UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and IT

Department of Computer Science

MIT C (Big Data Science)

MIT 808

Assignment 1 - Project Proposal Document

31 May 2019

Marea Gwyneth Sing

u10500449

Mentor: Dr Conrad Beyers **Signature:**

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

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

Contents

1 Project Motivation **1**

2 Project Summary **1**

2.1 Problem Statement 1

2.2 Goals, Scope and Objectives 2

2.3 Assumptions 2

2.4 Limitations 3

3 Project Methodology **3**

3.1 The Software Process Model 3

3.2 Data to be used and the data pre-processing process 4

3.3 Platforms and infrastructure to be used 6

3.4 Experimental process 6

3.5 Evaluation process 7

4 Project Report Structure **8**

5 Project Plan **8**

1 Project Motivation

Systemic financial crises are known to have real, adverse impacts on the socio-economic well-being of not only the originating country's citizens, but also those whose financial systems are affected through cross-border contagion ([1]). In particular, systemic banking crises are found to increase the probability of both currency and sovereign crises ([4], [11]).

It has been shown that with timeous intervention by national governments, the severity and duration of this impact can be lessened ([1]). However, effective policy intervention requires the existence of a reliable early-warning system (EWS).

While the literature on the development of early-warning systems is fairly well established, the use of machine-learning in their development, and the field of economics in general, is still in its infancy ([3]). Recent contributions to the literature have suggested that it is these particular types of "policy prediction" problems, that hold the most promise for meaningful contributions from the application of machine-learning to macro-financial problems ([12], [2]).

Financial systems and their inter-linkages with the real macroeconomy constitute a complex dynamic system, replete with complex, non-linear economic relationships. Thus, the sources of banking crises are varied, and their transmission mechanisms through domestic and international financial markets are not as well-established in the economic literature. Gaytan et al. ([11]) classify the sources of banking crises into (1) macroeconomic epidemics, (2) microeconomic deficiencies, and (3) endemic crises. And while these categories broadly capture the variables needed to build early-warning systems, the variables employed in the existing literature varies from study to study - both in terms of inclusion and statistical significance.

Thus, one of the first steps in building an EWS is to determine which subset of variables, from a high-dimensional full information set, should be included in the final model specification.

2 Project Summary

2.1 Problem Statement

In general, the variables chosen to develop early-warning systems for financial crises are based on economic theory and known transmission mechanisms. However, given the complex, non-linear nature of these systemic events, and the importance of early policy intervention, it is

not well-established whether (1) this approach captures the full set of variables that could be used to improve the accuracy, precision, and recall of an EWS, and (2) this set of variables, or perhaps a subset thereof, are consistent across the considered countries and time periods.

2.2 Goals, Scope and Objectives

The goal of this project is to produce a rigorous data analysis report, using newer machine-learning techniques, to identify the variables that appear to signal imminent financial crises, or the nature of ongoing crises, and how these variables may change across countries and/or time.

This is in support of a larger research effort to establish the statistical superiority (or lack thereof) of a variety of methodological approaches to building an EWS, ranging from extant statistical approaches to newer machine-learning approaches.

Specifically, the scope of this project is two-fold, utilizing self-organising maps (SOMs) to:

1. Identify per country, which variables appear to indicate periods of imminent and ongoing financial crisis. (This analysis will employ a supervised SOM.)

2. Identify across time, which of the countries are most similar, and analyse the cross-border transmission of financial crises (that is, whether they cluster together on the map when a crisis occurs in any one of them). (This analysis will use an unsupervised SOM.)

In addition, this will produce a unique visual representation of financial crises in two-dimensional space that might assist in furthering the understanding of intra- and cross-border transmission mechanisms.

A final objective of this project is to conduct this analysis such that it can be easily monitored and edited by supervisors and co-authors, as part of a larger effort to create high-quality reproducible research. In future, this analysis may be shared publicly as a stand-alone, interactive data analysis application.

2.3 Assumptions

The explicit objective of this project is to venture beyond the set of variables usually analysed and identified in the literature when trying to understand and predict financial crises, deliberately reducing the impact of the assumptions implied by economic theory.

