



DOCTORAL PROGRAMME

Information Management

Specialization in Statistics and Econometrics

**Data Science for Finance: Targeted Learning
from (Big) Data to Economic Stability and
Financial Risk Management**

Afshin Ashofteh

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor in Information Management

September 2017 - May 2021

NOVA Information Management School

Universidade Nova de Lisboa

Afshin Ashofteh

**Data Science for Finance: Targeted Learning from (Big) Data to
Economic Stability and Financial Risk Management**

Dissertation submitted to Information Management School da
Universidade Nova de Lisboa

to obtain the degree of

**Doctor Philosophiae in Informatics Management Specialization
in Statistics and Econometrics**

Approved by:

President: Professor Doutor Tiago Oliveira.

Referee: Professor Doutor Manuel Vilares.

Referee: Professor Doutor Patrício Ricardo Soares Costa.

Referee: Professor Doutor José Carlos Dias.

Supervisor: Professor Doutor Jorge Miguel Ventura Bravo.

(Jorge Miguel Ventura Bravo)

**Data Science for Finance: Targeted Learning from (Big) Data
to Economic Stability and Financial Risk Management**

Work Supervised by:

Professor Doutor Jorge Miguel Ventura Bravo, Supervisor.

Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa
NOVA Information Management School

Copyright © by

[Afshin Ashofteh](#)

All rights reserved

TO MY BELOVED FAMILY...

ACKNOWLEDGMENTS

With was the help of so many that made this dissertation possible. First of all, my deepest gratitude goes to my supervisor, Professor Jorge Bravo, for his continuous guidance, motivation, and constructive feedback during the Ph.D. Program. With his great support, I was able to deliver the results presented in this dissertation. He was not only my supervisor but also a good, supportive, and kind friend. Professor Jorge Bravo also helped me realize how valuable is to produce work that can be appreciated by Academia but also bringing contributions to Society.

I am also most grateful to Professor Dr. Fernando Bação, Professor Dr. Marco Painho, and Professor Dr. Paulo Rita for their guidance and support during the Research Seminars and my colleagues as peer reviewers of my research in these courses that provided the right guidance to execute a Ph.D. research.

I would also like, to mention my cordial gratitude to all my teachers, colleagues, Nova athletics sports team, Nova Doctoral School, and Nova IMS staff that together make this institution a wonderful place to conduct Ph.D. research. Additionally, I would make a special mention of gratitude to José L. Cervera-Ferri (devstat), Steve MacFeely (UNCTAD & WHO) for their invitation as an invited speaker during my Ph.D. career, which motivate and support me mentally in my research career. Also, the Risk Analytics Committee of ISI for selecting my manuscript for the first prize of the International Conference on Risk Analysis and Design of Experiments.

I would like to thank, all people that I might not even know about their kindnesses, but I could feel them by my heart.

To all my family and friends for their support, care, incentive, understanding, and love. Particularly to my wife, and my son.

Finally, I also want to express my gratitude to the scientific community. The authors of the many scientific papers I have read, especially those cited in this dissertation, and the anonymous reviewers for their constructive feedback.

*Afshin Ashofteh
aashofteh@novaims.unl.pt*

ABSTRACT

The modelling, measurement, and management of systemic financial stability remains a critical issue in most countries. Policymakers, regulators, and managers depend on complex models for financial stability and risk management. The models are compelled to be robust, realistic, and consistent with all relevant available data. This requires great data disclosure, which is deemed to have the highest quality standards. However, stressed situations, financial crises, and pandemics are the source of many new risks with new requirements such as new data sources and different models.

This dissertation aims to show the data quality challenges of high-risk situations such as pandemics or economic crisis and it try to theorize the new machine learning models for predictive and longitudes time series models.

In the first study (Chapter Two) we analyzed and compared the quality of official datasets available for COVID-19 as a best practice for a recent high-risk situation with dramatic effects on financial stability. We used comparative statistical analysis to evaluate the accuracy of data collection by a national (Chinese Center for Disease Control and Prevention) and two international (World Health Organization; European Centre for Disease Prevention and Control) organizations based on the value of systematic measurement errors. We combined excel files, text mining techniques, and manual data entries to extract the COVID-19 data from official reports and to generate an accurate profile for comparisons. The findings show noticeable and increasing measurement errors in the three datasets as the pandemic outbreak expanded and more countries contributed data for the official repositories, raising data comparability concerns and pointing to the need for better coordination and harmonized statistical methods. The study offers a COVID-19 combined dataset and dashboard with minimum systematic measurement errors and valuable insights into the potential problems in using databanks without carefully examining the metadata and additional documentation that describe the overall context of data.

In the second study (Chapter Three) we discussed credit risk as the most significant source of risk in banking as one of the most important sectors of financial institutions. We proposed a new machine learning approach for online credit scoring which is enough conservative and robust for unstable and high-risk situations. This Chapter is aimed at the case of credit scoring in risk management and presents a novel method to be used for the default prediction of high-risk branches or customers. This study uses the Kruskal-Wallis non-

parametric statistic to form a conservative credit-scoring model and to study its impact on modeling performance on the benefit of the credit provider. The findings show that the new credit scoring methodology represents a reasonable coefficient of determination and a very low false-negative rate. It is computationally less expensive with high accuracy with around 18% improvement in Recall/Sensitivity. Because of the recent perspective of continued credit/behavior scoring, our study suggests using this credit score for non-traditional data sources for online loan providers to allow them to study and reveal changes in client behavior over time and choose the reliable unbanked customers, based on their application data. This is the first study that develops an online non-parametric credit scoring system, which can reselect effective features automatically for continued credit evaluation and weigh them out by their level of contribution with a good diagnostic ability.

In the third study (Chapter Four) we focus on the financial stability challenges faced by insurance companies and pension schemes when managing systematic (undiversifiable) mortality and longevity risk. For this purpose, we first developed a new ensemble learning strategy for panel time-series forecasting and studied its applications to tracking respiratory disease excess mortality during the COVID-19 pandemic. The layered learning approach is a solution related to ensemble learning to address a given predictive task by different predictive models when direct mapping from inputs to outputs is not accurate. We adopt a layered learning approach to an ensemble learning strategy to solve the predictive tasks with improved predictive performance and take advantage of multiple learning processes into an ensemble model. In this proposed strategy, the appropriate holdout for each model is specified individually. Additionally, the models in the ensemble are selected by a proposed selection approach to be combined dynamically based on their predictive performance. It provides a high-performance ensemble model to automatically cope with the different kinds of time series for each panel member. For the experimental section, we studied more than twelve thousand observations in a portfolio of 61-time series (countries) of reported respiratory disease deaths with monthly sampling frequency to show the amount of improvement in predictive performance. We then compare each country's forecasts of respiratory disease deaths generated by our model with the corresponding COVID-19 deaths in 2020. The results of this large set of experiments show that the accuracy of the ensemble model is improved noticeably by using different holdouts for different contributed time series methods based on the proposed model selection method. These improved time series models provide us proper forecasting of respiratory disease deaths for each country, exhibiting high correlation (0.94) with Covid-19 deaths in 2020.

In the fourth study (Chapter Five) we used the new ensemble learning approach for time series modeling, discussed in the previous Chapter, accompany by K-means clustering for forecasting life tables in COVID-19 times. Stochastic mortality modeling plays a critical role in public pension design, population and public health projections, and in the design, pricing, and risk management of life insurance contracts and longevity-linked securities. There is no general method to forecast the mortality rate applicable to all situations especially for unusual years such as the COVID-19 pandemic. In this Chapter, we investigate the feasibility of using an ensemble of traditional and machine learning time series methods to empower forecasts of age-specific mortality rates for groups of countries that share common longevity trends. We use Generalized Age-Period-Cohort stochastic mortality models to capture age and period effects, apply K-means clustering to time series to group countries following common longevity trends, and use ensemble learning to forecast life expectancy and annuity prices by age and sex. To calibrate models, we use data for 14 European countries from 1960 to 2018. The results show that the ensemble method presents the best robust results overall with minimum RMSE in the presence of structural changes in the shape of time series at the time of COVID-19.

In this dissertation's conclusions (Chapter Six), we provide more detailed insights about the overall contributions of this dissertation on the financial stability and risk management by data science, opportunities, limitations, and avenues for future research about the application of data science in finance and economy.■

Keywords:

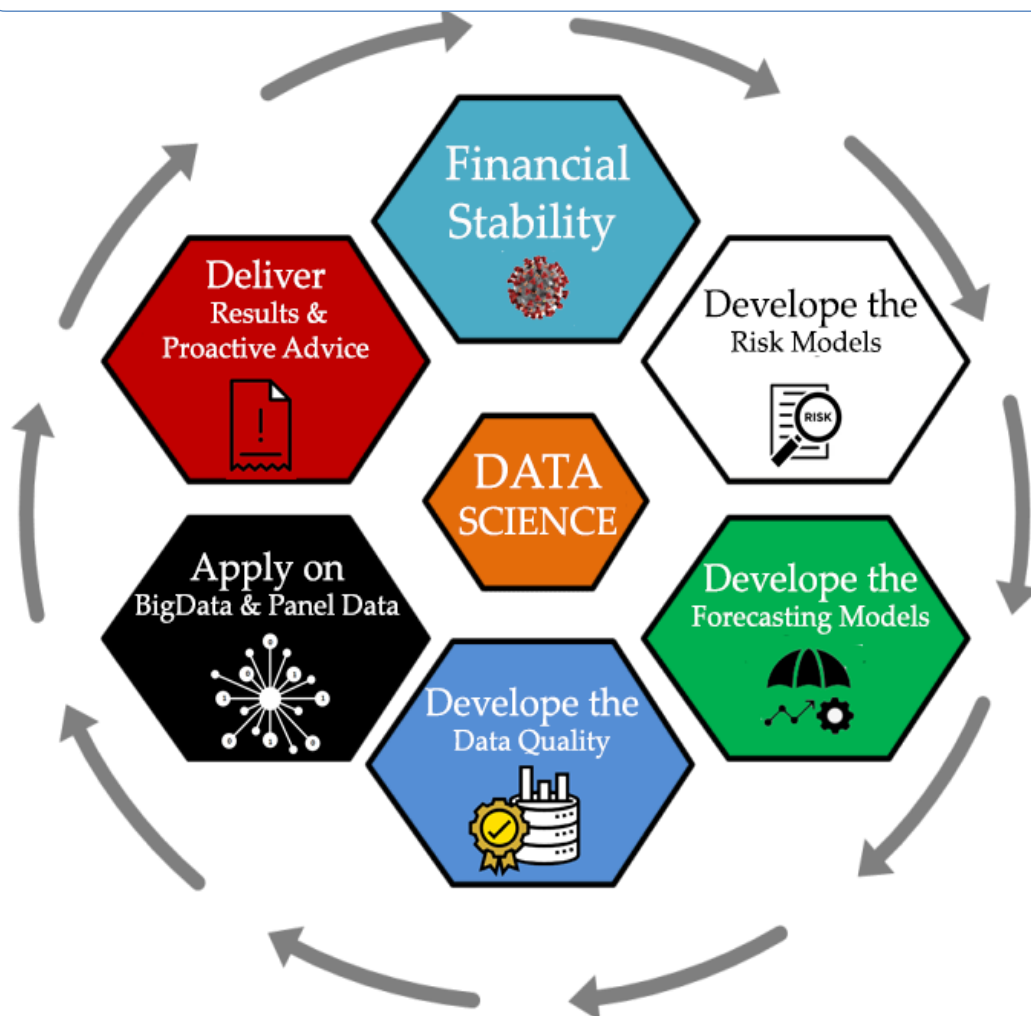
Data Science; Machine learning; Big Data; PySpark; Layered learning; Ensemble learning; Multiple learning process; Panel Time series; Ensemble Bayesian Model Averaging (EBMA); Measurement error; Data quality; Risk Analysis; Online credit scoring; Credit Risk; Finance; Economic Stability; Financial Risk Management; Excess Mortality; Mortality forecasting; Annuities; Life expectancy; Banking; Insurance; Open banking; SARS-CoV-2; Coronavirus (COVID-19).

