UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Engineering, Built Environment and IT

Department of Computer Science

MIT C (Big Data Science)

MIT 808

Experimental Process Document

15 June 2019

Marea Gwyneth Sing

u10500449

Mentor: Dr Conrad Beyers **Signature:**

Co-mentor: Mrs Anna Bosman **Signature:**

Identifying and Analysing Patterns of Financial Crises Using Self-Organising Maps

Contents

1 Introduction

This research project aims to use self-organising maps (SOMs) to determine if the standard economic variables used in building early warning systems (EWS) for financial crises is generally supported by data, and whether any additional variables may be useful in improving the accuracy, precision and recall of an EWS. In particular, this project uses the ECB/ESRB EU crisis database which covers all financial crises in all in the European union during the period 1960 to 2016.

A more detailed description of the project's problem statement, goals, scope and objectives, is available in the project proposal (assignment 1) and requirements document (assignment 2). For informational purposes, these are re-stated in brief below:

1.1 Problem Statement

Given the complex, non-linear nature of financial crises, it has not been ascertained whether all useful variables (for improving accuracy, precision and recall) are identified in the economic literature. It has also not been ascertained whether these variables are generally consistent across countries and time periods.

1.2 Goals, Scope and Objectives

Therefore, the goal of this project is to confirm the variables selected in the extant economic literature and/or to identify new variables that may prove useful for building an EWS. In addition, the project aims to describe how these variables may/may not differ over time and per country.

The scope of this project is to build two types of self-organising maps (SOMs) to address this goal, as this form of machine-learning is known to uncover correlations between large datasets that may not otherwise be easily identifiable.

1. A supervised SOM:
   One produced per country, to identify a country's financial stability distribution over time, together with which variables appear to indicate periods of imminent and ongoing financial crisis.

2. An unsupervised SOM:
   One produced per country, to analyse the similarity in distribution of countries, and

possibly to show how countries begin clustering together on a map during periods of financial distress (an indication of contagion effects).

Finally, this project will be conducted in such a way as to improve monitoring and collaboration, and to eventually create publicly available, reproducible research.

2 Implementation Details

Given the supplementary objective of creating publicly available, reproducible research, it was decided to write this project in an open-source software language. Specifically, the project is coded in R since the language is widely-used, and high-quality existing packages exist for implementing machine-learning models, such as Random Forests (RF) and SOMs.

For the RF models and data splitting and pre-processing, the caret (Classification And REgression Training) package [4] will be used, whilst for fitting SOMs the Kohonen package [10] will be used. The former also allows for variable importance estimation (from the RF model), while the latter is able to fit both supervised and unsupervised SOMs.

All data, code, and results will be uploaded to a private GitHub repository, which is linked to a project, and shared with project mentors. This allows for the monitoring of progress and general project management, and also provides a simple platform for making the project publicly available in future.

3 Experimental Process

3.1 Introduction

This project aimed to use newer machine-learning approaches to analyse systemic financial crises in the EU over the period 1970t0 2016. In particular, it proposed (1) to use a supervised SOM to create economic variable "feature planes", in order to determine whether any variables indicate periods of imminent/extant crises; and (2) to use an unsupervised SOM, to determine (per period) which countries were similarly distributed, and if they become even more so during periods of financial distress (i.e. the systemic/contagious nature of crises).

This section will explain in more detail what a SOM is, and how it was fit to the data in th is project.

3.2 SOM Analysis

A SOM is a simple feed-forward neural network, with only one input layer and one output layer. The outcome of a SOM is a dimension reduction technique similar to principal component analysis or k-means clustering, but one that tends to perform better in non-linear environments [2], making them particularly applicable to analysing systemic financial crises.

A SOM is estimated iteratively, passing each data observation to the neural network a number of times (parameter: *rlen*). For each observation, its best matching unit (BMU) is found in the output layer, which is then "moved" towards the input vector, along with its neighbours (albeit to a lesser extent). The iterative process then adjusts the position of these nodes delivering a final result which is a type of clustering where the topographical features of the input dataset is preserved. This allows for a unique visual interpretation of how, in this case, different countries or time periods are mapped; as well as, through the feature planes, how each economic variable is distributed over the map.

