4	X_5	Fail
X_1	1					
X_2	0.108	1				
X_3	0.128	0.511	1			
X_4	0.279	-0.021	0.136	1		
X_5	0.355	0.066	0.102	-0.055	1	
Fail	0.040	-0.010	-0.046	-0.043	0.046	1

Notes: All data is sourced from Compustat. Financial ratios are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). Fail is a dummy for deletion from the Compustat dataset in the subsequent year owing to either bankruptcy or liquidation. Sample from 1961 to 2015, $n = 110668$

ratios capture the liquidity (X_1), profitability (X_2), productivity (X_3), leverage (X_4) and asset turnover (X_5). Mitigating the effect of extreme values we remove the firms with values of any of the five X variables in the highest and lowest 0.5% of the observations for that variable in the full data set. We further remove any observations for which there is missing data and are left with 110668 firm-years of which 3.7% are failed firms.

Table 2 provides full sample correlations between the five explanatory ratios of Altman (1968) and the firm failure dummy. Amongst these there are high correlations between profitability, X_2, and productivity, X_3, as well as between liquidity , X_1, and asset turnover, X_5. None of these values touch the 0.7 in absolute value that would be seen as a sign of multicollinearity, but the correlation is high and so regression analyses should note that in their exposition.

It is recognised that effects may vary from year to year and so Table 3 summarises the five ratios for each year considered in this paper. For brevity the number of years considered is just 6; being in 10 year intervals to the most recent data and 2008 to capture what was happening at the start of the global financial crisis. Failure proportions were much higher during the early years, whilst most recently the failure proportion has been very low. Even at the height of the financial crisis the percentage of firms that failed was just 1%.

4 Methodology

Analysis of the shape of the data begins with the construction of the point cloud. For credit default modelling using Altman (1968) this is achieved by plotting each firm as a point in five dimensions. Each coordinate being one of the X_j's used in the formation of the Z-score. In this paper different clouds are formed for each of the years studied. Firms which are proximate in the five dimensional space must have similar values for all of the considered financial ratios. A theoretical introduction to the method follows, with consideration then given to the representation that might be expected to emerge.

Table 3: Annual Summary Statistics

Year	Financial Ratios					Failure
	X_1	X_2	X_3	X_4	X_5	(%)
1975	0.273	0.259	0.111	1.229	1.169	6.91%
	(0.196)	(0.250)	(0.091)	(1.792)	(1.721)	
1985	0.234	0.073	0.045	2.492	2.519	6.26%
	(0.22)	(0.544)	(0.169)	(3.497)	(4.105)	
1995	0.214	-0.101	0.033	3.682	3.735	2.90%
	(0.225)	(0.779)	(0.174)	(4.951)	(5.541)	
2005	0.196	-0.361	0.021	4.081	4.076	2.01%
	(0.224)	(1.222)	(0.180)	(4.992)	(5.118)	
2008	0.188	-0.327	-0.013	2.446	2.278	1.01%
	(0.225)	(1.181)	(0.216)	(3.796)	(3.738)	
2015	0.17	-0.370	-0.013	3.072	3.163	0.03%
	(0.214)	(1.228)	(0.191)	(4.225)	(5.485)	

Notes:Financial ratios following Altman (1968) are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). Failure classified as de-listing from Compustat owing to either bankruptcy or liquidation. Data from Compustat.

4.1 TDA Ball Mapper

Representation of the multi-dimensional point cloud is achieved using the BM algorithm of Dłotko (2019) as implemented in the R package *BallMapper* (Dlotko, 2019). There are a number of advantages of the Dłotko (2019) approach over the original mapper algorithm developed by Singh et al. (2007) and implemented in the *BallMapper* package of R (Pearson et al., 2015). Whilst the BM algorithm selects landmark points at random, the representation of the point cloud produced by any selection will be fully consistent with any other for appropriately large features. As the ball radius changes so the representation adjusts, but it does so in a smooth way, providing consistency throughout the adjustment. For a given point cloud the construction of the BM diagram is subject only to the selection of the radius epsilon. These significant improvements are an important advance from the traditional mapper and its requirements of the user oversight of multiple functional inputs.

Formally, the BM algorithm of Dłotko (2019) starts with the point cloud, X, and a constant $\epsilon > 0$. It selects a subset $C \subset X$ having the property that the set of balls $B(C) = \bigcup_{x \in C} B(x, \epsilon)$ contains the whole set X. Such a subset C is referred to as an ϵ net. Algorithm 1 of the Dłotko (2019) paper identifies neatly how the ϵ-net C is formed. Algorithm 1 ends when all of the points in X are covered by at least one ball.

Algorithm 1: Greedy ϵ-net(Dłotko, 2019)

Input: Point cloud X, $\epsilon > 0$
$C = \emptyset$;
Mark all points of X as *uncovered*;
while *There exist uncovered* $p \in X$ **do**
  $C = C \cup p$;
  Mark every point $x \in B(p, \epsilon)$ as *covered*;
end
Output: C

In this construction the ball radius ϵ is the only exogenous input. Choosing ϵ recognises the competing forces of maintaining detail and producing a representation upon meaningful inference can be made. The sequential process of Algorithm 1 can produce slightly different results based on the random selection of the next uncovered point p, but because all possible ϵ nets are close to each other, the impact of this randomness to the overall output is marginal[4]. Owing to the way that the balls are formed the maximal distance of

[4]In the *BallMapper* package the selection of the next landmark point is done by taking the first uncovered data point from the dataset in the order that it is provided to the algorithm. By changing the order of the data points it may be readily verified that only slight differences appear in the resulting graphs.

points from the ball's centre is bounded by ϵ. This may be the entire ϵ on one axis and zero distance on others. The total quantity of the distance can also be shared out across all the axes. To think about this, consider the unit circle centered at the point $(0, 0)$ being drawn on a two dimensional plane. In this case every point (x, y) that satisfies $x^2 + y^2 \leq 1$ will be in that unit circle. In general it is not possible to decide the optimal value of the parameter ϵ and so it is left to the researcher to determine how big to set the radius.

Algorithm 2: Construction of a BM graph (Dłotko, 2019, Algorithm 3)

Input: C, X, ϵ
V = abstract vertices, one per each element of C;
$E = \emptyset$, ;
for $p_1, p_2 \in C$ *such that there exist* $x \in X \cap B(p_1, \epsilon) \cap B(p_2, \epsilon)$ **do**
  $E = E \cup \{p_1, p_2\}$;
end
Result: BM Graph, G=(V,E)

Conversion of the output from Algorithm 1 into a BM graph requires a further stage of graph construction. Algorithm 2 provides such a step when an abstract graph to summarize the shape of X is constructed. As defined by the algorithm an edge is drawn between the centroids of every two balls which have data points in their intersection; such lines help to identify where in the cloud each ball sits relative to the others. Because of the way that the graph is constructed it would be expected that more vertices would appear in the BM than when using conventional mapper of Singh et al. (2007). Consequently, there may be additional information which is visualised in the BM graph.

An important decision in the construction of graphs is whether the variables should be normalised. In this paper we do normalise all axes onto the range $[0, 1]$ to recognise the variability in ranges identified in Tables 1 and 3. Here the smallest value observed for any characteristic is normalised to 0 and the largest to 1. In other applications normalisation might not be appropriate.