GRAPHICAL ABSTRACT

Introduction | Research Context | Motivation | Methodological Approach



Research Question - How to Improve Machine Learning Techniques and Time Series Models to Harness the Power of Data Science in Finance and in the Economy for Better Financial Stability and Risk Management?



Conclusion | Future Research

PUBLICATIONS, DATASETS, CODES & AWARDS

Journal Articles (Scopus and ISI indexed)

Ashofteh, A., & Bravo, J. M. (2021). A conservative approach for online credit scoring. *Expert Systems with Applications*, 176, 114835. <https://doi.org/10.1016/j.eswa.2021.114835>

(Retrieved from:

<https://www.sciencedirect.com/science/article/pii/S0957417421002761?via%3Dihub>)

Ashofteh, A., & Bravo, J. M. (2020). A study on the quality of novel coronavirus (COVID-19) official datasets. *Statistical Journal of the IAOS*, 36(2), 291–301. <https://doi.org/10.3233/SJI-200674>

(Retrieved from:

<https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji200674>)

Ashofteh, A., Bravo, J. M., Ayuso, M., (submitted on a top scholarly journal), A New Ensemble Learning Strategy for Panel Time-Series Forecasting with Applications to Tracking Respiratory Disease Excess Mortality during the COVID-19 pandemic.

Scopus indexed conference papers

Ashofteh, A., Bravo, J. M., Ayuso, M. (2021). A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic. In *Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao 2021- 21th Conference of the Portuguese Association for Information Systems, CAPSI 2021*.

Ashofteh, A., & Bravo, J. M. (2021). Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach. 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), 2021, pp. 1-6, DOI: <https://doi.org/10.23919/CISTI52073.2021.9476583>.

(Available at: <https://ieeexplore.ieee.org/document/9476583>)

Ashofteh, A., & Bravo, J. M. (2019). A non-parametric-based computationally efficient approach for credit scoring. In *Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao 2019- 19th Conference of the Portuguese Association for Information Systems, CAPSI 2019; Lisboa; Portugal; 11 October 2019 through 12 October 2019* (pp. 19).

https://www.novaims.unl.pt/uploads/imagens_ficheiros/CAPSI-ata-da-conferencia/ebook-capsi-2019.pdf

Ashofteh, A. (2018). Mining Big Data in statistical systems of the monetary financial institutions (MFIs). International Conference on Advanced Research Methods and Analytics (CARMA 2018). ISBN: 9788490486894. <https://doi.org/10.4995/carma2018.2018.8570>

Dataset

Ashofteh, Afshin; M. Bravo, Jorge (2020). "Corona-virus disease (COVID-19) Data-set with Improved Measurement Errors of Referenced Official Data Sources ", Mendeley Data, V3.
<http://dx.doi.org/10.17632/nw5m4hs3jr.3>

Code & Software

Ashofteh, Afshin; M. Bravo, Jorge (2021). Spark Code: A Novel Conservative Approach for Online Credit Scoring [Source Code]. CodeOcean.
<https://doi.org/10.24433/CO.1963899.V1>

Press/Media & Public Impact

Ashofteh, Afshin; M. Bravo, Jorge (2020). Óbitos Negativos ou Datas Falsas. Detetados Erros nas Maiores Bases de Dados da Pandemia. Television (TVI24). <https://tvi24.iol.pt/internacional/estudo/detetados-erros-em-base-de-dados-da-oms-sobre-pandemia-obitos-negativos-e-um-deles>

Ashofteh, Afshin; M. Bravo, Jorge (2020). Covid-19: Estudo revela imprecisões em bases de dados sobre novo coronavírus. News Agency (LUSA).
<https://www.lusa.pt/article/k3WvWHCnrM0LrEQFdBkJ7jMSZM5iuSI1/covid-19-estudo-revela-imprecis%C3%B5es-em-bases-de-dados-sobre-novo-coronav%C3%ADrus-top-five-news>

Prize

First Prize for the presentation of "A non-parametric-based computationally efficient approach for credit scoring using non-traditional data", at the 8th International Conference on Risk Analysis and Design of Experiments. (Best Paper Award)



Winner's Certificate

It is hereby confirmed that **Afshin Ashofteh** has won the first prize for his presentation entitled

"A Non-Parametric-Based Computationally Efficient Approach for Credit Scoring Using Non-Traditional Data"

at the

**Eighth International Conference on Risk Analysis
and Design of Experiments**

University of Natural Resources and Life Sciences

Vienna, Austria

April 23rd - 26th, 2019

UNIVERSITÄT FÜR BODENKULTUR WIEN
Dept. Raum, Landschaft & Infrastruktur
Institut für Angewandte Statistik und EDV
Peter-Jordan-Straße 82, 1190 Wien
Tel. +43/1/47654-5060, Fax +43/1/47654-5069

A.o. Prof. Dr. Karl Moder
Head of the LOC

Afshin Ashofteh ([Email](#))

- | | |
|--------------------|---------------------------------|
| ▶ Research Profile | ▶ ORCID: 0000-0001-5075-9822 |
| ▶ Academic Profile | ▶ Ciência ID: 5213-E4D6-CC60 |
| ▶ ResearchGate | ▶ Scopus Author ID: 57217177194 |
| ▶ LinkedIn | ▶ ResearcherID: AAI-3996-2021 |

| | |
|---|-------------|
| ACKNOWLEDGMENTS | VII |
| ABSTRACT | IX |
| GRAPHICAL ABSTRACT | XIII |
| PUBLICATIONS, DATASETS, CODES & AWARDS..... | XV |
| | |
| 1 CHAPTER ONE - INTRODUCTION | 1 |
| 1.1 RESEARCH CONTEXT | 1 |
| 1.2 MOTIVATION | 2 |
| 1.3 RESEARCH FOCUS..... | 3 |
| 1.4 RESEARCH GOALS..... | 3 |
| 1.5 METHODS AND TECHNOLOGIES..... | 4 |
| 1.6 RESEARCH PATH..... | 5 |
| 2 CHAPTER TWO - A STUDY ON THE QUALITY OF OFFICIAL DATASETS IN DISTRESS SITUATIONS SUCH AS NOVEL CORONAVIRUS (COVID-19) TIMES..... | 7 |
| 2.1 INTRODUCTION | 7 |
| 2.2 OFFICIAL COVID-19 DATASETS: AN OVERVIEW..... | 11 |
| 2.2.1 World Health Organization (WHO) reports..... | 11 |
| 2.2.2 European Centre for Disease Prevention and Control dataset (ECDC) | 12 |
| 2.2.3 Chinese Center for Disease Control and Prevention (Chinese CDC) | 13 |
| 2.3 METHODS | 15 |
| 2.3.1 Data..... | 15 |
| 2.3.2 Errors and Outliers | 16 |
| 2.4 RESULTS..... | 21 |
| 2.5 CONCLUSION..... | 25 |
| 3 CHAPTER THREE - A CONSERVATIVE MACHINE LEARNING APPROACH FOR ONLINE CREDIT SCORING UNDER DISTRESS SITUATIONS SUCH AS COVID-19 TIMES 27 | |
| 3.1 INTRODUCTION | 28 |
| 3.2 LITERATURE REVIEW..... | 31 |
| 3.2.1 Credit scoring models using traditional data sets..... | 31 |
| 3.2.2 Non-traditional data sets for credit scoring | 34 |
| 3.3 DYNAMIC NONPARAMETRIC CREDIT SCORE..... | 38 |
| 3.3.1 Kruskal-Wallis statistic for online features reduction | 38 |
| 3.3.2 Kruskal-Wallis statistic for empowering features | 42 |
| 3.3.3 Credit score formulation..... | 44 |

| | | |
|----------|---|------------|
| 3.4 | EXPERIMENTAL DESIGN | 46 |
| 3.4.1 | Small data set..... | 46 |
| 3.4.2 | Big Data set..... | 52 |
| 3.5 | IMPACT OF RESEARCH | 61 |
| 3.5.1 | Confidentiality and privacy | 61 |
| 3.5.2 | Financial inclusion | 61 |
| 3.5.3 | Compliance risk impact..... | 62 |
| 3.6 | CONCLUSION..... | 62 |
| 4 | CHAPTER FOUR - A NEW ENSEMBLE LEARNING STRATEGY FOR PANEL TIME-SERIES FORECASTING WITH APPLICATIONS TO TRACKING RESPIRATORY DISEASE EXCESS MORTALITY DURING THE COVID-19 PANDEMIC..... | 65 |
| 4.1 | INTRODUCTION | 66 |
| 4.2 | MATERIALS AND METHODS | 71 |
| 4.2.1 | Layered learning and the proposed ensemble learning strategy | 73 |
| 4.2.2 | The learning algorithms..... | 78 |
| 4.3 | EMPIRICAL EXPERIMENTS..... | 82 |
| 4.3.1 | Data selection and cleansing..... | 82 |
| 4.3.2 | Results | 84 |
| 4.4 | DISCUSSION AND CONCLUSION | 96 |
| 5 | CHAPTER FIVE - LIFE TABLE FORECASTING IN COVID-19 TIMES: AN ENSEMBLE LEARNING APPROACH..... | 99 |
| 5.1 | INTRODUCTION | 100 |
| 5.2 | MATERIALS AND METHODS | 102 |
| 5.2.1 | GAPC stochastic mortality models..... | 102 |
| 5.2.2 | K-Means Clustering of time trend indices | 103 |
| 5.2.3 | Ensemble learning of time series | 104 |
| 5.2.4 | Life expectancy and life annuity computation..... | 106 |
| 5.3 | RESULTS..... | 106 |
| 5.4 | CONCLUSIONS | 111 |
| 6 | CHAPTER SIX - CONCLUSIONS..... | 113 |
| 6.1 | SUMMARY OF FINDINGS AND CONTRIBUTIONS | 113 |
| 6.2 | LIMITATIONS..... | 114 |
| 6.3 | FUTURE RESEARCH | 114 |
| | REFERENCES..... | 117 |

| | |
|---|------------|
| APPENDIX..... | 127 |
| APPENDIX-1 PYSPARK CODE | 127 |
| APPENDIX-2 CORRECTIONS IN THE DATASET | 138 |
| APPENDIX-3 R CODE..... | 139 |