From this brief explanation, it should be apparent that a number of parameters and hyper-parameters need to be defined:

- The **size and shape** of the map is determined by the number of output nodes and how they are arranged.

  A general rule of thumb for selecting the number of output nodes is presented in [9]. It suggests that the optimal size of the SOM grid can be approximated by $5 * \sqrt{n}$, where n is the number of observations in the sample. This is the rule employed in this project, although, should this map prove to be too large for some periods for which less data is available, the constant (5) will be reduced until no further issues are encountered.

  As far as the shape of the map is concerned, in the extant literature, a hexagonal-shaped grid is preferred over a square-shaped grid, as it allows for more neighbours for each node. The entire grid may then be shaped in a number of ways, of which the rectangular shape is the simplest and most common [1]. As stated by [1], the rectangular arrangement is more appropriate when there are more extreme observations in the sample. These will then be mapped to the outer edge of the map, as the extreme nature of the observations will necessitate fewer neighbours. This is what we expect when applying this methodology for analysing financial crises. Thus, a hexagonal-cell, rectangular-shaped map is employed in this project.

- The size of the **neighbourhood**

  The size of the neighourhood affects how the map is fit to the data. In the limit, a neigh-

bourhood of zero results in standard k-means clustering, whereas a large neighbourhood results in a stiff map that maintains the topology of the data, at the cost of quantization error (how well each node fits the observations, on average) [6]. This project makes use of the kohonen package in R, which allows for a linearly decreasing neighbourhood radius, starting at a value that "covers 2/3 of all unit-to-unit distances"[11]. That is, the map is first fit prioritising topology over quantization error, and linearly produces more fine-tuned results as the model iterates through all the observations.

- The measure of **distance**
  In general, the Euclidean distance is the most well-known and widely-used measure of distance. However, in this project the Euclidean squared distance (sum of squares) is used, as this reportedly increases efficiency in clustering problems [3].

- The **learning parameter**, α
  The learning rate in a SOM determines how close the BMU (and its neighbours) move towards the input vector. In this project, this learning parameter is set to decline linearly from 0.05 to 0.01. This allows BMU nodes to initially move quickly towards the input vector, and then slows down this adjustment (fine-tuning).

- The **number of times** the data is passed through the map
  In general, the number of times the data is passed to the model is determined by analysing whether the mean distance to the closest BMU for each observation has stabilised at a apparent minimum. Based on previous SOM analysis, this project will start with an *rlen* value of 500, and will increase it if necessary.

In addition, each map will be specifically coloured in order to aid map evaluation. Each map will be coloured by grouping: this grouping is generally undefined in the unsupervised SOM, and is grouped according to estimated financial state[1] in the supervised SOM.

As outlined above, the dynamic nature of the parameters, and the generally accepted starting and ending points for each, allows this analysis to be conducted on the entire data set (unsupervised SOM) or on a simple training-test split (70-30, supervised SOM). That is, no specific validation set is used in this project[2].

Using these specifications, two types of SOMs will be employed on annual data for the EU covering the period 1970-2016. This deviates from previous assignments, where it was

[1]The financial state is either "0", normality, or "1", crisis. However, any values produced by the map between these extreme values will be identified as "pre-crisis" nodes.

[2]This deviates from the previous expectations outlined in past assignments.

envisioned that quarterly data would be used, however, this new approach will continue to meet the functional and non-functional requirements of this project. Another minor deviation from the project proposal, is that the data used will be loaded as percentage changes, whereas it was previously suggested that the data would be loaded in levels and transformed using log differences. This process is equivalent at a first-order approximation.

In addition, it should be noted that due to the use of annual data, missing values are no longer interpolated as they tend to occur at the beginning or end of a data series. Instead, the only requirement for using a variable in either SOM is that it is not missing for more than half of the sample being considered. The SOM algorithm is then adjusted accordingly, to allow for a similar threshold of missing data.

It is also worth noting explicitly here, that while the data is centered and scaled, as this tends to improve the performance of neural networks in general [7], the data used is centered and scaled per country. This is as EU countries are in at different stages of development and will consequently have systemic differences in macroeconomic variables, for example, GDP growth and inflation is generally higher in developing countries. If this is not accounted for, it might bias the results obtained for this project. Furthermore, the centering and scaling of the data is conducted independently on training and test sets to ensure that information from each set does not "spillover" to the other, which may also bias the results obtained.