4.2 Interpreting TDA Ball Mapper Graphs

BM graphs have several key features that aid the understanding of the data they plot. Although necessarily abstract, BM graphs are topologically faithful to the underlying dimensions. Stemming from the notion that data points are random draws from an underlying manifold, rather like statisticians may view data points as draws from underlying distributions, having sufficiently many points enables the homology to be recovered with high confidence (Niyogi et al., 2008). BM uses an $\epsilon - net$ to capture the space, meaning that any landmark in that net is ϵ away from the considered point cloud. Consequently the set of landmark points and initial point cloud are at most epsilon away from each other in a Hausdorff Metric. Hence, provided ϵ is smaller than the injectivity radius from Niyogi et al. (2008), we get the topological connectivity information from the underlying manifold with high confidence[5]. This is what is referred to as topologically faithful when discussing TDA, and here the BM graph. Thus the BM graph approximates well the dataset, and underlying manifold, that it seeks to represent.

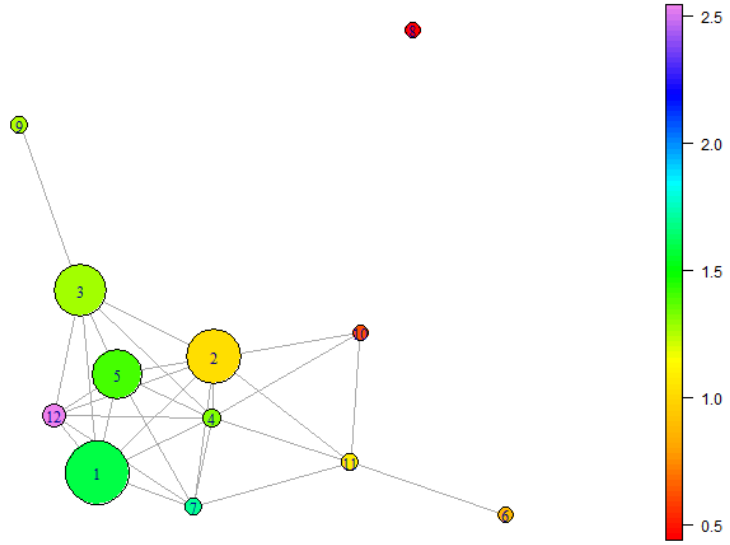
As an illustration of the properties a plot using 1975 data is provided in Figure 1.

Firstly, the colouration of the graph allows analysis of the distribution of an outcome of interest across the space. This may simply be the average value for all the data points contained within the ball, as is done in this paper, but it is also possible to use counts, standard deviations, minima, maxima, etc. The choice of function is methodologically left to the user to define. Because in most instances it is the average outcome that is considered most representative of a ball, it is this which is the default function in the *BallMapper* package. A scale to the right of the plot shows the values, here Z-scores from the Altman (1968) model. In this way it can be seen that the lowest scores sit to the right of the plot, with the only ball averaging over 2 being ball 12.

Secondly, the size of the balls gives indication of the number of data points located within that part of the plot. Bigger balls mean more points and a denser data concentration within that ϵ radius of the central

[5]Full details of the injectivity radius may be found in Niyogi et al. (2008). For the purposes of this paper we simply recommend users to choose an ϵ low enough to be able to usefully gain information from the BM graph.

Figure 1: Interpreting Ball Mapper graphs



Notes: Example BM plot created using *BallMapper* (Dlotko, 2019). Axes following Altman (1968) are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). Data is from Compustat and represents the value of these variables in 1975. Colouration is the Z-score as calculated using $Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5$. All axes are normalised to 0,1. $\epsilon = 0.5$.

point of the ball. In Figure 1 it is ball number one that has the most points contained within it, closely followed by 2, 3 and 5. There are a number of less populated balls spreading out to the right of the figure. Some like 4 and 12 are very close to these larger balls.

Connectivity between balls represents the third useful feature of a BM graph. Figure 1 contains a large number of edges emanating from most of the balls. In this way we would conclude that there is a lot of overlap and hence the whole graph is covering a cloud of similar data. There are however smaller arms sticking off to balls 6 and 9 that would represent parts of the cloud that extend out from the main collection of points. Finally where a ball is not connected we are identifying potential outliers. Ball 8 to the top right of Figure 1 is such a point. Note that the location of any disconnected balls does not show where they sit relative to the main connected components since the plot is abstract.

Fourthly the BM plots demonstrate correlation between the axis variables. For example in the two dimensional case a set of correlated points occur within a narrow band; this extends into multiple dimensions such that the BM graph will itself be closer to being a long thin shape. Note that the BM graph is abstract so it is unlikely that the line drawn would be truly straight, but rather it would bend round to fit within the plot. Where variables are less correlated the cover will need to spread out, giving the more net-like appearance seen at the lower left of Figure 1.

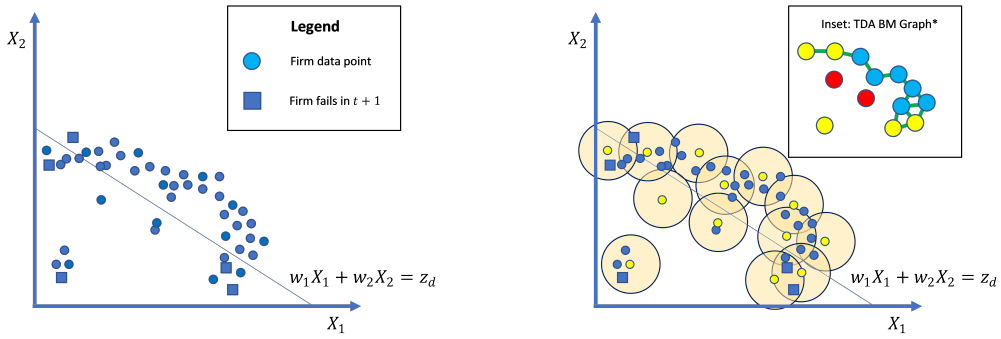
Although not reported immediately by the plot, the number of balls will give the analyst a feeling for the level of detail contained. Smaller ball radius parameters, lower ϵ, will lead to more balls being needed to cover the set of data points. Precise determination of ϵ for any given application is a matter for the analyst to determine, but we might conclude that the choice made in Figure 1 is too high as there is not much detail being gained at the centre of the plot. In what follows a smaller ϵ is used.

BM thus has a number of useful features that can help interpret the link between firm characteristics and firm failure. As with all methodologies, the final choice of inputs will be the defining factor for the value of the analysis performed.

4.3 Theoretical Linkages Between TDA Ball Mapper and Altman (1968)

In developing the intuition behind the TDA BM algorithm in this paper, careful attention has been paid to using examples from credit scoring models in the spirit of Altman (1968). However, to fully appreciate

Figure 2: Artificial Two-Dimensional Motivation of Altman (1968) TDA Ball Mapper link



(a) Point cloud as scatterplot

(b) Construction of BM representation

Notes: Points indicate individual firms values of X_1 and X_2, where these are ratios of the type constructed in Altman (1968). In both figures points towards the top right are omitted to focus on the area in and around the distress zone. Squares are used to indicate that the firm fails in year $t+1$. Panel (a) shows only the datapoints and the line which links all points on the boundary of the distress zone. Where z_d is the critical value identified as defining the distress zone, w_1 and w_2 are the weights given to X_1 and X_2 respectively. Panel (b) draws balls in line with Algorithms 1 and 2, connecting those balls where there are points in the intersection. Points used as the centre of the balls are coloured yellow. The resulting BM plot is shown as an inset to panel (b). Note that in this panel ball sizes have not been adjusted. Further, the colouration simply shows whether a point was in the distress zone, coloured red, contains firms that subsequently failed, coloured yellow, or did not contain either firms in the distress zone or any firm that failed, coloured blue. Because of these reductions in the detailing the inset is labelled with an *. A colour version of this plot is available in the online version of the paper.

the messages emerging, it is useful to think more about the theoretical underpinnings the approach that the original Altman (1968) paper took to identifying the "distress zone". Considering the intuition of LDA in two-dimensions again simplifies the correspondence between LDA and BM. It does so by allowing us to draw readily understood diagrams with axes. All that is discussed here in two-dimensions immediately generalises to any number of axes.