List of Tables

| | |
|---|-----|
| TABLE 2.1 – SAMPLE OF DATA FROM SITUATION REPORTS OF THE WORLD HEALTH ORGANISATION (WHO)..... | 12 |
| TABLE 2.2 – TOP 10 ROWS OF AGGREGATED ATTRIBUTES OF COVID-19 FOR CHINA - WESTERN PACIFIC REGION. | 13 |
| TABLE 2.3 - TOP 10 ROWS OF NEW ATTRIBUTES OF COVID-19 FOR CHINA - WESTERN PACIFIC REGION. | 14 |
| TABLE 2.4 - ANALYTICAL BASE TABLE (ABT) OF JOINED DATA SOURCES..... | 15 |
| TABLE 2.5 - NEGATIVE VALUES IN DATASETS..... | 17 |
| TABLE 2.6 - ROOT MEAN SQUARE ERRORS OF ATTRIBUTES OF DIFFERENT REPORTS..... | 19 |
| TABLE 2.7 - COMPARING DISTRIBUTIONS BASED ON RMSE. | 23 |
| TABLE 2.8 - IMPORTED CASES TO CHINA FROM OUTSIDE..... | 24 |
| TABLE 2.9 - COMPARING DISTRIBUTIONS BASED ON RMSE WITH CORRECTION FOR IMPORTED POSITIVE CASES. | 25 |
| TABLE 3.1 - . H_{JT} VALUES AND ITS EQUIVALENT IN G_{JT} , W_{JT} AND Φ_{JT} | 43 |
| TABLE 3.2 - PSEUDO CODE OF PROPOSED METHODOLOGY. | 45 |
| TABLE 3.3 - - DESCRIPTIVE STATISTICS OF “TYPE OF HOME-OWNERSHIP.” | 47 |
| TABLE 3.4 - DESCRIPTIVE STATISTICS OF “EXISTENCE OF RECORDS” & “TYPE OF JOB.” | 47 |
| TABLE 3.5 - DESCRIPTIVE STATISTICS OF “MARITAL STATUS.” | 47 |
| TABLE 3.6 - DESCRIPTIVE STATISTICS OF SCALE VARIABLES..... | 47 |
| TABLE 3.7 - K-W, $WK1$ AND $\Phi1$ | 49 |
| TABLE 3.8 - CREDIT SCORE VALIDATION..... | 49 |
| TABLE 3.9 - CLASSIFICATION TABLE AND STATISTICAL MODEL PERFORMANCE (AUC) FOR ORIGINAL VARIABLES. | 50 |
| TABLE 3.10 - CLASSIFICATION TABLE AND STATISTICAL MODEL PERFORMANCE (AUC) FOR WEIGHTED VARIABLES AND CRI..... | 51 |
| TABLE 3.11 - K-W, $WK1$ AND $\Phi1$ | 55 |
| TABLE 3.12 - CLASSIFICATION TABLE FOR LOGISTIC REGRESSION WITH RIDGE PENALTY. IN THE TABLE 0, 1 INDICATE NON-DEFAULT LOAN AND DEFAULT LOAN/PAYMENT ARRIARS, RESPECTIVELY, AND LOSS STANDS FOR THE SUBTRACTION OF NET REMAIN OF (0,0) FROM (1,0) COMBINATIONS..... | 57 |
| TABLE 3.13 - CLASSIFICATION TABLE FOR RANDOM FOREST CLASSIFIER. IN THE TABLE 0, 1 INDICATE NON- DEFAULT LOAN AND DEFAULT LOAN/PAYMENT ARRIARS, RESPECTIVELY, AND LOSS STANDS FOR THE SUBTRACTION OF NET REMAIN OF (0,0) FROM (1,0) COMBINATIONS..... | 58 |
| TABLE 3.14 - CLASSIFICATION TABLE FOR LINEAR SUPPORT VECTOR MACHINE IN THE TABLE 0, 1 INDICATE NON- DEFAULT LOAN AND DEFAULT LOAN/PAYMENT ARRIARS, RESPECTIVELY, AND LOSS STANDS FOR THE SUBTRACTION OF NET REMAIN OF (0,0) FROM (1,0) COMBINATIONS..... | 59 |
| TABLE 3.15 - THE BEST MODEL OF EACH CLASSIFIER. | 60 |
| TABLE 4.1. PSEUDO CODE OF THE PROPOSED ENSEMBLE STRATEGY. | 76 |
| TABLE 4.2 - ALGORITHMS AND HYPER-PARAMETERS CHOICES..... | 80 |
| TABLE 4.3. DIFFERENT LEVELS OF QUALITY ALLOCATED FOR THE REPORTED RESPIRATORY DISEASE DEATHS BY COUNTRIES. | 82 |
| TABLE 4.4. METADATA OF CODE OF DISEASES CATEGORIZED AS RESPIRATORY DISEASE. | 83 |
| TABLE 4.5. RANKING THE MODELS AND ENSEMBLES ACCORDING TO THE ACCURACY MEASURE. | 86 |
| TABLE 4.6. CONTRIBUTION RATE OF THE MODELS IN THE ENSEMBLE. | 88 |
| TABLE 4.7. THE MODEL’S EXCLUSION FREQUENCY FOR THE ENSEMBLE WITH DYNAMIC HOLDOUTS..... | 88 |
| TABLE 4.8. THE METHODOLOGY EFFECT ON THE RUN-TIME AND COMPUTATIONAL EFFICIENCY..... | 90 |
| TABLE 4.9 - COMPARISON BETWEEN FORECASTING DEATHS FOR RESPIRATORY DISEASES AND ACTUAL COVID-19 DEATHS. | 91 |
| TABLE 5.1 - SUMMARY OF THE LEARNING ALGORITHMS..... | 105 |
| TABLE 5.2 - LIFE ANNUITY, LIFE EXPECTANCY AND PROBABILITY OF DEATH BASED ON ARIMA FORECASTING OF KAPPA FOR THE YEAR 2020..... | 110 |

| | |
|---|-----|
| TABLE 5.3 - LIFE ANNUITY, LIFE EXPECTANCY AND PROBABILITY OF DEATH BASED ON ENSEMBLE FORECASTING OF KAPPA FOR THE YEAR 2020..... | 111 |
|---|-----|

List of Figures

| | |
|---|-----|
| FIGURE 2-1 - SCATTER PLOT AND CORRELATION BETWEEN CCDC REPORTS WITH WHO, AND ECDC REPORTS. | 17 |
| FIGURE 2-2 - CORRELATION OF NCC BETWEEN WHO AND ECDC REPORTED DATA. | 18 |
| FIGURE 2-3 - DAILY SUM OF SQUARE ERROR AGGREGATED FOR ALL ATTRIBUTES, COUNTRIES AND DATASETS. ... | 20 |
| FIGURE 2-4 - POSITIVE TREND IN THE ROOT MEAN SQUARE ERROR AGGREGATED FOR ALL ATTRIBUTES RELATED TO NEW CASES FOR ALL COUNTRIES IN THE THREE REFERENCE DATASETS. | 21 |
| FIGURE 2-5 - POSITIVE TREND IN THE ROOT MEAN SQUARE ERROR AGGREGATED FOR TOTAL CONFIRMED CASES OF ALL COUNTRIES IN THE THREE REFERENCE DATASETS. | 21 |
| FIGURE 2-6 - THE LOGNORMAL DISTRIBUTION FOR ATTRIBUTED NEW CONFIRMED CASES IN CHINA..... | 23 |
| FIGURE 3-1- CHALLENGES IN CREDIT SCORING..... | 29 |
| FIGURE 3-2 - EMERGENCE OF NON-TRADITIONAL DATA ANALYSIS IN CREDIT SCORING. | 35 |
| FIGURE 3-3 - THE DENSITY PLOT OF IJT. | 41 |
| FIGURE 3-4 - NORMALIZED IMPORTANCE OF ATTRIBUTES IN MODELING STAGE. | 56 |
| FIGURE 4-1 - GRAPHICAL ABSTRACT OF THE PROPOSED DYNAMIC ENSEMBLE LEARNING STRATEGY. | 72 |
| FIGURE 4-2 - PROPOSED STRATEGY OF ENSEMBLE LEARNING. | 74 |
| FIGURE 4-3 - COMPARING THE ACCURACY OF THE MODELS..... | 87 |
| FIGURE 4-4 - BME MODEL CONFIDENCE SET AND ESTIMATED WEIGHTS PER COUNTRY. | 89 |
| FIGURE 4-5 - RESPIRATORY DISEASES DEATHS AND COVID-19 DEATHS FOR EUROPE AND NORTH AMERICA IN 2020..... | 93 |
| FIGURE 4-6 - RESPIRATORY DISEASES DEATHS AND COVID-19 DEATHS FOR EACH COUNTRY IN 2020..... | 95 |
| FIGURE 5-1 - CHANGES IN THE TOTAL LOG MORTALITY RATES WITH RESPECT TO BOTH AGE AND YEAR OVER THE PERIOD 1960-2018 IN PORTUGAL..... | 107 |
| FIGURE 5-2 - THE PARAMETERS OF THE POISSON LEE-CARTER MODEL OVER THE PERIOD 1960 - 2018. PORTUGAL WITH MAXIMUM BETA AT AGE 72 AND MAXIMUM KAPPA AT THE YEAR 1969..... | 107 |
| FIGURE 5-3 - LIFE EXPECTANCY/ANNUITY PRICES; BEFORE (BLUE) AND AFTER (RED) THE COVID19. | 108 |
| FIGURE 5-4 - CLUSTER DENDROGRAMS OF THE BASELINE TIME SERIES PARAMETERS BY COUNTRY AND GENDER..... | 109 |

Abbreviations & Acronyms

| | |
|---------------|--|
| AIC | Akaike information criterion |
| ARIMA | Auto-Regressive Integrated Moving Average |
| AUC | Area under the receiver operating curve |
| BCBS | Basel Committee on Banking Supervision |
| BIC | Bayesian information criterion |
| BME | Bayesian Model Ensemble |
| CDF | Cumulative distribution function |
| CDR | Call detail records (Cellphone data) |
| CECL | Current Expected Credit Loss |
| Chinese CDC | Chinese Center for Disease Control and Prevention |
| COVID-19 | Coronavirus 2019 |
| CRI | Credit risk index |
| DA | Discriminate analysis |
| DBN | Deep belief networks |
| DGCEC | Deep Genetic Cascade Ensembles of Classifiers |
| DTW | Dynamic time warping |
| ECDC | European Centre for Disease Prevention and Control |
| ELM | Extreme Learning Machine methods |
| ETS | Exponential Smoothing State Space Model |
| FASB | Financial Accounting Standards Board |
| FQSSVM | Fuzzy Quadratic Surface Support Vector Machine |
| FR | Fatality rate |
| GA | Genetic and evolutionary algorithms |
| GAPC | Generalized Age-Period-Cohort |
| GBDT | Gradient boosting decision tree classifier |
| GDDS | General Data Dissemination System |
| GDPR | General Data Protection Regulation |
| GNM | Generalized nonlinear models |
| HACT | Hybrid associative classifier with translation |
| HWA | Holt-Winters' additive method |
| HWM | Holt-Winters' multiplicative method |
| IFRS9 | International Financial Reporting Standard |
| IRB | Internal Ratings Based approach |
| K-W statistic | Kruskal-Wallis statistic |
| LBs | Distance and its corresponding lower bounds |
| LR | Logistic regression |
| ML | Maximum-likelihood methods |

| | |
|------------|---|
| MLP | Multilayer Perceptron for time series |
| MR | Mortality rate |
| NCC | New confirmed cases |
| NCC | New confirmed cases |
| ND | New deaths |
| NN | Neural network |
| NNETAR | Neural network autoregression model |
| PHIS | Public health information systems |
| Pop | Population |
| RD_TD | Total deaths for respiratory diseases |
| RF | Random Forests |
| ROC | Receiver operating curves |
| RWF | Random Walk with drift model |
| SARS-CoV-2 | Severe acute respiratory syndrome coronavirus 2 |
| SDDS | Special Data Dissemination Standard |
| SDDS+ | Special Data Dissemination Standard plus |
| SMAPE | Symmetric mean absolute percentage error |
| SMOTE | Synthetic minority oversampling technique |
| SNAIVE | Seasonal Naïve random walk forecast model |
| SSA | Singular spectrum analysis |
| SSVM | Semi-supervised Support Vector Machines |
| STL | Seasonal Trend Decomposition uses Loess |
| SVM | Support Vector Machine |
| TCC | Total confirmed cases |
| TD | Total deaths |
| WHO | World Health Organization ■ |

CHAPTER ONE – INTRODUCTION

This Chapter contextualizes the research motivation and main research goals. Here is also described the methodological approach and dissertation structure.