- The Supervised SOM

  This SOM will use a merged dataset: a number of world bank indicators per country (identified by country and time period[3]), together with the ECB/ESRB database on financial crises. In general, the macroeconomic indicators cover broad areas of the economy, such as GDP, savings, trade, debt and the labour market.

  It follows the general approach of [6], and will produce maps that can be identified as crisis, normal or pre-crisis periods. The underlying feature planes for each economic variable can then be used to understand how each economic variable is distributed across these periods, and whether there are any apparent relationships (correlations) that hold between variables during these periods.

- The Unsupervised SOM

  This SOM will only use the world bank indicator dataset, but will estimate a mapping for each year in the sample.

  This more generally follows the approach of [2], producing a map per time period that

[3]This deviates from previous expectations, as annual data does not allow for a more disaggregated per-country analysis at this stage.

allows for the visualisation of how similar countries are, during years of tranquility and years of crisis. This may also show, when comparing annual maps, how countries become more similarly distributed during periods of financial contagion. In addition, the feature planes will allow for the analysis of which economic variables are distributed across these clustered countries.

Finally, as an additional evaluation metric (discussed below), a random forest (RF) model will be fit to the same dataset[4] used to estimated the supervised SOM. This will be used to calculate variable importance measures for the economic variable inputs. RF models require a number of parameters and hyperparameters to be set[5], the most important of which is the *mtry* parameter, which controls the numbers of features sampled at each tree split, and *ntree* which controls the number of trees in the forest. The number of trees is left at the default value of 500, while the caret package inbuilt option to try 10 different values for *mtry* is used[6]. The Gini index (*impurity* option) is used to calculate variable importance, as this is less computationally intensive than using a measure of entropy, and an applied study suggested that there is no conclusive evidence to suggest that either measure is superior to the other [5].

4 Evaluation Process

As discussed in assignment 1, there are a number of approaches that will be employed in this project to evaluate the fit of the SOMs:

1 Qualitative Methods
In addition to analysing the identified relationships through the lense of economic theory, each map can be visually assessed using three SOM-specific qualitative methods:

- For each SOM a "changes" plot is produced, which shows how the mean-distance to each BMU decreases as the model iterates through all of the dataset. As a rule of thumb, if this measure has not stabilised at a low value, the model requires more iterations to be run.

[4]The data will not be split into training, validation, and test sets, as the purpose of this model is not to test the RF model for predictive power.

[5]The majority of these are left at the caret and ranger packages default values.

[6]The recommended default for this parameter is $floor(sqrt(ncol(x)))$, where x is the input matrix.

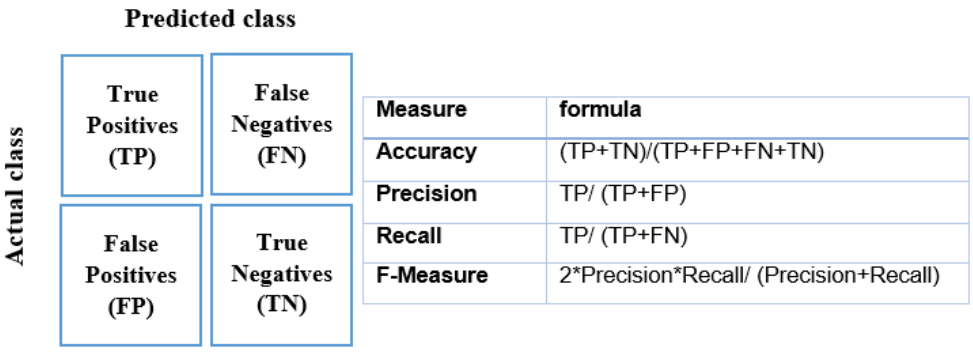
– For each SOM a "unified distance matrix" plot is produced, which shows how closely clusters are distributed. This will effectively show how significant the clusters identified by the SOM are.

– For each SOM "quality" plots are produced. These maps show the average distance between each input mapped onto a particular node (output BMU), and the centroid of that BMU. It is in some sense, a more disaggregated view of the the "changes" plot, and shows how well the model fits the data.