However, it is still assumed that the variables that may improve the statistical and practical performance of an EWS will come from standard macroeconomic databases.

In addition, while variation is expected across both countries and time, it is assumed that there exists a common core set of variables that indicate current or imminent financial crises. That is, both the supervised and unsupervised SOM algorithms require that there exists a common set of features that can be clustered, and thereby, mapped.

2.4 Limitations

One of the first considerations when building an EWS, is the definition of what constitutes a crisis (for example, financial, banking, currency, sovereign), and how to determine its start date, end date, and therefore, duration.

Given this requirement, and due to the limited availability of internally consistent databases, the scope of this project will focus this analysis on European countries only.

While the extension of this analysis to cover a more globally representative list of countries will be considered for future projects, it is noted from the outset that this project will produce analysis that may be biased to fit the macroeconomic structure and inter-linkages of the European Union. Consequently, issues of transfer learning fall outside of the scope of this project.

In addition, it is acknowledged that, as with all data analyses, the well-known saying of "correlation does not equal causation" applies. That is, this project is consciously descriptive and exploratory in nature.

3 Project Methodology

3.1 The Software Process Model

When defining this project's software development life cycle (SDLC) model, a few key considerations were used to determine which model would most accurately describe the process that would be implemented[1]:

1. The stability of the project's product definition and objectives

   This project forms part of a broader research agenda, and consequently, its scope,

[1]These were considered based on a number of formal ([8], [18], [16], [6]) and informal([9], [10],[13]) resources.

objectives and final desired product (a data analysis report) is well-defined and unlikely to change.

2. The nature of the project's user and system requirements
   The methodology that is to employed in order to achieve the project's objectives, has been clearly pre-defined and agreed upon during the planning (specification) phase of the broader research agenda.

   Similarly, the objective of ensuring the production of reproducible research, that can be easily monitored and edited by supervisors and co-authors, is unambiguous and fixed.

3. The size and length of the project
   Finally, the project is small (as reflected by the fact that it is to be completed by a single individual), and fairly short in duration (final submission by 15 June 2019).

Given these considerations, it was decided that this project could be best described by the waterfall model, when viewed in isolation. That is, the process steps of defining the requirements, designing the system and software solution to meet those requirements, implementing and testing that solution, and the operation and maintenance of the final product, are clearly separate and distinct. For example, even with the iterative nature of model building (including parameter tuning), this stage is self-contained, and will neither influence which variables are tested (implementation does not impact design), or the structure of the final reproducible, and publicly available, research report (implementation does not impact operation).

When considering that this project is only one part of a larger research agenda, and may be extended to include more countries and/or time periods in future, its implementation may be seen as part of an evolutionary, or incremental, development model. However, given that this larger research agenda is well-defined and of an academic nature, this would be considered to be more of a plan-driven, rather than an agile, approach.

3.2 Data to be used and the data pre-processing process

The data used in this project can be summarised into two sets. A dependent variable set consisting of: a crisis variable (binary) $\times$ country $\times$ time, and an explanatory variable set: macro-financial variable $\times$ country $\times$ time.

There are two main data sources that will be utilized in this project:

1. ECB/ESRB EU Crises Database [7]

   This dataset covers all EU Member States[2] and Norway for the period 1970-2016. It identifies 50 systemic crises using a consistent methodological approach. In addition, it will be regularly updated, allowing for the continued use of this database for project maintenance and/or extensions.

2. The World Bank DataBank

   A number of World Bank indicators will be collected for each country included in the ECB/ESRB EU crises database, covering the topics of the economy and growth, the financial sector, external debt, trade, the public and private sector, demographics, education, and poverty. These are presented in Table 1 in the appendix.

Given the sources of this data, its validity, veracity and general accuracy can be assured. In addition, the World Bank's database is made available under a Creative Commons Attribution 4.0 International License (CC BY 4.0), making the data used in this project freely and easily accessible.

The data pre-processing process for this project can be categorised into three objectives: (1) data cleaning, (2) data integration, and (3) data standardization/transformation.