1.1 RESEARCH CONTEXT

This dissertation is within the context of information management, with a special emphasis on data science, statistics, and econometrics. This research seeks effective technical solutions of targeted learning, Big Data, and ensemble learning to be implemented into the financial data, especially at the time of distress situations such as financial crises or pandemics.

Targeted learning focuses on efficient machine-learning-based substitution estimators of parameters that are defined as features of the probability distribution of the data. This research area is very challenging in the area of classification algorithms of imbalanced big data.

Big Data posits the challenges of Volume, Velocity, and Variety. In addition, other attributes have been linked to Big Data such as Veracity or Value, among others. The scalability issue of financial online data must be properly addressed to develop new solutions or adapt existing ones for Big Data case studies. Spark has emerged as a popular choice to implement large-scale Machine Learning applications on Big Data and sound statistical and machine learning procedures are required to be scalable computationally and to deal with massive datasets of financial corporations.

Ensemble learning methods for longitudinal data analysis play a critical role in many areas of econometrics and risk management. Developing panel time series models and making them robust enough against distress situations is a recent study area, in which few pieces of research have been conducted.

1.2MOTIVATION

Policymakers depend on complex models that are compelled to be robust, realistic, defensible, and explainable to make decisions on basic and complex problems and shocks that affect the economy and society.

The international financial crisis and the COVID-19 pandemic outbreak prompted several statutory and supervisory initiatives that require great data disclosure by financial firms. Recently, open banking, new banking, and data-driven insurance organizations are producing big data sources, which could be used to monitor different kinds of risks.

Data Science and new Non-traditional datasets provide the banking and finance industry a chance to manage the volume, variety, and velocity from different sources to boost business outcomes. This new approach of business data analysis plays a great competitive advantage for the main functions of monetary and financial institutions and the whole economy (Ashofteh & M. Bravo, 2021b).

For using this potential, it is of extreme importance to design novel approaches to deal with the challenges that come with big, complex, and dynamic data. It is essential to present a scientific roadmap to translate machine-learning applications into challenging practical applications such as credit scoring in the risk management systems and financial stability.

These new approaches should be able to deal with distress situations, which are relatively infrequent events and there is usually very limited information for distinguishing them in an extremely sparse and imbalanced financial data environment. It makes financial stability more and more challenging.

Being still a recent discipline, few pieces of research has been conducted on using Big Data and ensemble machine learning approaches for financial and economical problems, which at the same time could be explainable and fit enough with regulations. The reasons behind this are mainly the difficulties in adapting standard techniques, regulations, and complicated financial concepts to the new machine learning approaches and MapReduce programming style. Additionally, inner problems of imbalanced data, namely the lack of data and

small disjuncts, are accentuated during the data partitioning in the MapReduce programming style when we are dealing with Big Data.

1.3 RESEARCH FOCUS

This dissertation focuses on modelling and risk management in banking, insurance, and pension funds, which are key players in every country's financial system. For the methodology, the focus is on machine learning, Big Data, and ensemble time series models. To comprehend the role of data science in financial stability, it is critical to:

- Develop a new machine learning approach for better risk management of online banking by using Big Data.
- Develop a new ensemble time series modeling approach for better risk management of insurance companies and pension funds by using panel data.
- Evaluate the potential problems of big data sources at the time of financial crisis or pandemics.

We expect that this dissertation will improve the knowledge of data science in finance, by suggesting new models that will include the most relevant topics in risk management and forecasting in finance.

1.4 RESEARCH GOALS

The main goal of this dissertation is to propose new methodologies of risk management by developing machine learning algorithms for financial Big Data, longitudinal data, and panel data, appropriate for distress situations.

With this purpose in mind, the dissertation is structured in separate chapters as follows.

In the second Chapter, we investigate the data quality challenges in COVID-19 times as the ultimate and most recent distress situation and the reliability and validity of datasets from several international organizations, which are following restrict standards in producing data and statistics.

In the third Chapter, we discuss credit risk as the single most significant source of risk for banks and loan providers. Improved credit risk management is important for both individual bank risk management and the systemic

modelling, measurement, and management of financial stability. We develop a new machine learning approach for online credit scoring which is conservative and robust against unstable and high-risk situations. This Chapter is aimed at the case of credit scoring in risk management and presents a novel method to be used for the default prediction of high-risk branches or customers. The chapter develops an online non-parametric credit scoring system, which can reselect effective features automatically for continued credit evaluation.

In the fourth Chapter, we focus on insurance companies and pension funds as another domain in financial sectors with a prominent impact on financial stability. For this purpose, we first develop a new ensemble learning strategy for panel time-series forecasting. We use a layered learning approach, which is a solution related to ensemble learning to address a given predictive task by different predictive models when direct mapping from inputs to outputs is not accurate. We adopt a layered learning approach to an ensemble learning strategy to solve the predictive tasks with improved predictive performance and take advantage of multiple learning processes into an ensemble model.

In the fifth Chapter, we combine our new ensemble learning approach for time series modeling, discussed in the previous chapter, with K-means clustering for forecasting mortality and life table construction in COVID-19 times. There is no general method to forecast mortality rate applicable to all situations especially for unusual years such as the COVID-19 pandemic. In this Chapter, we investigate the feasibility of using an ensemble of traditional and machine learning time series methods to empower forecasts of age-specific mortality rates for groups of countries that share common longevity trends. We use Generalized Age-Period-Cohort stochastic mortality models to capture age and period effects, apply K-means clustering to time series to group countries following common longevity trends, and use ensemble learning to forecast future longevity markers and annuity prices.

1.5 METHODS AND TECHNOLOGIES

This dissertation's methodological approach follows innovation in targeted learning from (Big) Data and operators of Spark as a novel Big Data programming framework. Taking into account the different data characteristics

in finance and econometrics, we consider a new machine learning approach for streaming data and a new ensemble time-series approach for longitudinal data.

- We focus on Targeted learning and its pre-specified analytic plans and algorithms to make flexible nonparametric or semiparametric machine learning models.
- To achieve this goal, the different super learning methods are taken into account for the sake of maintaining the robustness of the modeling when seeking a higher level of scalability and predictive performance.
- Focus on the Map-Reduce workflow. First, we can act on the learning classifier itself with each Map task. Because instances in the small disjuncts are likely to be difficult to predict, we could use Boosting algorithms to improve their classification performance. Second, we can also take advantage of the Map-Reduce programming scheme focusing on the Reduce stage. Specifically, we must analyze two different schemes for the classification techniques: (1) carrying out a model aggregation (fusion) from the outputs of every Map process, or (2) building an ensemble system and combine their predictions during the inference process.
- For the time series, we investigate the feasibility of using an ensemble of traditional and machine learning time series methods. We improve the quality of forecasting for high-volatile time series with rapid changes in their trends, which could be appropriate for financial data in extreme situations.

For the technology, we use SAS, R software, Python, Spark, and Microsoft PowerBI.

1.6 RESEARCH PATH

This dissertation gathers the findings of several research projects, reported separately, including three papers published in journals with double-blind review process (indexed in Scimago and ISI Thomson Reuters). One journal is among the top 5% of indexed journals in the field of Artificial Intelligence.

Additionally, it is supported by three conferences presentations, of which two conference proceedings were also indexed at Scopus. A part of this dissertation won the first prize for the best paper presented at the 8th International Conference on Risk Analysis and Design of Experiments.

Except for the introduction and the conclusion sections, all other chapters are supported by work published in scholarly publications with a double-blind review process, including first quartile (Q1) journals. This can be regarded as a

positive indication of the work quality that supports this dissertation. The highest quartile range reported to each journal concerns the latest available Scimago ranking (2020).

| Ch. | Study Title | Current State |
|-----------|--|---|
| Chapter 2 | A study on the quality of novel coronavirus (COVID-19) official datasets | Published in Statistical Journal of the IAOS. ▶ |
| | Corona-virus disease (COVID-19) Data-set with Improved Measurement Errors of Referenced Official Data Sources | Published in Mendeley Data. ▶ |
| Chapter 3 | A conservative approach for online credit scoring | Published in Expert Systems with Applications. ▶ |
| | Spark Code: A Novel Conservative Approach for Online Credit Scoring [Source Code] | Published in CodeOcean, awarded Reproducible Badge with Expert Systems with Applications. ▶ |
| | Mining Big Data in statistical systems of the monetary financial institutions (MFIs). | International Conference on Advanced Research Methods and Analytics, CARMA2018 ▶ (SCOPUS indexed proceeding) |
| Chapter 4 | A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic | Published in the CAPSI 2021. ▶ (SCOPUS indexed proceeding) |
| | A New Ensemble Learning Strategy for Panel Time-Series Forecasting with Applications to Tracking Respiratory Disease Excess Mortality during the COVID-19 pandemic | Submitted for publication. ▶ |
| Chapter 5 | Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach | Published in the CISTI 2021 - 16th Iberian Conference on Information Systems and Technologies. ▶ (SCOPUS indexed proceeding) |

CHAPTER TWO - A STUDY ON THE QUALITY OF OFFICIAL DATASETS IN DISTRESS SITUATIONS SUCH AS NOVEL CORONAVIRUS (COVID-19) TIMES

Policymakers depend on complex epidemiological models that are compelled to be robust, realistic, defensible, and consistent with all relevant available data disclosed by official authorities, which is deemed to have the highest quality standards. This Chapter analyses and compares the quality of official datasets available for COVID-19. We used comparative statistical analysis to evaluate the accuracy of data collection by a national (Chinese Center for Disease Control and Prevention) and two international (World Health Organization; European Centre for Disease Prevention and Control) organisations based on the value of systematic measurement errors. We combined excel files, text mining techniques, and manual data entries to extract the COVID-19 data from official reports and to generate an accurate profile for comparisons. The findings show noticeable and increasing measurement errors in the three datasets as the pandemic outbreak expanded and more countries contributed data for the official repositories, raising data comparability concerns and pointing to the need for better coordination and harmonized statistical methods. The study offers a COVID-19 combined dataset and dashboard with minimum systematic measurement errors, and valuable insights into the potential problems in using databanks without carefully examining the metadata and additional documentation that describe the overall context of data¹.