2 Quantitative Methods

– As another measure of quality, each map can be analysed by its quantization error. This is conceptually linked to the "changes" and "quality" plots described in the previous section. It is calculated as the average distance from each input vector to it's BMU centroid.

– For the supervised SOM, the map can also be assessed using the standard confusion matrix, which produces measures such as accuracy, precision and recall. For this, a cut-off of 80% is used to classify the SOM predictions as a crisis. In reality, this threshold is going to have to be set by policy-makers according to their preferences between making type I (a missed crisis) and type II (a falsely predicted crisis) errors. Increasing this threshold will lead to more missed crises, while decreasing this threshold will lead to more false predictions.
Specifically, this confusion matrix and measures can be represented as:



Measure	formula
Accuracy	(TP+TN)/(TP+FP+FN+TN)
Precision	TP/ (TP+FP)
Recall	TP/ (TP+FN)
F-Measure	2*Precision*Recall/ (Precision+Recall)

Figure 1: Source: [8]

That is, precision in this application accounts for the number of correctly predicted crises as a ratio of all predicted crises, whereas recall represents the ratio of all correctly predicted crises to the number of actual crises.

3 Supporting Machine-learning Methods

- Finally, a random forest model can be used to confirm SOM findings, by producing a measure of variable importance. It should be noted that RF models, unlike SOMs, are not able to handle missing values. As a result, any missing values are removed from the dataset when estimating this model. While this is not ideal, it should not negatively impact the estimation of variable importance.

5 Addressing the Requirements

5.1 Functional requirements

The use of a private Github repository, shared with the project mentors addresses the functional requirements of (1) providing a shared space for collaboration and monitoring of progess; (2) the need of a version control system to record changes to model data, code, and results; (3) allowing for offline modifications that can be synced to an online shared space; and (4) providing a space for future researchers to access the project details and results.

The experimental and evaluation processes will address the requirements of (1) producing two types of SOMs; (2) evaluating the fit of these SOMs using a mixture of qualitative and quantitative measures; and (3) using a random forest model to verify input variable importance.

5.2 Non-functional requirements

The created Github repository also addresses the non-functional requirements of (1) providing a space to store old models, data, and results; and (2) ensuring that the information is accessible from all modern browsers.

Finally, the use of the R languages and associated packages meets the non-functional requirement of using open source software, which will enable sharing and the general production of reproducible research.

References

[1] Richard G Brereton. Self organising maps for visualising and modelling. *Chemistry Central Journal*, 6(2):S1, 2012.

[2] Smaranda Cimpoeru. Using self-organizing maps for assessing systemic risk. evidences from the global economic crisis. *Economic Computation & Economic Cybernetics Studies & Research*, 49(2), 2015.

[3] IOS. Euclidean and euclidean squared distance metrics. http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Euclidean_and_Euclidean_Squared_Distance_Metrics.htm. (Accessed on 06/10/2019).

[4] Max Kuhn. The caret package. http://topepo.github.io/caret/index.html, 2019 March. (Accessed on 06/03/2019).

[5] Laura Elena Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93, 2004.

[6] Peter Sarlin and Tuomas A Peltonen. Mapping the state of financial stability. *Journal of International Financial Markets, Institutions and Money*, 26:46–76, 2013.

[7] Murali Shanker, Michael Y Hu, and Ming S Hung. Effect of data standardization on neural network training. *Omega*, 24(4):385–397, 1996.

[8] Safae Sossi Alaoui, Yousef Labsiv, and B Aksasse. Classification algorithms in data mining. *International Journal of Tomography and Simulation*, 31:34–44, 08 2018.

[9] Juha Vesanto, Esa Alhoniemi, et al. Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3):586–600, 2000.

[10] Ron Wehrens, Lutgarde MC Buydens, et al. Self-and super-organizing maps in r: the kohonen package. *Journal of Statistical Software*, 21(5):1–19, 2007.

[11] Ron Wehrens and Johannes Kruisselbrink. kohonen.pdf. https://cran.r-project.org/web/packages/kohonen/kohonen.pdf, December 2018. (Accessed on 06/10/2019).