Figure 2 plots just two axes, X_1 and X_2, upon which firms' characteristics are charted. The equivalent of the z-score for firm i, z_i is simply a weighted sum of each firm's coordinates on the two axes, $z_i = w_1X_{i_1} + w_2X_{i_2}$. In Altman (1968) MDA is used to determine the weights w_1 and w_2. The distress zone is defined as firms having $z_i < z_d$, the equivalent of having a z-score less than 1.8 in Altman (1968). It is immediate that such a region would be the triangle nearest the origin[6]. This paper asks whether it is sufficient to simply think of the distress zone in this way, or whether BM might be able to provide more direction to those assessing firms. In essence this is because BM is considering the whole space, including being able to talk about all of the combinations of axis variables. In this two dimensional space we can talk about any combinations of X_1 and X_2, such as areas where both are high. The linear model shows X_1 being high but X_2 low would produce a z-score outside the distress zone ($> z_d$), adding a requirement that both X_1 and X_2 be larger than some critical values, or that their product must be large, would rule out areas close to either axis. These terms may enter the model and have weights estimated thereupon, but to so do we would increase the degrees of freedom. As we do not consider statistical models, the impact of the additional degrees of freedom is of less importance than if, as in statistical modelling, it were to affect the accuracy of our estimates; we must recognise the issue in the comparison though.

As a second stage, balls may be constructed around the points and the BM representation constructed. Panel (b) of Figure 2 overlays a BM coverage with a small filtration radius. Joins are made between balls that have points in their intersection. An inset to panel (b) shows the effect of removing the other data points and the axes. Representing firms that subsequently fail as squares enables colouration of the BM graph using yellow to show balls that contain failed firms, red for balls within the distress zone that do not contain failed firms, and finally blue for all other balls. Further abstraction of the plot, and the resizing of

[6]Note that because we perform normalisation, and the axes were on different ranges, the shape may be very different to that which would otherwise be seen. Consistency of mapping between the scales and the normalised version means this is not an issue for interpretation.

the balls to represent the number of points within is not carried out to maintain the intuition.

This exercise shows how BM is able to split the distress zone and, with balls coloured by firm failure, say more of the incidence of observed failure within the space. In this artificial illustration it has been designed that the two balls with high X_1 and low X_2, and that with low X_1 and high X_2, contain failures, but that the other balls centred towards the edge of the distress zone do not see failures. A failure is placed in the low X_1 and low X_2 zone to recognise that firms this low in the space are more liable to fail. In such a situation the truly concerning balls would be attached to the main shape of panel (b). No linear segregation of the space would be able to isolate the five balls that contain failed firms from the others. Therein lies the challenge for models like Altman (1968). ML models may use non-linear classification rules, but such would necessarily be complicated by their reliance on small numbers of observations. Given the low numbers of failures there are issues of over-fitting if trying to use ML. The ultimate ML technique for classification is the random forest; this uses decision rules to classify and continues to segregate the data into groups according to outcome until near perfection is reached. This is often criticised as over-fitting because new data would be unlikely to fit exactly to the patterns in the data that was fitted, as Cawley and Talbot (2010), Mullainathan and Spiess (2017) and many others note; this poor out-of-sample performance is a common argument against complex non-linear ML. Z-scores are proportional to the distance to the origin, but failure need not be. BM here is representing the data, plotting failure rates or the z-score. As illustrated it does not follow that balls with similar z-scores would be connected, or that those with similar failure rates would join.

5 Altman (1968) Z-score Model

Altman (1968) proposed the Z-score model for predicting firm failure as:

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \tag{1}$$

concluding that a Z-score of larger than 2.99 would place the firm in the safe zone and unlikely to suffer distress. A Z-score between 1.8 and 2.99 places the firm in a "grey" zone where failure cannot be ruled out. Should the Z-score be below 1.8 then Altman (1968) assigns the firm to a "distress" zone. This section uses this model to inform a BM analysis. First we look at 2015, the most recent year for which full data is available. Considering the time-varying position we then exposit analyses of 1975, 1985, 1995, 2005 and 2008; being 10 year intervals with 2008 added for its correspondence with the global financial crisis. Within the BM algorithm the choice of ball radius parameter is the only user decision, hence we end this section with a look at the effect of changing the ball radius.