2.1 INTRODUCTION

The local outbreak of pneumonia detected in December 2019 in Wuhan (Hubei, China), later determined to be caused by a novel coronavirus denominated severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has since spread rapidly to every province of mainland China as well as more than 200

¹ Please cite this chapter as: Ashofteh, A., & Bravo, J. M. (2020). A study on the quality of novel coronavirus (COVID-19) official datasets. Statistical Journal of the IAOS, 36(2), 291–301. <https://doi.org/10.3233/SJI-200674>

other countries/regions, with more than 3,4 million confirmed cases as of 2 May 2020, threatening human lives and significantly disrupting the world economy and society (World Health Organization, 2021). The special characteristic of this new virus is how it spread undetected for weeks, which exposed the tardiness and unpreparedness of health systems since its outbreak. Governments and public health systems need accurate and agile information about the characteristics and behaviour of COVID-19 to respond to this ongoing public health emergency appropriately. Researchers, public health authorities, and the general public will benefit from reliable and expeditious data to evaluate the impact of the Coronavirus pandemic on health care systems and to plan for an appropriate policy response at all levels of government (Cheng et al. 2009). Currently, governments and policymakers throughout the world are being forced to make decisions and take actions based on alternative mathematical models developed for other diseases and/or the experience of other countries in which the outbreak has been detected early and developed. In this situation, high-quality institutional-based datasets are the prerequisite of necessary analysis for public health, which is inherently a data-intensive domain (WHO, 2008). Effective data quality assessment in the data collection process would guarantee concordant outcomes from different studies worldwide.

There are several institutional-based repositories of public health data with the capability of electronic data collection and dissemination such as the datasets of public health information systems (PHIS), with various data quality assessment methods and standards (WHO, 2008). However, poor data quality or coding errors in PHIS is not a new issue and can lead to inaccurate inferences of health interventions (Chen et al. 2014). For COVID-19, multi-source datasets of the “World Health Organization (WHO)”, “European Centre for Disease Prevention and Control” and “Chinese Center for Disease Control and Prevention (Chinese CDC)” are reputable references for global BI dashboards and academic research, comprising measures of confirmed, deaths, severe, suspected and recovered cases. These resources are widely used to monitor trends in the virus outbreak and assess the risks of the pandemic in several countries and regions.

This study assesses the systematic measurement errors, completeness, accuracy, and timeliness of the mentioned official datasets for COVID-19 by

using text-mining, reviewing reports, metadata and reference data to extract the essential information for qualitative and quantitative assessment. As we are in the primary stage of this world pandemic, our goal is to investigate and compare the official COVID-19 datasets for data-quality assessment to identify potential improvements and to provide a novel combined dataset with minimum systematic measurement errors to be used by researchers and decision-makers. The findings show noticeable and increasing measurement errors in the three datasets as the pandemic outbreak expanded and more countries contributed data for the official repositories, raising data comparability concerns and pointing to the need for better coordination and harmonized statistical methods. The presence of measurement errors causes biased and inconsistent parameter estimates and leads to erroneous conclusions to various degrees in epidemiological analysis. We provide a corrected dataset incorporating our findings of the necessary corrections of these data sources, imputation of missing values, outlier treatment, and adjusting the date attribute, which we concluded were suffering from a one or two-day lag. This data set with 11,838 rows and 37 attributes and minimal measurement error is available for further research and the users of these official data sources (Ashofteh & M. Bravo, 2020). The authors provide also a dedicated data dashboard for an online visual summary of the main findings of this article, which is available online as a graphical abstract (Ashofteh & M. Bravo, 2020).

The description of the dataset comparisons provides valuable insights into the potential problems in using databanks that are the repository of information from many countries without carefully examining the metadata and additional documentation that describe the content and the overall context of data. Developing guidelines, standards, and ontologies for data documentation is crucial for researchers and policymakers in terms of understanding the context of data creation and collection. Moreover, the altering way in which confirmed cases and deaths have been classified in China points to similar problems which may arise in other countries which require careful forensic analysis regularly to understand how definitions are applied and to what extent data are comparable. There is a growing need for harmonization and standardization of the data gathering, reporting and data analysis processes.

In epidemic modelling, there is an increasing need to exploit information from multiple conventional and non-conventional sources, ensuring decision-making on public health policies geared to control epidemics is progressively data and model driven (De Angelis et al., 2015; Rutherford et al., 2010). Several epidemiological models of COVID-19's outbreak and spread have been used to provide a preliminary assessment of the magnitude and timeline for confirmed cases, long-term predictions of deaths or hospital utilization, the effects of quarantine, stay-at-home orders and other social distancing measures, travel restrictions or the pandemic's turning point. The accuracy and validity of these models crucially depends on data availability and quality. The impact on epidemiological models of the errors that can be found in the international databases is of matter of great concern since these models will continue to be used worldwide to inform national and local authorities on how to implement an adaptive response approach to re-opening the economy, re-open schools, alleviate business and social distancing restrictions, allow sports events to resume. To highlight these problems, we provide a brief study of the impact of imported cases on model fitting considering the data for China and to underline the implications for models developed in countries where imported cases have been prominent in triggering the pandemic there.

Although this analysis is being conducted at a relatively early stage of the epidemics and, in the course of time, additional data sets have become available, the Chapter approach on the identification of measurement errors remains timely, useful, and important. Indeed, we show that the significant challenges posed by the epidemic context offer a renovated opportunity to improve the quality of official statistical methodology, particularly where several datasets may be needed to inform an epidemiological model. The Chapter also contributes to the ongoing discussion triggered by the Statistical Journal of the IAOS (SJIAOS) on the need for good (old and new) official statistics in the preparation of the important political decisions required to tackle the problems that will be at the top of the agenda in the next phases of the crisis management (e.g, economic recovery plans, unemployment, collateral illnesses (depression, suicide), domestic violence), as well as to address all the topics that were given lower priority in the short-term crisis (e.g., UN Sustainable Development Goals, reducing poverty and inequality, climate change and biodiversity challenges) that will shape the world of

tomorrow.² The current experience also shows that the preparation and dissemination of official statistics contributes to reduce the “pandemics of fear” and “fake news” that either try to minimize or overstate the severity of the public health threat, eroding trust in public health authorities, potentially reducing compliance with essential protective guidance. The structure of the remaining of this Chapter is as follows: Section 2 provides a brief description of the official COVID-19 datasets and how the data was handled. Section 3 describes the data and methods used in this study. Section 4 presents and discusses the main results of the investigation. Finally, Section 5 concludes.

2.2 OFFICIAL COVID-19 DATASETS: AN OVERVIEW

2.2.1 World Health Organization (WHO) reports

The World Health Organization (WHO) has been in regular and direct contact with Chinese as well as authorities in other countries since the reporting of their cases. It provides daily situation reports for within and outside of mainland China. These situation reports include the raw data and the metadata, in pdf format files, to represent the numbers and inform the developments of public health policies such as quarantine and the establishment of priorities such as urgent research for implementing surveillance of this new disease (World Health Organization, 2021). The first report was published on January 21 2020, with a small table consisting of four countries and included four territories or areas of China with reported confirmed cases of 20 January 2020. There are informative details about the reported cases, Wuhan City, and the surveillance and preparedness in all infected countries. We loaded the data by using a semi-automated table recognition strategy for the WHO pdf files and read the contents of the reports for additional data or information by purpose. The structure of pdf files was not similar, and the number of tables was not fixed. Therefore, it was difficult to read their data fully automatically, and we interfered manually to adjust the program several times. The result was a table with 11,838 rows of time-series data referring to countries and nine columns consisting of attributes, namely, Row, Date, Country Code, WHO Region,

² See www.officialstatistics.com for details on the ongoing discussion on the role of Official Statistics in the context of the COVID-19 crisis and in shaping the world of tomorrow.

Country/Territory/Area, Confirmed Cases, New Cases, Total Deaths, and New Deaths, a sample of which is shown in Table 2.1.

TABLE 2.1 – SAMPLE OF DATA FROM SITUATION REPORTS OF THE WORLD HEALTH ORGANISATION (WHO).

| row | Date | Code | Area | Country | confirmed cases | confirmed new cases | deaths | new deaths |
|------|----------|------|--------------------------------------|-----------|-----------------|---------------------|--------|------------|
| 2501 | 20200316 | CN | Western Pacific Region | China | 81077 | 29 | 3218 | 14 |
| 2514 | 20200316 | IT | European Region | Italy | 24747 | 3590 | 1809 | 368 |
| 2569 | 20200316 | ID | South-East Asia Region | Indonesia | 117 | 0 | 4 | 0 |
| 2577 | 20200316 | IR | Eastern Mediterranean | Iran | 14991 | 2262 | 853 | 245 |
| 2594 | 20200316 | US | Region of the Americas | USA | 1678 | 0 | 41 | 0 |
| 2628 | 20200316 | ZA | African Region | S. Africa | 51 | 13 | 0 | 0 |
| 2652 | 20200316 | * | Cruise ship Diamond Princess (Japan) | Other | 712 | 15 | 7 | 0 |

* JPG11668 is considered as the code of Diamond Princess Cruise Ship.

Source: Author's preparation based on the WHO.

Data entry and number verification took several days to avoid systematic data collection errors. This process could have been fully automated had the number of tables in the different pdf files, and the structure of the tables been fixed by the WHO. However, as the outbreak evolved the manual collecting and reporting process became unsustainable.

2.2.2 European Centre for Disease Prevention and Control dataset (ECDC)

A data file in Excel format and the appropriate R software code to read the file from its source are available on the ECDC website (ECDC, 2020). It is updated daily and contains the latest available public data on COVID-19. This data file put the attributes of Date, Day, Month, Year, Confirmed Cases, Death, Name of the Country, population in 2018 (Population Division - United Nations, 2019) and alpha-2/alpha 3 Country code (International Organization on Standardization, 2006; Statistics Division - United Nations, n.d.) in columns. The date and alpha-2 country code attributes are useful to be concatenated as a single code for merging different databases and putting the numbers in the

corresponding rows of the query dataset. We used this strategy to find the unique rows in the different datasets and to make a unique dataset for our further analysis. In this case, the date and code of countries should be accurate to allow users to manipulate the data and use it for statistical analysis or reporting purposes. Name of countries is not recommended, because they might be written in different ways, especially for countries with separate names, which could be compiled with dashes, parentheses, or blanks.

2.2.3 Chinese Center for Disease Control and Prevention (Chinese CDC)

The Chinese CDC Weekly website makes daily reports available for the public via their online portal (National health commission - China). This platform has started to publish COVID-19 reports, by using various national data sources from 19 January, 22:00 CST (UTC+8). Some crucial information is in the contents of the reports, and an important point is the report dates. The website provides the statistics of the previous 24-hour day, every day. However, in the summary statistics at the top of the webpage, this one-day lag is not mentioned.

TABLE 2.2 – TOP 10 ROWS OF AGGREGATED ATTRIBUTES OF COVID-19 FOR CHINA - WESTERN PACIFIC REGION.

| row | Date | Total confirmed cases in PLADs | Total severe cases | Total deaths | total recovered and discharged | total suspected cases | Total confirmed cases in Regions | Total Confirmed in PLADs and Regions |
|-----|----------|--------------------------------|--------------------|--------------|--------------------------------|-----------------------|----------------------------------|--------------------------------------|
| 1 | 20200118 | 62 | | | | | 0 | 62 |
| 2 | 20200119 | 198 | | | | | 0 | 198 |
| 3 | 20200120 | 291 | | | | | 0 | 291 |
| 4 | 20200121 | 440 | | 9 | | | 0 | 440 |
| 5 | 20200122 | 571 | 95 | 17 | | | 3 | 574 |
| 6 | 20200123 | 830 | 166 | 25 | | | 8 | 838 |
| 7 | 20200124 | 1287 | 237 | 41 | 38 | 1965 | 18 | 1305 |
| 8 | 20200125 | 1975 | 324 | 56 | 49 | 2684 | 28 | 2003 |
| 9 | 20200126 | 2744 | 461 | 80 | 51 | 5794 | 45 | 2789 |
| 10 | 20200127 | 4515 | 976 | 106 | 60 | 6973 | 65 | 4580 |

Source: Author's preparation based on the Chinese CDC.