The first step of the data pre-processing process will identify any missing values in the database, although, due to the high quality of the data sources, it is not foreseen that this would be a major concern. In the event of a missing data point in any given macroeconomic series, the value will be interpolated.

The next step of data transformation/standardization is expected to be more important given the nature of economic data and the requirements of neural networks, in general, and SOMs, in particular.

All data will be aggregated to a quarterly frequency for this project. All level variables will be logged, and quarterly and annual differences will be created for all variables, where applicable. While this should moderate the non-normality often observed in macro-financial data, the data will still need to be centered and scaled, given that neural networks (and therefore, SOMs) deliver superior results when run on standardized observations [17].

Finally, the data will need to be split into training, validation, and test data sets (70/15/15) for both the supervised and unsupervised SOMs that will be implemented in this analysis. This is to allow for weight estimation, hyper-parameter tuning, and to verify

[2]Austria, Italy, Belgium, Latvia, Bulgaria, Lithuania, Croatia, Luxembourg, Cyprus, Malta, Czechia, Netherland, Denmark, Poland, Estonia, Portugal, Finland, Romania, France, Slovakia, Germany, Slovenia, Greece, Spain, Hungary, Sweden, Ireland, and the United Kingdom.

the validity of these SOMs' use on future data.

3.3 Platforms and infrastructure to be used

While SOMs are generally computationally-heavy, due to the smaller scale of the project and the numeric data that will be used, this project will be completed on a laptop with a dual-core 2.20 GHz CPU, 12 GB of RAM, and 0.5 TB of hard drive space, running on a Windows operating system.

In addition, this project will be hosted on Github for easy sharing, collaboration and monitoring. While most of the project will be conducted on a local repository, it will require an internet connection to sync progress, and also allow the use of the "wbstats" R package to directly access the required data through the World Bank's API. While this data will be stored locally in standard csv-format, Google sheets may be used in future for further online developments.

Finally, the "kohonen" R package ([20]) will be used to fit the supervised and unsupervised SOMs, respectively.

3.4 Experimental process

As outlined earlier in this document, the purpose of this project is two-fold: (1) to understand which economic variables may indicate, through clustering, periods of financial distress, and (2) to understand how macroeconomic variables co-move over time and across borders.

Standard SOMs are a dimension reduction technique based on the use of a neural network model, with one input layer and one output layer (the map). The algorithm attempts to match each input onto its "best matching unit" in the output layer, and then "moves" that node, and its neighbours to a lesser extent, closer to the input. In this way, SOMs perform clustering similar to k-means clustering or principal components analysis, but maintaining the topographical features of the input dataset. This is done iteratively in order to minimise the distance between the map nodes and the inputs. While SOMs are similar to principal component analysis, they are less sensitive to the distribution of the underlying data, and so perform better in non-linear environments ([5]).

Two types of SOMs, each applied to two of the three dimensions of the explanatory variable data set, will be used in order to achieve the project's objectives:

- Supervised SOM

  This SOM will be run across countries (i.e. a map per country) using the time $\times$

variable dimensions of the data. The dependent variable will be modified slightly to indicate whether a country was in a period 6 months before the onset of a crisis, and this information will be used to fit the map. This follows the approach of [15], and should produce similar maps for each country. The feature planes from these maps can then be used to analyse how each economic variable is distributed across crisis, pre-crisis and non-crisis (tranquil) periods. And will also indicate whether these distributions appear correlated with each other.

- Unsupervised SOM
  This SOM will be run across time periods using the country $\times$ variable dimensions of the data. This unsupervised approach will cluster countries according to their similarity, which can then be analysed over time. In particular, as with the approach of [5], this can be used to visualise the co-movement of these countries' economic state before a crisis, during a crisis, and during more tranquil periods. It might also allow for the visualisation of crisis spillovers.

Each map will be utilize a hexagonal grid, and require the number of nodes (grid size), size of the neighbourhood (radius), the number of iterations (rlen), and learning rate (α) to be set. The validation set will be used to tune these latter hyperparameters. In general, however, a linearly declining trend for the learning rate and neighbourhood radius will be employed. Given that all of the explanatory variables are likely to be continuous in nature, the default "kohonen" package option of a "sum of squares" distance function will be used. In future, the euclidean distance function may be used as a robustness check and to investigate the performance of each option.