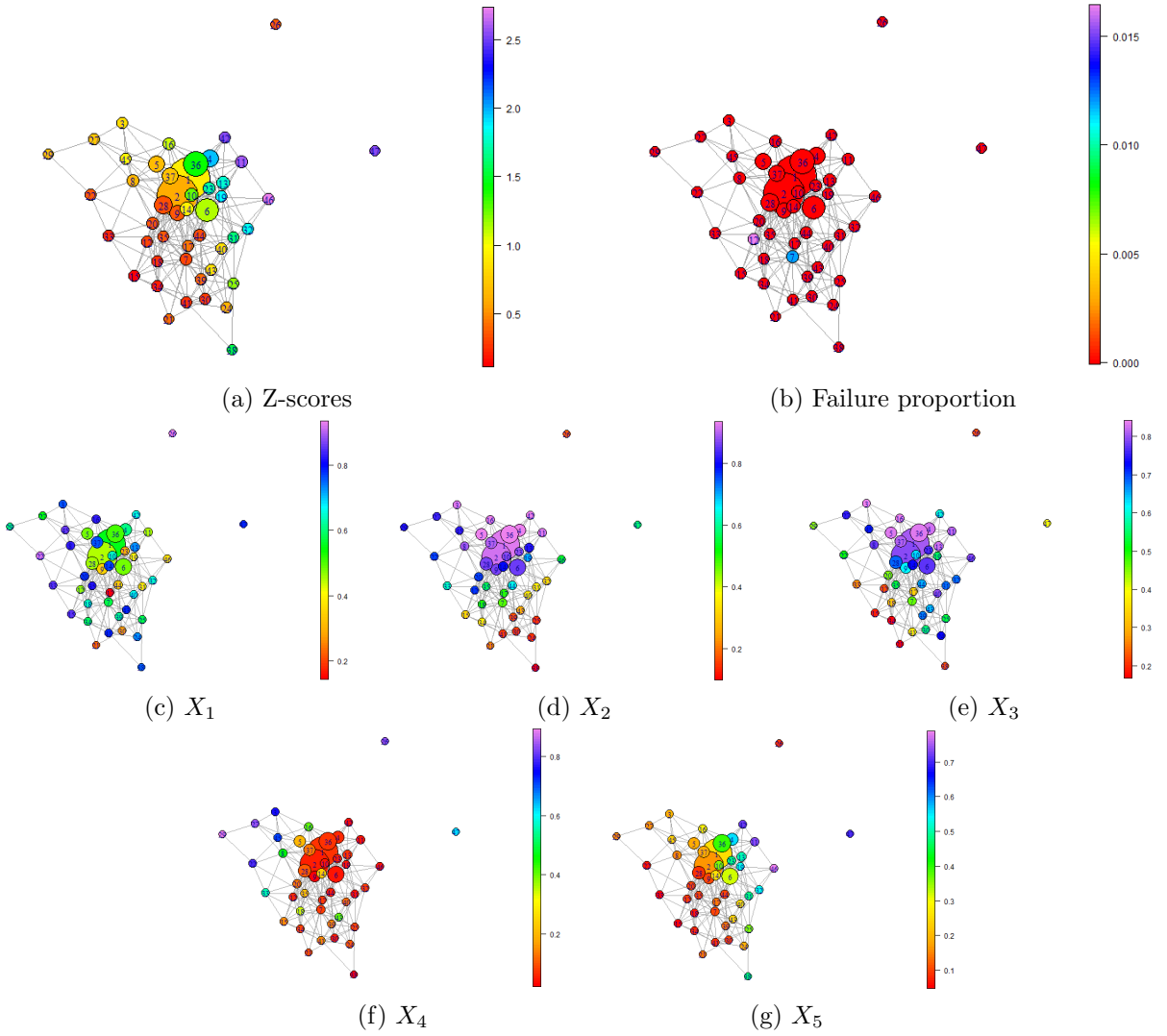
5.1 Default Prediction with Ball Mapper

As an illustration of the benefit of TDA consider the space defined by the original Altman model. BM is used to construct an abstract representation of the firm characteristics space, here in the five dimensions set out as X_1 to X_5 in Table 1. In this way the segments of the point cloud with low values for the Z-score will be clearly visible. It is then asked whether the firms that failed were indeed in this part of the space. If the model is effective then it would be expected that the proportion of firms within a given ball that fail would be highest in the balls identified as having low Z-scores. Further it should follow that the proportion of failure in balls with high Z-scores should be 0. Figure 3 examines this for 2015. To construct a BM plot a ball radius must be selected. For this purpose $\epsilon = 1.4$ is used, meaning that if all other variables are the same the largest distance between two points in a ball on the single differential axis will be 0.8 in the normalised space. This may seem large but since there are five axes that means the average distance from the central point is just $0.4/5 = 0.08$. Moving to lower numbers produces a very large set of balls and makes inference more challenging[7].

Figure 3 is divided into three key parts. Firstly the Z-scores predicted using equation (1) are plotted in panel (a). Lower values, associated with predictions of failure, are located to the bottom left of the plot and are denoted by reds and oranges in the shading. Higher values are found to the right and towards the top represented by the blues and purples. In Altman (1968) a Z-score below 1.80 is considered as placing a firm in the "distress" zone. In the plot the "distress" zone will also include the big balls at the centre right.

[7]Results from other filtrations are available on request from the authors.

Figure 3: Altman Z Scores and Firm Failures: Original Model 2015 ($\epsilon = 0.4$)



(a) Z-scores (b) Failure proportion

(c) X_1 (d) X_2 (e) X_3

(f) X_4 (g) X_5

Notes: BM (Dłotko, 2019) plots of the five dimensions of the original Altman (1968) model generated using *BallMapper* (Dlotko, 2019). Axes are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). All axis variables are normalised to the range 0 to 1 for consistency. Panel (a) is coloured according to the z-score calculated by equation (1). Panel (b) is shaded according to the proportion of observations within the ball that did fail. Panels (c) to (g) are coloured based upon the variables used in the construction of the Z-score. Here we the abstract nature of the plots. Diversity in colour stems from the normalisation process as evidenced in comparison with the actual value plots. Figures are available in colour in the on-line version of the paper.

Panel (a) also reveals that no ball has an average Z-score above 2.99 meaning that no ball is considered entirely within the "safe" zone. Panel (b) is coloured according to the proportion of firms within a given ball that suffer failure in the following year. The highest proportion is 1.5% and occurs to the lower left of the big mass. A comparison with panel (a) shows that these were indeed low Z-score balls. In this way the plots are suggestive that the original Altman (1968) model does identify potential failures in 2015. However, the model states that all those with Z-scores below 1.8 should be considered as being in the "distress" zone and it is clear that it is only a subset of these firms that did go on to fail. To diagnose the source of this colouration according to the five axis variables may be undertaken.

Correspondence between the failure proportions and the values of X_1 is much closer, both of the balls with higher proportions of failure correspond to higher values of X_1. In (1) the coefficient on X_1 is small and so it has a much smaller effect on the Z score. There is some evidence in these plots that a higher coefficient would be beneficial to represent the 2015 data. X_2, shown in panel (d), is higher towards the top. There is some correspondence with the overall Z-score, but where the latter is low on the top left X_2 is not. Explaining the low Z-Score in the top left is best done by looking at X_3. Indeed panel (e) confirms this. Panel (g), X_5 also has a strong correlation with the Z-Score with higher values in the top right. Panel (f) shows a much greater diversity of spread for X_4, only in the top left of the plot is any consistency observed. Failure in 2015 appears to be most associated with X_1, but again there are a lot of high values of X_1 where failure is not seen.

An immediate observation from the 2015 data is that there are many stories behind the data which the linear discriminate models are not bringing out. Indeed the use of the variables without interactions would still not be able to make the discriminations suggested by Figure 3.

5.2 Time-Variation?

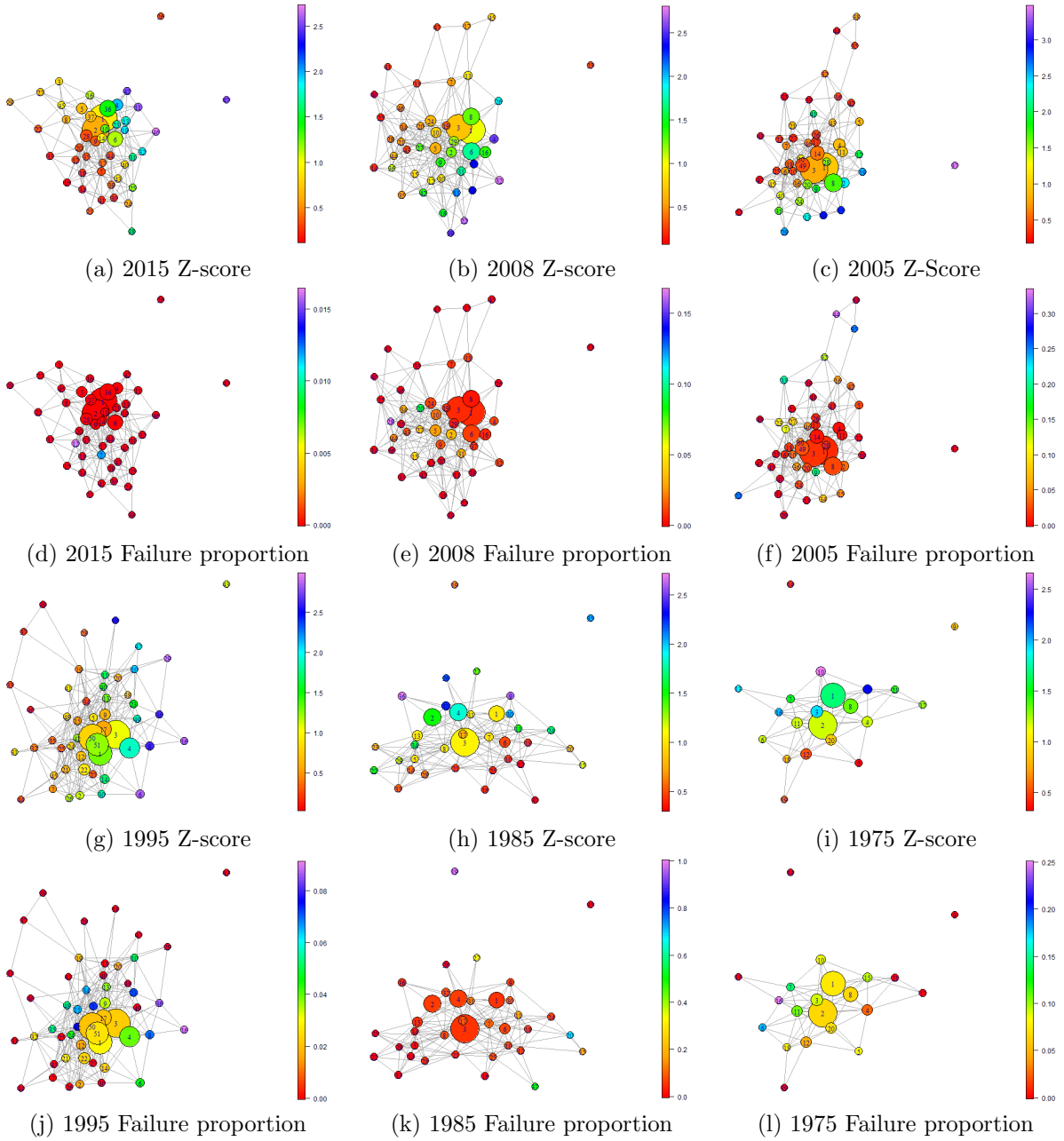
Appraising the fit of the original model in earlier years, Figure 4 has the Z-score and failure proportion plots for 1975, 1985, 1995, 2005 and the height of the global financial crisis in 2008. All plots have a similar lattice format to the 2015 case. However, a narrower, longer, shape in 2005 suggests that the data was more correlated that year. It has already been noted that a strength of the TDA Ball Mapper approach is that it can continue to be applied in cases like this.