TABLE 2.3 - TOP 10 ROWS OF NEW ATTRIBUTES OF COVID-19 FOR CHINA - WESTERN PACIFIC REGION.

| row | Date | New confirmed cases in PLADs | New severe cases | New deaths | New recovered and discharged | New suspected | New confirmed cases in Regions | New confirmed cases in PLADs and Regions | close contacts | have been released | under medical observation |
|-----|----------|---------------------------------|------------------|------------|---------------------------------|---------------|-----------------------------------|---|----------------|--------------------|------------------------------|
| 1 | 20200118 | | | | | | | 0 | | | |
| 2 | 20200119 | 136 | | | | | | 136 | | | |
| 3 | 20200120 | 77 | | | | | | 77 | | | |
| 4 | 20200121 | 149 | | 3 | | | 0 | 149 | | | |
| 5 | 20200122 | 131 | | 8 | | 257 | 3 | 134 | 5897 | 969 | 4928 |
| 6 | 20200123 | 259 | | 8 | 6 | 680 | 5 | 264 | 9507 | 1070 | 8437 |
| 7 | 20200124 | 444 | | 16 | 3 | 1118 | 10 | 454 | 15197 | 1230 | 13967 |
| 8 | 20200125 | 668 | 87 | 15 | 11 | 1309 | 10 | 678 | 23431 | 325 | 21556 |
| 9 | 20200126 | 769 | 137 | 24 | 2 | 3806 | 17 | 786 | 32799 | 583 | 30453 |
| 10 | 20200127 | 1771 | 515 | 26 | 9 | 2077 | 20 | 1791 | 47833 | 914 | 44132 |

Source: Author's preparation based on the Chinese CDC.

Therefore, if users try to extract data by web scraping or simply look at the data in the summaries at the top of the website and do not pay enough attention to the metadata in the full reports or the references, presented at the bottom of the webpage or links, then the day of extracted data will be biased for the one-day lag. As a result, for extracting the data from CCDC's reports, we used text mining along with reading full reports and references to make a reliable base for checking the two other official data sources mentioned (namely WHO and ECDC) for China.

The extracted dataset of China includes 23 attributes in a time-series format. A sample view with the top 10 rows is shown in Tables 2.2 and 2.3.

2.3 METHODS

2.3.1 Data

The data used in this study is from the repositories of the World health organisation (WHO), the European Centre for Disease Prevention and Control dataset (ECDC) and the Chinese Center for Disease Control and Prevention (Chinese CDC). First, we performed the text mining and loaded the data of the reports from the pdf files and websites along with the perusal of the full reports. Then, by reading the first eight characters of the country names, the alpha-2 codes were added to all rows of these datasets, combined with the Date variable for each row to make a unique primary key for each country and each day. This primary key was used to combine these three datasets into one. A manual search of the reports and dataset metadata was conducted to improve accuracy and to identify new attributes and statistics inside the text of the reports together with some new information referenced by other publications or well-known communities. For instance, data referring to 17 November and 20 December 2019, were added to those mentioned datasets. An Analytical Base Table of the combined data sources is shown in Table 2.4 (Ashofteh & M. Bravo, 2020).

TABLE 2.4 - ANALYTICAL BASE TABLE (ABT) OF JOINED DATA SOURCES.

| Attribute | Description | Additional Information |
|-------------|--|--|
| Row | Row number | It is useful to sort the dataset to its original order. |
| Date | Date of the referenced day | Date in the yyymmdd format referenced to the past 24 hours of the date mentioned. |
| Year | Year of referenced day | Year in the yyyy format. |
| Month | Month of referenced day | Month in the mm format. |
| Day | Referenced day | Day in the dd format. |
| Area | WHO region | The World Health Organization divides the world into six WHO regions, for the purpose of reporting, analysis and administration. |
| Country | Name of country | Name of countries based on WHO reports. |
| Country_Num | M49 code | Standard country or area codes for statistical use. |
| Alpha-2 | Abbreviation code of the country – Two letters | Includes two letters for each country, except for JPC11668, which is allocated to the Diamond Princess Cruise Ship (Japan). |
| Alpha-3 | Abbreviation code of the country – Three letters | Includes three letters for each country, except for JPC11668, which is allocated to the Diamond Princess Cruise Ship (Japan). |
| latitude | Latitude of the country | |
| longitude | Longitude of the country | |
| Population | Total population of the country (thousands) | From World Population Prospects 2019, United Nations, Department of Economic and Social Affairs. |

| | | |
|--------------------|---|--|
| WHO_TCC | WHO Total confirmed cases | Total confirmed cases are the aggregation of confirmed cases during the time, including both laboratory-confirmed and clinically diagnosed cases in WHO reports. |
| WHO_NCC | WHO New confirmed cases | New confirmed cases is similar to WHO_TCC but for new cases in WHO reports. |
| WHO_TD | WHO Total deaths | Cumulative aggregation of deaths in WHO reports. |
| WHO_ND | WHO New deaths | Number of new deaths in WHO reports. |
| CCDC_TCC | CCDC Total confirmed cases | Cumulative aggregation of confirmed cases includes both laboratory-confirmed and clinically diagnosed cases in CCDC reports. |
| CCDC_NCC | CCDC New confirmed cases | New confirmed cases are similar to CCDC_TCC but for new cases in CCDC reports. |
| CCDC_TD | CCDC Total deaths | Cumulative aggregation of deaths in CCDC reports. |
| CCDC_ND | CCDC New deaths | Number of new deaths in CCDC reports. |
| ECDC_TCC | ECDC Total confirmed cases | This column is calculated from ECDC_NCC by author's. |
| ECDC_NCC | ECDC New confirmed cases | New confirmed cases in the ECDC public dataset. |
| ECDC_TD | ECDC Total deaths | This column is calculated from ECDC_ND by author's. |
| ECDC_ND | ECDC New deaths | Number of new deaths reported in the ECDC public dataset. |
| TCC_authors | Corrected total confirmed cases | Total confirmed cases with measurement error correction by authors. |
| NCC_authors | Corrected new confirmed cases | New confirmed cases with measurement error correction by authors. |
| TD_authors | Corrected total deaths | Total deaths with measurement error correction by authors. |
| ND_authors | Corrected new deaths | New deaths with measurement error correction by authors. |
| MR | Mortality rate | Mortality rate (TD_authors/Population) based on measurement error correction by authors. |
| FR | Fatality rate | Fatality rate (TD_authors/TCC_authors) based on measurement error correction by authors. |
| TCC/Pop | Corrected TCC adjusted for Population (thousands) | Corrected total confirmed cases with an adjustment for population. |
| NCC/Pop | Corrected NCC adjusted for Population (thousands) | Corrected new confirmed cases with an adjustment for population. |

Source: Author's preparation. DOI: 10.17632/nw5m4hs3jr.3

2.3.2 Errors and Outliers

We checked the new dataset for negative numbers and discovered four negative values in the attribute of new confirmed cases in the ECDC dataset, as shown in Table 2.5.

Chapter II – QUALITY OF DATASETS IN DISTRESS SITUATIONS

TABLE 2.5 - NEGATIVE VALUES IN DATASETS.

| Date | Code | Area | Country | WHO_TCC | WHO_NCC | ECDC_NCC |
|----------|----------|--------------------------------|-----------|---------|---------|----------|
| 20200310 | KH | Western Pacific Region | Cambodia | 2 | 0 | -9 |
| 20200310 | JPG11668 | Cruise ship (Diamond Princess) | Other | 696 | 0 | -9 |
| 20200419 | ES | European Region | Spain | 191726 | 3658 | -1430 |
| 20200429 | LT | European Region | Lithuania | 1449 | 0 | -105 |

Source: Author's preparation.

As evident in the first row of Table 2.5, the value of minus nine is not possible when the total infected is two. Some official statistics authorities usually use the digit 9 for unknown situations; however, in this case, we did not find any evidence of this tradition. Also, the WHO reported zero new confirmed cases for the Diamond Princess Cruise ship on 10 March 2020. Therefore, we corrected these four negative values, according to the WHO reported values.

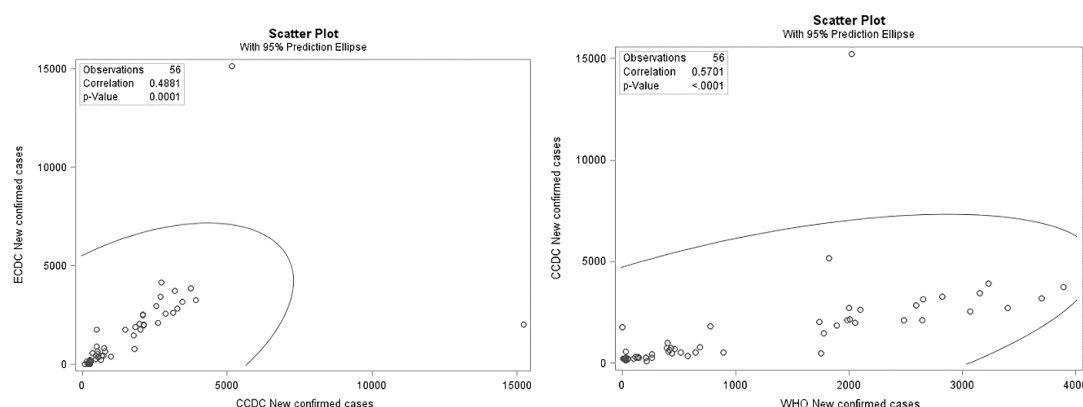


FIGURE 2-1 - SCATTER PLOT AND CORRELATION BETWEEN CCDC REPORTS WITH WHO, AND ECDC REPORTS.

In Figure 2-1, we can see the correlation between the three datasets for new confirmed cases in China, which is less than 0.60. Because China (Wuhan, Hubei) was the first place to face the COVID-19 outbreak, one might expect the Chinese data to be completer and more robust when compared to other countries. Nevertheless, the correlations among the CCDC dataset and the two other official datasets are very low as presented in Figure 2-1, especially for attributes which should have almost the same values. As discussed, the authors extracted the CCDC data directly from the official CCDC website, which is

assumed to be a reliable source for the comparisons. These corrections were not enough to significantly reduce the distortions in these datasets. Indeed, the correlation between new confirmed cases reported by the WHO and ECDC (Figure 2-2) continues to be less than 0.60, which is still considered to be a small number, but we can now observe that the distortion is slightly smaller than that in Figure 2-1 and the correlation is almost linear.