3.5 Evaluation process

The quality of a SOM is most often assessed through the use of the "distortion" measure ([14]), as there is no general cost function associated with the algorithm ([19]). This measure was shown to be a combination of both the SOMs' quantization error (related to the classical vector quantization error, how well the nodes fit the data) and its topographical error (how well the nodes preserve the topographical relationship present in the data) ([19]).

The evaluation of the supervised SOM will vary slightly from this as it also allows its performance to be quantified using a standard classification model framework which includes measures of accuracy, precision (the ratio of correctly identified crisis to all predicted crises) and recall (the ratio of correctly identified crisis to all actual crises).

Finally, SOM discrimination can be used to identify the economic variables that are most important for understanding periods of financial crises. This will be used to evaluate the findings from a more qualitative perspective, by determining whether these variables are in line with current economic theory, whether any new variables identified make intuitive sense, and whether these variables vary across countries.

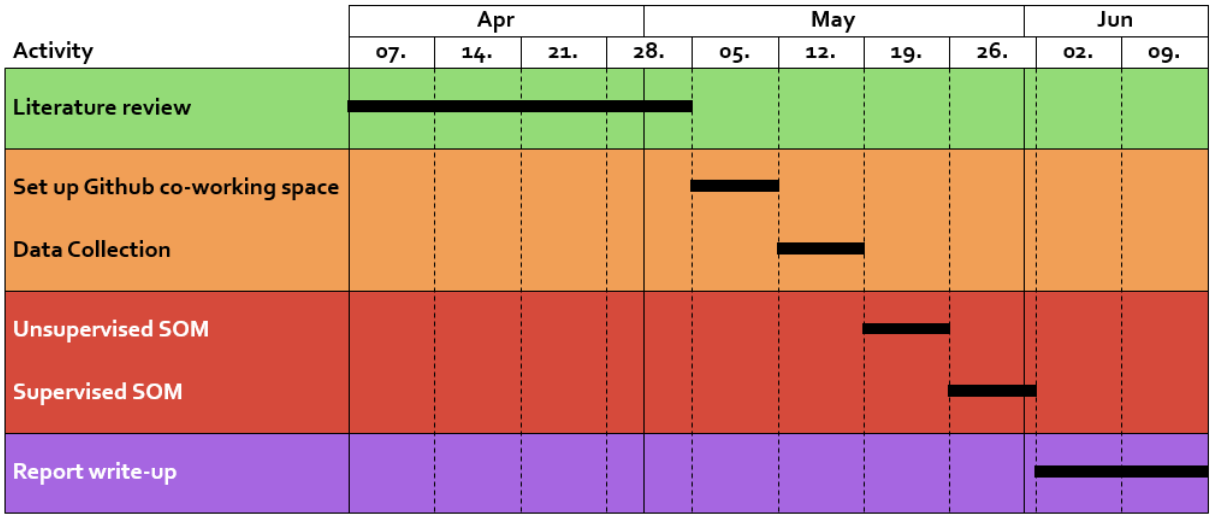
4 Project Report Structure

The final project report will consist of a reproducible (R Markdown), scientific document that contains the following sections:

- A brief introduction of the project and its problem statement;
- A standard statistical description of the data used in the project;
- A comprehensive write-up of the methodology employed, including parameter tuning;
- A section presenting the SOMs produced; and
- An analysis of the most important findings across countries and time periods. This should include a discussion of the identified signficant variables and their commonality or differences across countries and time periods.

5 Project Plan

The first month of the project will be spent on identifying and understanding existing economic literature on crises and EWS, as well as methodological literature on the use of SOMs to visualise relationships in high-dimensional data. Thereafter, a week will be spent on setting up the co-working space, collecting the required data, and estimating the SOM models for each country and each time period (1 week each). The remaining two weeks of the project will be spent on writing up the results (assignments 3 and 4) and finalising the report.