Panels (a) and (d) of Figure 4 show the Z-scores and firm failure proportions for 2015 and are included for reference. Interest begins with panels (b) and (e) which show these two outcomes for 2008. This was the start of the global financial crisis so theoretically may have the most surprising exits from the Compustat database. Compared to the 2015 plot only a few firms are obtaining the highest Z-scores. The region coloured green, yellow, orange and red, covers the majority of the space. The largest balls in the plot are also in the distress region according to the Altman Z-score. Failures are indeed seen across the space, but the larger proportions are concentrated in the lower centre of the graph. Again this informs that an interaction between the variables will be better to identify where exactly exiting the Compustat listing will occur.

For 2005 panels (c) and (f) reveal a similar story of the "distress" zone covering a much larger proportion of the space than the other levels. There are failures in the lower part of the plot that correspond with the high Z-scores and low Z-scores. The most intense of the failure proportions appear in the top of the plot, far from the high Z-score end. Panels (g) and (j), plotting 1995, show that there is a smaller coverage of balls with low Z-scores. Failure proportions in panel (j) here correspond more with the high Z-scores to the right of the plot; the Altman (1968) model does not perform well for 1995. By contrast panels (h) and (k) for 1985 have the failures primarily concentrated in the bottom right, an area with very low Z-scores. There are also low failure proportions within the biggest balls, here again the average Z-score is well below the 1.8 cut off for the "distress" zone. Going back through time the same filtration produces fewer balls, the 1975 plots in panels (i) and (l), are particularly simplified relative to the others. Here the failures sit to the left of the plot in an area where there were some very high Z-scores noted.

Overall the TDA Ball Mapper plots have usefully shown that the firms that failed have characteristics in, or around the boundary of, the "distress" zone. This will explain the high accuracy of prediction from the Altman (1968) model. However, there are also many cases where the failed firms sat in areas of the plot where Z-scores were high and financial distress was not expected. Two important messages thus emerge. Firstly there is a need to split the "distress" zone using the interactions between variables. Secondly, non-linearity between the financial ratios and outcomes apply across the space.

Figure 4: Altman Z Scores and Firm Failures: Original Model



(a) 2015 Z-score (b) 2008 Z-score (c) 2005 Z-Score

(d) 2015 Failure proportion (e) 2008 Failure proportion (f) 2005 Failure proportion

(g) 1995 Z-score (h) 1985 Z-score (i) 1975 Z-score

(j) 1995 Failure proportion (k) 1985 Failure proportion (l) 1975 Failure proportion

Notes: BM plots generated using the Dłotko (2019) algorithm as implemented in *BallMapper* (Dlotko, 2019), for the original Altman (1968) model. Axes are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). Panels (a) to (c) and (g) to (i) are coloured according to the z-score calculated by equation (1). Panels (d) to (f) and (j) to (l) are shaded according to the proportion of observations within the ball that did fail. Figures are available in colour in the on-line version of the paper.

Table 4: Choice of Ball Radius and Ball Numbers

Ball Radius (ϵ)	1975	1985	1995	2005	2008	2015
0.25	683.1	1608	2391	2374	2008	1836
	(7.11)	(8.43)	(9.54)	(8.70)	(7.54)	(7.72)
0.50	182.6	545.1	890.8	995.1	790.1	786.3
	(3.73)	(6.33)	(7.61)	(7.42)	(7.43)	(6.85)
0.75	84.36	255.2	424.8	488.7	376.6	397.4
	(2.71)	(4.71)	(6.13)	(6.37)	(5.56)	(5.53)
1.00	51.34	144.4	238.8	272.6	211.8	224.9
	(2.17)	(3.46)	(4.76)	(4.82)	(4.60)	(4.78)
1.50	25.29	64.45	100.9	117.3	85.55	94.74
	(1.42)	(2.46)	(3.08)	(3.36)	(2.56)	(3.10)
2.00	15.8	35.34	53.15	60.6	49.99	51.77
	(1.10)	(1.88)	(2.16)	(2.22)	(2.06)	(2.36)
2.50	10.73	23.23	34.45	40.08	32.45	33.62
	(0.97)	(1.52)	(1.69)	(1.92)	(1.84)	(1.84)
3.00	7.99	16.83	23.97	27.99	22.98	22.96
	(0.93)	(1.24)	(1.54)	(1.62)	(1.51)	(1.52)
4.00	5.35	10.35	13.69	15.33	13.31	13.58
	(0.63)	(0.98)	(1.10)	(1.20)	(1.06)	(1.06)
5.00	3.91	7.7	9.81	10.4	9.17	9.78
	(0.41)	(0.75)	(0.81)	(0.96)	(0.78)	(0.80)

Notes: Numbers report the average number of balls estimated on the data from the stated year at the ϵ given in the first column. Figures in parentheses report the standard deviation from the 1000 estimates for each ϵ-year combination. Axes are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). All BM graphs constructed using the BallMapper R package of Dlotko (2019).

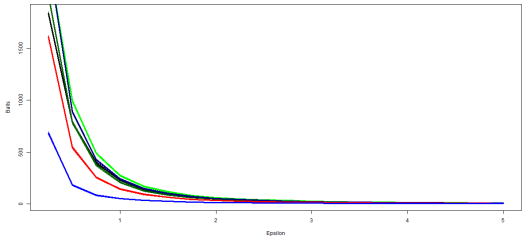
5.3 Role of Ball Radius

In the preceding analysis we chose ball radii, ϵ, values that optimised the appearance of the graphs, trading off the accuracy afforded by low filtration with the benefits of understanding connectivity that come from higher ϵ. In this section we evaluate the role of ϵ on the number of balls plotted, doing so for each of the six years that have been considered in this paper. In order to demonstrate the consistency of estimation we produce 1000 graphs for each ϵ. First Table 4 reports the numbers of balls for selected ϵ together with the standard deviation from the 1000 estimates. Figure 5 then features two plots of ball numbers and three BM graphs with differing ϵ. On these plots 95% confidence intervals are added as thinner lines based upon the 1000 estimates.