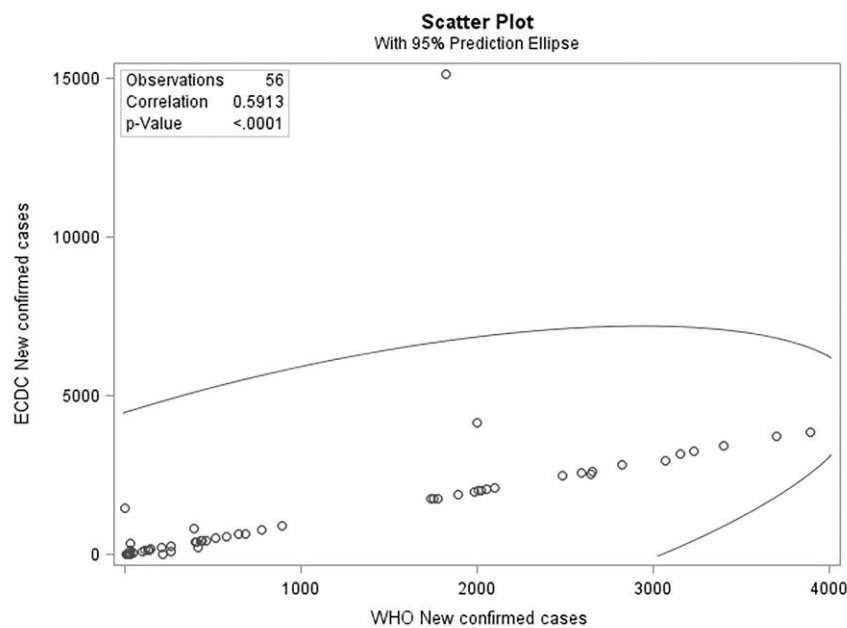


FIGURE 2-2 - CORRELATION OF NCC BETWEEN WHO AND ECDC REPORTED DATA.

One attribute which could make this situation possible is the calendar date variable. Therefore, we checked the date variable and corresponding values in the three datasets. We determined that the values of this variable suffer from a one-day lag between the different datasets as follows. The WHO reports were initiated on 21 January 2020 and, as mentioned, in the first report that date refers to the occurrences on 20 January. Subsequently, the January 22nd report communicated the January 21st statistics. However, in the January 23rd report, the date as reported was also 23 January and included the information reported to the WHO Geneva at 10 AM CET. It means that the WHO has no data for 22 January or it is aggregated with the January 23rd data. However, we detected a one-day lag in the WHO statistics compared to the correspondent values from China, based on the CCDC daily reports. It means that the WHO daily situation reports were shifted forward for one day on 23 January and should

Chapter II – QUALITY OF DATASETS IN DISTRESS SITUATIONS

consequently be corrected from this date. Similarly, the ECDC dataset manifested the same systematic measurement error.

This distortion was judged to need correction because, as mentioned, it is common to use the date attribute and country codes to create a primary key for these kinds of datasets. Furthermore, the exact report dates were essential to evaluate the outcomes of policy interventions and the effectiveness of public health measures to reduce the disease severity. In this regard, even a small error in the date of clinical reports can change the clinical data analysis explanations and results and wrongly inform decision makers.

The data analysis also identified some outliers, which are shown in Figures 2.1 and 2.2. Finally, in the first four days, the values presented in the reports were dramatically different, and there were especially acute different values for some other days in some parts of datasets. The root mean square errors of attributes in the paired comparison of datasets were noticeable and increasing with time as the pandemic outbreak expanded and more countries contributed data for the official datasets (Table 2.6). This points increasing risks on the use of inaccurate datasets as the pandemic develops and global modelling and comparisons is made.

TABLE 2.6 - ROOT MEAN SQUARE ERRORS OF ATTRIBUTES OF DIFFERENT REPORTS.

| | | TCC ¹ of WHO & CCDC | NCC ² of WHO & CCDC | NCC of WHO & ECDC | NCC of CCDC & ECDC | TD ³ of WHO & CCDC | ND ⁴ of WHO & CCDC | ND of WHO & ECDC | ND of CCDC & ECDC |
|--------|------|--------------------------------------|--------------------------------------|-------------------------|--------------------------|-------------------------------------|-------------------------------------|------------------------|-------------------------|
| 31 Jan | RMSE | 73.44 | 432.96 | 123.24 | 28.22 | 123.23 | 25.75 | 7.38 | 1.04 |
| | N | 12 | 12 | 147 | 12 | 12 | 12 | 147 | 12 |
| 29 Feb | RMSE | 2444.9 | 2166.94 | 419.46 | 190.85 | 71 | 53.44 | 11.08 | 16.89 |
| | N | 41 | 41 | 1050 | 41 | 41 | 41 | 1050 | 41 |
| 31 Mar | RMSE | 4965.82 | 1663 | 300.63 | 343.45 | 53.69 | 40.34 | 10.41 | 12.8 |
| | N | 72 | 72 | 5689 | 72 | 72 | 72 | 5689 | 72 |
| 30 Apr | RMSE | 1591.67 | 393.3 | 805.33 | 137.16 | 132.8 | 50.75 | 128.82 | 1591.67 |
| | N | 101 | 11836 | 101 | 101 | 101 | 11836 | 101 | 101 |

Source: Author's preparation. **Notes:** 1-Cumulative aggregation of confirmed cases; 2-New confirmed cases; 3-Cumulative aggregation of deaths; 4-Number of new deaths.

As a result, we reviewed the resources and looked for the logic behind the irregular values of these attributes. From Figure 2-3, we noted the first problematic dates are 12 and 13 February.

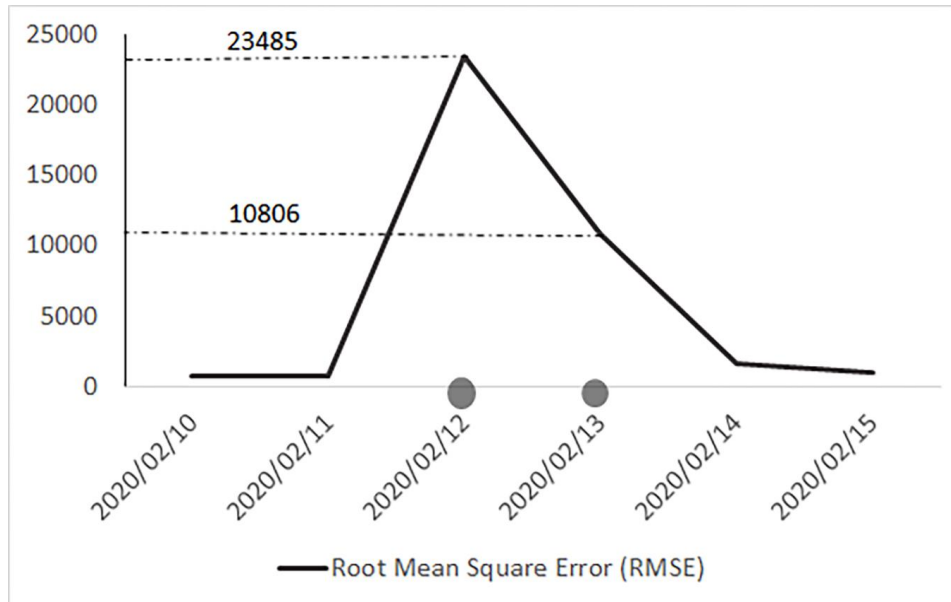


FIGURE 2-3 - DAILY SUM OF SQUARE ERROR AGGREGATED FOR ALL ATTRIBUTES, COUNTRIES AND DATASETS.

We discerned that the structure of the WHO reports was changed several times on these dates. For instance, the report structure was changed on 13 February 2020, and total deaths and total new deaths were no longer reported. By comparing to the other reports, we could conclude that the WHO became aware of the fact that the Chinese data only referred to laboratory-confirmed cases and did not include clinically diagnosed ones. As a result, in the next report, the report structure was changed once again. On 14 February 2020, instead of reporting China as a whole in the table of countries, the table of Chinese provinces, regions and cities was extended with additional information for laboratory-confirmed and clinically diagnosed cases, and a total number for China could be read from the column aggregates. From this report and comparing the numbers, we could conclude that the numbers, which were previously reported under the “Confirmed Cases” nomenclature only included laboratory-confirmed incidents and not clinically diagnosed ones. Therefore, we could observe a jump in confirmed cases in these three official data sources on 12 and 13 February. This time series leap is what analysts should not consider as a real surge, showing a special treatment of

COVID-19 or a real pick in the distribution of data. The use of smoothing techniques could be recommended to researchers for this part of the data sets.

Again, in the 17 February 2020 report, the Chinese table structure was changed to one aggregated column in the WHO situation report, including “reported laboratory-confirmed” and “clinically diagnosed”. Finally, in the 2 March 2020 report, the structure of countries table was changed yet again and the number of new cases and new deaths, which were previously reported in parentheses in front of total confirmed cases and total deaths in the same columns, were separated into new columns. As a result, for the purpose of this research and using the WHO data as one of the main resources, data entry for these days was done manually by the researchers and the missing total deaths and total new deaths relative to 13 February were imputed by using interpolation and available information from 12 to 14 February.

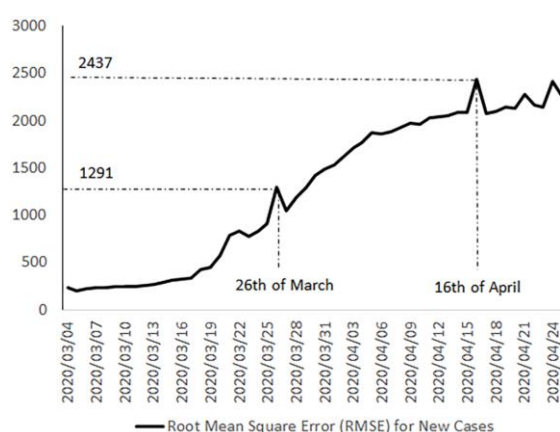


FIGURE 2-4 - POSITIVE TREND IN THE ROOT MEAN SQUARE ERROR AGGREGATED FOR ALL ATTRIBUTES RELATED TO NEW CASES FOR ALL COUNTRIES IN THE THREE REFERENCE DATASETS.

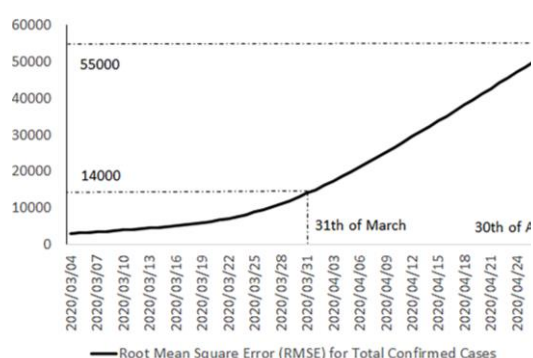


FIGURE 2-5 - POSITIVE TREND IN THE ROOT MEAN SQUARE ERROR AGGREGATED FOR TOTAL CONFIRMED CASES OF ALL COUNTRIES IN THE THREE REFERENCE DATASETS.

Finally, in Figures 2.4 and 2.5, we can see a positive trend for errors in recent last days, which could be considered as an alert for serious inhomogeneity of these three public official data sources. It seems that by increasing the reported positive cases and the epidemic of COVID-19 in more countries, the homogeneity of these data sets is decreasing.

2.4 RESULTS

The main outcome of our analysis is showing an increasing measurement error in the three datasets as the the pandemic outbreak expanded and more

countries contributed data for the official repositories, an estimation of the distribution of new positive cases in China, and an extracted, and corrected dataset from the WHO situation reports, the ECDC dataset and CCDC daily reports, plus one extra row at the beginning of the dataset, related to the first infected person as the COVID-19 Patient-Zero, which was reported on 17 November 2019 in China. The corrected dataset incorporates our findings of the necessary corrections of these data sources, imputation of missing values, outlier treatment and adjusting the date attribute, which we concluded were suffering from a one or two-day lag. For China, we considered the CCDC reports and the maximum of cumulative values by the WHO and ECDC for other countries. It includes the data from the Hong Kong Special Administrative Region of (China), Macau Special Administrative Region (China), and Taiwan (China).