Apr
May
Jun
Activity
07.
14.
21.
28.
05.
12.
19.
26.
02.
09.
Literature review
Set up Github co-working space
Data Collection
Unsupervised SOM
Supervised SOM
Report write-up

References

[1] Mr David Amaglobeli, Mr Nicolas End, Mariusz Jarmuzek, and Mr Geremia Palomba. *From Systemic Banking Crises to Fiscal Costs: Risk Factors*. Number 15-166. International Monetary Fund, 2015.

[2] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[3] Susan Athey. The impact of machine learning on economics. In *Economics of Artificial Intelligence*. University of Chicago Press, 2017.

[4] Jan Babeckỳ, Tomas Havranek, Jakub Mateju, Marek Rusnák, Katerina Smidkova, and Borek Vasicek. Banking, debt and currency crises: early warning indicators for developed countries. 2012.

[5] 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.

[6] Dr Prittish Dala. Mit 808 – big data science project, 2019.

[7] Marco Lo Duca. A new database for financial crises in European countries. (194):62, 2017.

[8] Institute Electrical, Electronics Engineers, and Electronics Industry Association. Iso/iec/ieee 12207:2017 systems and software engineering — software life cycle processes. 11 2017.

[9] Omar Elgabry. Software engineering — software process and software process models (part 2). https://medium.com/omarelgabrys-blog/software-engineering-software-process-and-software-process-models-part-2-4a9d06213fdc, March 2017.

[10] Existek. Sdlc models explained: Agile, waterfall, v-shaped, iterative, spiral — existek blog. https://existek.com/blog/sdlc-models/, August 2017.

[11] Alejandro Gaytán, Christian A Johnson, et al. *A review of the literature on early warning systems for banking crises*, volume 183. Citeseer, 2002.

[12] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction Policy Problems. *American Economic Review*, 105(5):491–495, May 2015.

[13] Tutorials Point. Sdlc quick guide. https://www.tutorialspoint.com/sdlc/sdlc_quick_guide.htm.

[14] Joseph Rynkiewicz. Self-organizing map algorithm and distortion measure. *Neural networks*, 19(6-7):830–837, 2006.

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

[16] Walt Scacchi. Process models in software engineering. *Encyclopedia of software engineering*, 2002.

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

[18] IEEE Computer Society. Guide to the software engineering body of knowledge, version 3.0. www.swebok.org, 2014.

[19] Juha Vesanto, Mika Sulkava, and Jaakko Hollmén. On the decomposition of the self-organizing map distortion measure. In *Proceedings of the workshop on self-organizing maps (WSOM'03)*, pages 11–16, 2003.

[20] 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.

Table 1: World Bank indicators used

Bank capital to assets ratio (%)
Bank nonperforming loans to total gross loans (%)
Broad money (% of GDP)
Central government debt, total (% of GDP)
Current account balance (BoP, current US$)
Deposit interest rate (%)
Domestic credit to private sector (% of GDP)
Energy imports, net (% of energy use)
Expense (% of GDP)
Exports of goods and services (% of GDP)
External debt stocks (% of GNI)
External debt stocks, private nonguaranteed (DOD, current US$)
External debt stocks, public and publicly guaranteed (DOD, current US$)
External debt stocks, short-term (DOD, current US$)
External debt stocks, total (DOD, current US$)
Fuel exports (% of merchandise exports)
GDP (current US$)
GDP per capita (current US$)
GINI index (World Bank estimate)
Gross capital formation (% of GDP)
Gross savings (% of GDP)
High-technology exports (% of manufactured exports)
Imports of goods and services (% of GDP)
Industry (including construction), value added (% of GDP)
Inflation, consumer prices (annual %)
Interest rate spread (lending rate minus deposit rate, %)
Literacy rate, adult total (% of people ages 15 and above)
Merchandise exports (current US$)
Merchandise imports (current US$)
Merchandise trade (% of GDP)
Official exchange rate (LCU per US$, period average)
Population ages 15-64 (% of total)
Short-term debt (% of total reserves)
Tax revenue (% of GDP)
Total debt service (% of exports of goods, services and primary income)
Total reserves (includes gold, current US$)
Total tax and contribution rate (% of profit)
Unemployment, total (% of total labor force) (modeled ILO estimate)