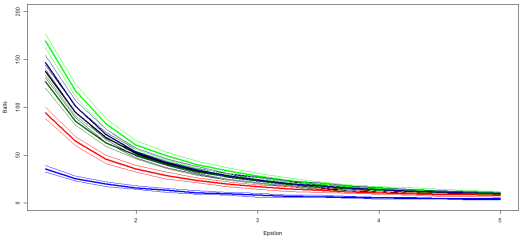
Our evidence shows how the number of balls quickly falls for the early increases in ϵ. The pattern is very similar across the years, but there are fewer balls in the earlier years owing to the lower number of firms and smaller value ranges. Table 4 shows this through the reporting of average numbers of balls for each ϵ-year pair in the columns relating to each year that has been considered in this paper. At the lowest level of ϵ considered there are as many as 2400 balls, but by $\epsilon = 1$ this number is just 10% of that figure. Figures in parentheses below each mean are the standard deviation from the 1000 estimates of each ϵ-year pair. In showing these numbers graphically in Figure 5 we present two plots, one with the full range and one which limits to $\epsilon \in [1, 5]$ to make the differences in this range more readable. In these plots the thick black line is 2015, whilst the remainder of the lines are drawn using 1975 (blue), 1985 (red), 1995 (navy blue), 2005 (green) and 2008 (dark green). We also provide thinner 95% confidence bands around the means in the same colour.

It should be noted that the construction of the graph is data dependent and so the variations between years reflect the different distributions of the data on each of the five axes, the joint distributions, and the number of observations that were included. Figure 5 also graphically illustrates the problem of a low ϵ by showing a value which renders the diagram unreadable. Moving to $\epsilon = 2$ in panel (d) produces a better plot, whilst $\epsilon = 4$ in panel (e) does not offer much information about the space due to the low number of balls.

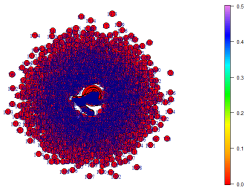
Figure 5: Choice of Ball Radius and Ball Mapper Graphs



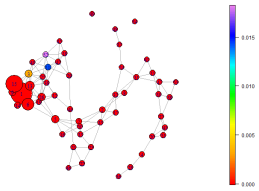
(a) ϵ and Ball Numbers



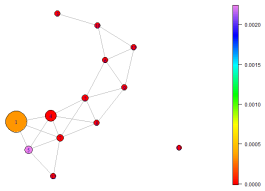
(b) Reduced ϵ range



(c) $\epsilon = 0.25$



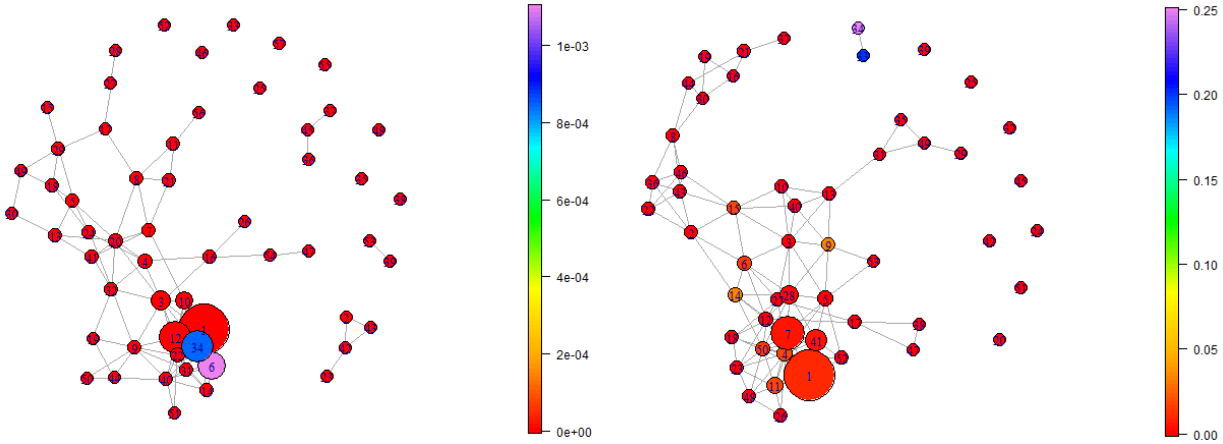
(d) $\epsilon = 2$



(e) $\epsilon = 4$

Notes: Panels (a) and (b) provide a plot of the number of balls produced by the BM algorithm as implemented in Dlotko (2019). In this plot the thick black line represents 2015, whilst the other years are shown as 1975 (blue), 1985 (red), 1995 (navy blue), 2005 (green) and 2008 (dark green). Thicker lines are the estimated means, whilst thinner lines of the same colour represent 95% confidence intervals from the 1000 estimations of each ϵ-year combination. Panel (b) is the same plot but with $\epsilon \in [0, 1)$ omitted. Below this we show the effect of ϵ on the appearance of the BM graph for 2015. Panel (c) shows what happens when $\epsilon = 0.25$, panel (d) when $\epsilon = 2$ and panel (e) when $\epsilon = 4$. Axes are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). A colour version of this plot is available in the online version of this paper.

Figure 6: Firm Failure Lagged Characteristics



(a) 2015 Failure Proportion (b) 2008 Failure Proportion

Notes: BM graphs constructed using the algorithm of Dłotko (2019) via the *BallMapper* R package Dłotko (2019). Axes are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover) with a further five axes added to capture the one year lag of X_1 to X_5. Balls are coloured according to the proportion of firms within the ball who failed in the subsequent year. This figure is available in colour in the online version of the paper.

As noted in the theoretical exposition, there is no algorithm which dictates the number of balls that is optimal and so it is left to the researchers discretion as to which value to continue with. Constructing graphs like that in Figure 5 may be informative, but it is better to focus on the informativeness of the BM graph as the main decision factor.

6 Modelling Extensions

From the BM analysis of Altman (1968) model it is apparent that the location of a firm within the point cloud is more important than linear classification permits. A beauty in the Altman (1968) representation is that it is simple to implement without relying on large amounts of additional data; given the financial data of just one firm an expectation on the likelihood of failure may be formed. This paper posits that identification of a firm's location on a BM map is a useful way to extend the analysis toolkit. However, the importance of location within the cloud also directs towards modelling extensions that may better fit firm failure. This section takes a brief look at two of these.

6.1 Historic Data

Logic dictates that the route to failure would typically involved periods of poor performance in which the firm manages to fight closure off. In terms of the model such would mean being in the distress zone for a number of years prior to failure. Alternatively there may be firms performing well who suffer large setbacks and fail quickly; such firms would not have been in the distress zone. Evaluation of the informativeness of past values of the X variables in the Altman (1968) model may be made by simply adding more axes to the BM plots. For brevity this extension is considered solely with the 2015 data.

As Figure 6 demonstrates, the inclusion of lagged financial characteristics increases the correlation between many axes and results in a shape that is longer and thinner than the equivalent in panel (b) of Figure 3. Continuing what was seen previously, failure is concentrated to just a smaller subset of the space, now within balls close to the main mass of the data. That these are so close to the main mass makes it more challenging to rule upon which firms are suitable for lending. Examining the data for other years reveals similar concentration, although typically this is in parts of the map that are not so proximate to the main mass. Panel (b) shows similar for 2008, showing again how a concentration of failed firms within the space

is achieved. 2008 also has failures whose performance had been very different to others, manifesting as two connected balls that are not joined to any other group, which 2015 did not.

From a practitioner angle the incorporation of further information has much to commend it, especially as this is data that is openly available to those assessing creditworthiness. However the extension into more lags requires more processing and does not generate a particularly large advantage over the previous use of just one year of data. Further we caution that the red colouration does not necessarily imply zero failure. Firms on their way to failure may simply fall quickly into the "distress zone" and therefore highlighting their existence is more effective if just using one year of data. Whether adding historic data, or otherwise, the ability of BM to represent the space neatly offers much.

6.2 Nearest Neighbours Modelling

A second potential addition to the set of X values for a firm is information about the performance of its neighbours. For this the proportion of failure amongst the firms within a fixed radius, or the failure proportion of the nearest k firms in the space may be used. Neither measure particularly accounts for the relative sparsity of a region so a scaling is performed based on the average distance through the space to each neighbour. In all cases distances are euclidean.

Logistic regressions are performed using the values of X_1 to X_5 from year t along with the measure of firm failure within the nearest neighbours of each point as explanatory variables. The dependent variable is failure at time $t+1$. p_i denotes the probability of failure for firm i. As with the additional axes variables, consideration is given to the 2015 data only. For comparison $k = 10$ and $k = 50$ are used. The proportion of failures within the nearest k observations being referred to as $NE10$ and $NE50$ for $k = 10$ and $k = 50$ respectively. Whether the spatial information adds value is then simply a test of the fit of the two models. Three models are thus estimated:

$$log\frac{p_i}{1-p_i} = \alpha + \beta\mathbf{X_i} + \epsilon_i \tag{2}$$

$$log\frac{p_i}{1-p_i} = \alpha + \beta\mathbf{X_i} + \gamma_{10}\texttt{NE10}_i + \epsilon_i \tag{3}$$

$$log\frac{p_i}{1-p_i} = \alpha + \beta\mathbf{X_i} + \gamma_{50}\texttt{NE50}_i + \epsilon_i \tag{4}$$

Where $\beta\mathbf{X_i}$ collects the vector of coefficients and the observed values of the five financial ratios for firm i. ϵ_i is an identically independently distributed error term with mean zero and constant variance.

Table 5 reports significant effects for the nearest neighbour estimates and shows that considering a firm's location in the point cloud is very important. Compared to Model 1, the likelihood ratios of Model 2 and Model 3 show significant movement towards zero. The Aikaike information criteria reduces notably moving from Model 1 to Models 2 and 3. Hence it is immediately concluded that adding nearest neighbours improves fit. This confirms the importance of spatial positioning that motivates this paper but does not mean that the points must be in a large connected component. Of the two models with nearest neighbours terms Model 2 is favoured, further evidencing the fact that we are not talking about a single connected set. In studying logistic regressions it is usual to consider marginal effects, or odds ratios, but the nature of the two nearest neighbour measures means that such effects are of a high order of magnitude, larger than those for the X_1 to X_5. Intuition would suggest that these spatial effects would dominate, but the order with which they do reaffirms the value of understanding the point cloud in every dimension.

7 Implications and Discussion

This paper seeks to evaluate the effectiveness of credit modelling in predicting firm failure, the lens applied making it possible to identify segments of the parameter space in which failures occur and, through the functionality of the *BallMapper* package, to identify the intensity of the failure rate in any given part of the space. In illustrating the power of the approach a simple dataset just using the five financial ratios of Altman (1968) was used. Therein lie a number of stories that create non-linearity in the link between the characteristics and failure; failures are not distributed evenly across that part of the space where the

Table 5: Logistic Regressions for Firm Failure: Nearest Neighbours Models

	Model 1	Model 2	Model 3
X_1	1.220	0.650	0.802
	(1.941)	(2.342)	(2.229)
X_2	0.010	0.006	0.114
	(0.158)	(0.263)	(0.615)
X_3	−0.285	−0.075	0.834
	(0.357)	(1.127)	(1.369)
X_4	−2.537	−1.261	−1.113
	(1.692)	(1.633)	(1.566)
X_5	0.235	0.151	0.105
	(0.162)	(0.369)	(0.214)
$NE10$		43.815***	
		(13.311)	
$NE50$			180.5***
			(65.658)
Constant	−5.661***	−7.400***	−7.755***
	(0.935)	(1.628)	(1.825)
Observations	5,057	5,057	5,057
Log Likelihood	−21.00	−13.27	−14.60
Aikaike Information Criteria	54.00	40.53	43.20

Notes: Logistic regression coefficient estimates for the roles of the five Altman (1968) variables in predicting firm failure. In all cases the dependent variable is an indicator which takes the value 1 if the firm fails in the subsequent year. These predictors are Axes are X_1 (liquidity), X_2 (profitability), X_3 (productivity), X_4 (leverage) and X_5 (asset turnover). $NE10$ is a variable which records the proportion of failed firms within the nearest 10 data points to each firm. $NE50$ is likewise but for the nearest 50 firms. Model 1 features only X_1 to X_5, with Model 2 including the average failure rate within the nearest 10 points and Model 3 containing the average failure rate within the nearest 50 points. Significance denoted by *p<0.05; **p<0.01; ***p<0.001

Altman (1968) model suggests financial distress. An overwhelming message from the analysis is that there is non-linearity and a clear importance to looking at the interactions between variables.

Altman (1983) suggested that the Z-score can be a useful means of informing managers about the ways in which they may turn their business around. The BM graph informs where a company sits in the space and hence where the nearest balls that are not associated with failure are. Using the information in the BM graph it is straightforward to look at which variables would have placed the firm in a safer part of the space and hence to understand the most effective route away from potential failure. In most instances this will mean moving along one of the edges of the graph since already there are overlaps that imply proximity of the two ball centres. It may be that such movements increase the Z-score.

Causality is not discussed within the context of the BM graph. What is seen is the actual occurrence of firm failure in the year after the Compustat financial ratios are observed. That there are no failures in a ball in a given year does not mean that it would have been impossible for a firm in that ball to fail. Likewise it is certainly not the case that any firm that finds itself in a ball where there have been failures will itself fail. In the examples provided here there is one case where a ball has 100% failure, but typically the rates are less than 25%. It can then be considered whether to build a causal model using the inference from the BM graph to guide variable inclusion.

Motivation for the use of BM came from the non-linearity criticisms of the MDA papers after Altman (1968). Linearity is also a feature of many of the probabilistic models that follow Ohlson (1980) also. By plotting the space using BM there is clear evidence that the failure outcome is not distributed evenly across the space in the way that a linear function, like the Z-score of Altman (1968) is. Including all of the interaction terms would identify the differences in the space and may be the next step for the construction of a better model. However, such inclusion of the full set of interactions becomes much harder as the original number of variables increases. To have all the interactions for the five axes used here necessitates a further 24 variables on top of the original 5. As the number of explanatory factors increases so there is a rapid increase in the number of parameters to be estimated. Depending then on the complexity of the modelling such large volumes of unknowns may be problematic for simultaneous explanation.

Two extensions were considered from the intuition of the BM approach; these offer indicative channels down which research could extend. Adding data is always liable to improve fit, and hence the BM plots with an additional year of data do show a clearer identification of failures as stemming from a particular zone. Because BM offers a representation of the full data space, the observation that bankruptcy is localised within that space suggests adding a measure which considers the neighbours of a firm. Our second extension used nearest neighbours and found the proportion of failures there amongst. A favouring of the immediate vacinity, 10 neighbours, over the wider reach of 50 neighbours suggests a fruitful direction for research. Our concern in advocating such an approach is that it requires further analysis beyond what is done in the basic Altman (1968) model. Representation of localised failure within the region of a point would do little to alter the BM graph and hence it is better to simply assess a candidate firm based upon its location on the BM graph constructed in this paper, with failure as the colouring variable.

Simple models additionally struggle when the data that is being used is highly correlated. Post normalisation the BM graphs do not have the narrow shape associated with strong correlation but, as seen in the summary statistics, the data does show strong correlation between some of the variables. Were the set of explanatory ratios to be expanded further then potential multicollinearity is a bigger issue. Consequentially the benefits of using BM would, as with the way that the number of interactions increases with potential factors to include, become more pronounced.

Contemporary growth in credit modelling has been driven by ML, seeking to bring the information from data through the search of specific patterns, classifications or simply the ability to fit non-linear functions. Motivating these advances is an improved fit that comes from being able to use models that are much freer in their functional form. Recent advances in ensemble learning and the use of synthetic factors have served to improve model fit even further. However all techniques have been subjected to criticism for their "black-box" nature because the precise details of the model fitting are not available in the way that the MDA and logistic models are. Classification models would aim to segregate the distress zone to provide relationships under which the failure regions shown in the BM plot might be split from other areas of the distress zone. As noted in Section 4.3 dividing the space may require non-linearity and lead to over-fitting. By looking at the stark difference between the predicted Z-score and the observed failure, the BM graphs have shown that to get a better fit it is necessary for these ML approaches to do a large amount of deviation from the linear models.

Hence the extent to which the unobservable functional form is relevant is more pronounced.

From a practitioner perspective the evidence presented here can help with either the decision to lend, or in the consideration of how best to escape from financial trouble. On the former a potential lender could identify where in the space the applicant for credit is, and therefore make a decision on how likely that candidate is to slip into distress. Presuming lenders gain profit from all successfully repaid loans this would open up huge parts of the "distress" region of Altman (1968) that would actually be solid opportunities. This is the motivation of the softspace clustering algorithms discussed in Liu et al. (2019); in that literature strand clustering algorithms are used to segment the space and different models are constructed for each cluster. Producing models for each ball would be an option but brings back the danger of over-fitting past data. Developing this strand, whilst maintaining the necessary generalizable results, will be an important next step. Having a location on the map for a candidate firm permits manual analysis of risk given positioning in the data cloud and remains an important, model free aid for practice.

8 Conclusions

Credit scoring models in the spirit of Altman (1968) seek to segment firms based upon the values of key financial ratios, identifying those in most danger of default. We may readily think of each ratio as an axis and the co-ordinates of a firm's location in that space as being defined by the ratio on each axis. Linear discriminant analyses like these necessarily treat each ratio independently and hence identify an enclosed part of the characteristic space in which linear criteria are satisfied. Segmenting in this way is arbitrary and has led to a number of non-linear classifiers being applied. Imposing such classifiers risks over-fitting as it makes claims about the shape of the data that may easily be time specific. An ever smaller number of observations are driving notions of causality with obvious danger for inference. Consequentially, adoption of the simple Altman (1968) form remains strong (Altman et al., 2017a). This paper exemplifies how contemporary data science may bring transparency in a world of "black box" models.

Introducing topological data analysis, and specifically the BM algorithm of Dłotko (2019), to simple datasets on financial defaults offers much to understanding the contribution of individual financial ratios to liquidation and bankruptcy. Mapping the financial ratio characteristic space it is shown that failure only occurs within a subset of the space. Representing what is there, rather than making any claim of causality, empowers BM as a tool. This paper has shown that interactions between liquidity, profitability, productivity, leverage and asset turnover should be explored further. BM has the ability to signpost exactly where firms are in the space, how close they are to areas where failure has been observed, and how we might then understand the decision to give credit based thereupon. Placing firms on the "map" is an important step to evaluating creditworthiness. Through two extensions we demonstrated the localised nature of failure and hence further underlined the value of mapping these localities. Because everything that appears in the plot is driven by the data there is no "black-box" criticism to any of the results that emerge; this sets BM apart from the machine learning literature which has yet to be fully trusted in practice.

A number of potential extensions emerge, with applications to other datasets, addition of further axes and consideration of wider time-frames being obvious next steps. However each of these is just a small increment relative to the demonstration of the power of the method. BM graphs are a map of the data cloud and hence have potential to guide the segmentation thereof; this may be a fruitful line of enquiry for subsequent research if it can be understood that there are no chances of failures occurring in any disregarded part of the space. At this stage shrinkage of the data cloud on the back of not having observed any bankruptcy outcomes would be premature. Notwithstanding, BM represents a new system that offers a great amount to the study of credit, moving us beyond the linear segmentation of space in a transparent and information driven way. This paper takes a critical first step on that representational journey.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

Altman, E. I. (1983). *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy*. New York: John Wiley & Sons.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., and Suvas, A. (2017a). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management & Accounting*, 28(2):131–171.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., and Suvas, A. (2017b). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management & Accounting*, 28(2):131–171.

Barboza, F., Kimura, H., and Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417.

Beaver, W. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4:71–111.

Beaver, W. (1968). Alternative accounting measures as predictors of failure. *The Accounting Review*, 43:113–122.

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107.

Choi, H., Son, H., and Kim, C. (2018). Predicting financial distress of contractors in the construction industry using ensemble learning. *Expert Systems with Applications*, 110:1–10.

Cortes, C. and Vapenik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

De Bock, K. (2017). The best of two worlds: Balancing model strength and comprehensibility in business failure prediction using spline-rule ensembles. *Expert Systems with Applications*, 90:23–30.

Dłotko, P. (2019). Ball mapper: a shape summary for topological data analysis. *arXiv preprint arXiv:1901.07410*.

Dlotko, P. (2019). *BallMapper: Create a Ball Mapper graph of the input data*. R package version 0.1.0.

Gidea, M. and Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications*, 491:820–834.

Li, H., Sun, J., Li, J. C., and Yan, X. Y. (2012). Forecasting business failure using two-stage ensemble of multivariate discriminant analysis and logistic regression. *Expert Systems*, 30(5):385–397.

Liu, C., Xie, J., Zhao, Q., Xie, Q., and Liu, C. (in press 2019). Novel evolutionary multi-objective soft subspace clustering algorithm for credit risk assessment. *Expert Systems with Applications*, 138.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nicolau, M., Levine, A., and Carlsson, G. (2011). Topology based data analysis identifies a group of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 107:7265–7270.

Niyogi, P., Smale, S., and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18:109–131.

Pearson, P., Muellner, D., and Singh, G. (2015). *TDAmapper: Analyze High-Dimensional Data Using Discrete Morse Theory*. R package version 1.0.

Petropoulos, A., Chatzis, S., and Xanthopoulos, S. (2016). A novel corporate credit rating system based on student's-t hidden Markov models. *Expert Systems with Applications*, 53:87–105.

Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100.

Son, H., Hyun, C., Phan, D., and Hwang, H. (in press 2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications*, 138.

Vejdemo-Johansson, M., Carlsson, G., Lum, P. Y., Lehman, A., Singh, G., and Ishkhanov, T. (2012). The topology of politics: voting connectivity in the us house of representatives. In *NIPS 2012 Workshop on Algebraic Topology and Machine Learning*.

Ziba, M., Tomczak, S. K., and Tomczak, J. M. (in press 2019). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*.