For other countries, we suggested the maximum values for aggregated attributes such as total confirmed cases, because of the time lag of the reports for the preceding 24 hours and the different updating time of reports which suggests the maximum as a most recent reported value by countries. If the difference between the CCDC and WHO reported values was more than double, we did not apply the maximum anymore but selected the WHO value as a reference instead. This data set with 11,838 rows and 37 attributes and minimal measurement error is available for further research and the users of these official data sources. The authors designed a data dashboard for an online visual summary of the main findings of this article, which is available online as a graphical abstract (Ashofteh & M. Bravo, 2020).

Another table with more COVID-19 attributes, which is extracted by text mining from the CCDC daily reports and its related metadata review and supporting documents with double-checking by the authors, was specified to China (Ashofteh & M. Bravo, 2020).

Finally, the distribution of new positive cases in China was studied by using our new dataset. We considered the attribute of date as our main variable and the number of new positive cases as corresponding frequencies. According to the shape of the data, we candidate some distributions such as Gamma,

Chapter II – QUALITY OF DATASETS IN DISTRESS SITUATIONS

Weibull and Lognormal (Table 2.7). Then we used the root mean square error to compare these candidate distributions.

TABLE 2.7 - COMPARING DISTRIBUTIONS BASED ON RMSE.

| Distribution | Gamma | Weibull | Lognormal |
|--------------|----------|-----------|-----------|
| Quantiles | Observed | Estimated | Estimated |
| 1% | 67 | 66.7031 | 63.6799 |
| 5% | 71 | 70.3026 | 67.9578 |
| 10% | 73 | 72.6388 | 70.9318 |
| 25% | 78 | 77.2546 | 76.8001 |
| 50% | 83 | 83.5129 | 84.2636 |
| 75% | 90 | 91.0554 | 92.2934 |
| 90% | 99 | 99.0123 | 99.7472 |
| 95% | 107 | 104.3273 | 104.2483 |
| 99% | 120 | 115.4559 | 112.6977 |
| RMSE | 1.84 | 3.24 | 1.13 |

Source: Author's preparation.

We identified that the distribution of new positive cases in China over time is very well expressed by the Lognormal distribution with threshold parameters of Theta equal to 52.4, scale parameter of Zeta equal to 3.43 and 0.32 for Sigma as shape parameter (Figure 2-6).

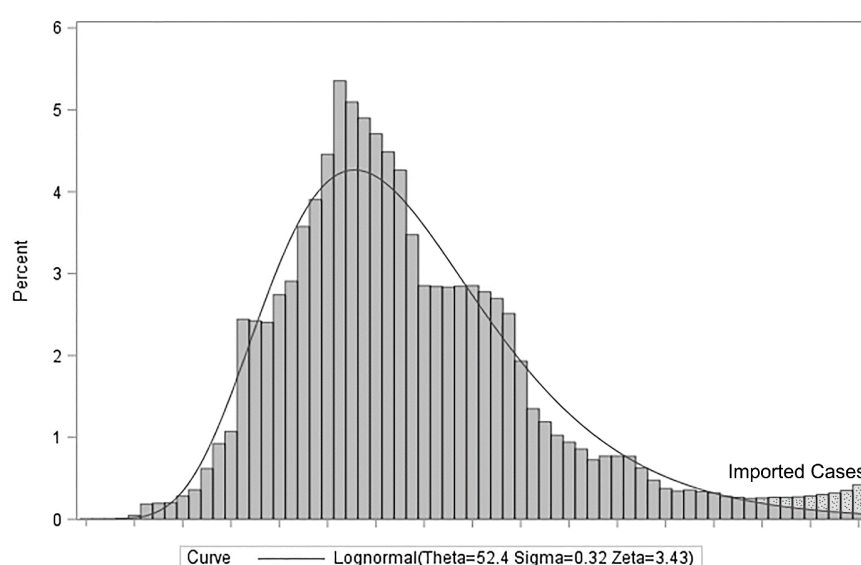


FIGURE 2-6 - THE LOGNORMAL DISTRIBUTION FOR ATTRIBUTED NEW CONFIRMED CASES IN CHINA.

As shown in Figure 2-6, the right tail of the distribution is not fitted appropriately. We investigated this situation by checking the CCDC daily reports and discovered a new paragraph that was added to the 3 March 2020, for the new imported cases from outside of China. These new cases do not belong to the country, and for the purpose of fitting a distribution to new confirmed cases in China, we should subtract these new imported cases from the corresponding new confirmed cases.

TABLE 2.8 - IMPORTED CASES TO CHINA FROM OUTSIDE.

| row | Date* | Total imported | New imported |
|-----|----------|----------------|--------------|
| 1 | 20200303 | 0 | 18 |
| 2 | 20200304 | 20 | 2 |
| 3 | 20200305 | 36 | 16 |
| 4 | 20200306 | 60 | 24 |
| 5 | 20200307 | 63 | 3 |
| 6 | 20200308 | 67 | 4 |
| 7 | 20200309 | 69 | 2 |
| 8 | 20200310 | 79 | 10 |
| 9 | 20200311 | 85 | 6 |
| 10 | 20200312 | 88 | 3 |

* Please note the one-day lag in the reference reports. One can find the corresponding numbers of row 2 (4 March) on the CCDC website under the date of 5 March.

Source: Author's preparation.

The number of imported cases to China from outside is shown in Table 2.8. As we can see in Table 2.7 and Table 2.9, the observed values for quantile 95% is changed from 107 to 106, and the New RMSE shows a better fitting of the distribution to this new data. However, the Lognormal distribution is still the best suggested one compared to the Gamma and Weibull distributions.

TABLE 2.9 - COMPARING DISTRIBUTIONS BASED ON RMSE WITH CORRECTION FOR IMPORTED POSITIVE CASES.

| Distribution | | Gamma | Weibull | Lognormal |
|--------------|----------|-----------|-----------|-----------|
| Quantiles | Observed | Estimated | Estimated | Estimated |
| 1% | 67 | 66.7031 | 63.6799 | 67.0697 |
| 5% | 71 | 70.3026 | 67.9578 | 70.6231 |
| 10% | 73 | 72.6388 | 70.9318 | 72.8641 |
| 25% | 78 | 77.2546 | 76.8001 | 77.2518 |
| 50% | 83 | 83.5129 | 84.2636 | 83.256 |
| 75% | 90 | 91.0554 | 92.2934 | 90.7295 |
| 90% | 99 | 99.0123 | 99.7472 | 99.0075 |
| 95% | 106 | 104.3273 | 104.2483 | 104.8004 |
| 99% | 120 | 115.4559 | 112.6977 | 117.6933 |
| NEW RMSE | | 1.7 | 3.16 | 0.95 |
| OLD RMSE | | 1.84 | 3.24 | 1.13 |

Source: Author's preparation.

2.5 CONCLUSION

This Chapter assessed the measurement error of three official datasets for COVID-19, currently used as the main references for researchers around the world and domain BI dashboards. These data sources will be used to model the COVID-19 pandemic and apply different methods such as machine learning and time-series algorithms to predict the future. As we know, most of these algorithms work based on computational linear algebra and linear space. This linearity is important to put machines to work. For instance, R software and Python utilise linear algebra packages such as BLAS and LAPACK. Therefore, researchers prefer linear space in comparison to the norm space to be able to take advantage of the different mathematical tools in a vector space and use multivariate analysis, measures of central tendency and variations. As a result, it would be possible to solve complex problems with easy additive univariate models in vector space without the need to create new algorithms. However, the accuracy of these data-driven tools is sensitive to distortions and measurement errors, especially when the dataset is small.

Although we can fit an approximate line, surface or high dimension solution to our data in vector space, on most occasions, we need to smooth the data to take advantage of many tools for optimising smooth functions such as derivatives for optimisation. This smoothness and averaging are also dramatically sensitive to measurement errors. Therefore, even minor measurement errors in official COVID-19 datasets could significantly impact the final outcomes of mathematical models used to forecast the development of this infectious disease. This matter shows the importance of the accuracy, timeliness and completeness of COVID-19 official datasets for better models and interpretations.

We studied three referenced COVID-19 datasets and tried to provide an improved dataset for further studies of researchers. Additionally, this study shows a positive trend in the risk of measurement errors in these official datasets, which could be prevented by responsible authorities with excogitating some precautions. Finally, the distribution of COVID-19 in China was estimated. Our results suggest that the best goodness of fit corresponds to a Lognormal distribution with threshold parameters of Theta equal to 52.4, a scale parameter of Zeta equal to 3.43 and 0.32 for Sigma as a shape parameter. A Gamma distribution with estimated parameters of 58.80 for Theta, 4.25 for Sigma and 6.13 for Alpha is another appropriate candidate, which could be tested into models by researchers. It could help understand the behaviour of COVID-19, considering as a prior for Bayesian methods and estimating the infection rate in different countries.■

CHAPTER THREE - A CONSERVATIVE MACHINE LEARNING APPROACH FOR ONLINE CREDIT SCORING UNDER DISTRESS SITUATIONS SUCH AS COVID-19 TIMES

As it was discussed in the last Chapter, the high-risk conditions are able to affect the standard data production quality and the normal mechanic of data. This Chapter is aimed at the case of credit scoring in risk management as one of the most important sources of risk for monetary institutions. We present a novel method to be used for the default prediction of high-risk branches or customers under high-risk conditions to preserve financial stability. This study uses the Kruskal-Wallis non-parametric statistic to form a conservative credit-scoring model and to study the impact on modeling performance on the benefit of the credit provider. The findings show that the new credit scoring methodology represents a reasonable coefficient of determination and a very low false-negative rate. It is computationally less expensive with high accuracy with around 18% improvement in Recall/Sensitivity. Because of the recent perspective of continued credit/behavior scoring, our study suggests using this credit score for non-traditional data sources for online loan providers to allow them to study and reveal changes in client behavior over time and choose the reliable unbanked customers, based on their application data. This is the first study that develops an online non-parametric credit scoring system, which can reselect effective features automatically for continued credit evaluation and weigh them out by their level of contribution with a good diagnostic ability³.

³ Please cite this chapter as: Ashofteh, A., & Bravo, J. M. (2021). A conservative approach for online credit scoring. *Expert Systems with Applications*, 176, 114835. <https://doi.org/10.1016/j.eswa.2021.114835>

3.1 INTRODUCTION

Credit scoring involves the use of analytical methods to transform relevant data into numerical measures that inform and determine credit decisions. In recent years the use of credit scoring tools has expanded beyond their original purpose of assessing credit risk, such as establishing the initial and ongoing credit limits available to borrowers, assessing the risk-adjusted profitability of account relationships, and assisting in a range of loan servicing activities, including fraud detection, delinquency intervention, and loss mitigation (Thomas, 2000).

The critical role of the lending market in causing the latest global financial crisis has increased academic research, policy interest, and bank regulation in this area. The banking regulatory framework changes brought by the revised Basel Committee on Banking Supervision (BCBS) Accords (later adopted by national legislation in many countries and regions, for instance, the European Capital Requirement Directives and the US Regulatory Capital Rules) introduced stronger risk management requirements for banks, with capital requirements tightly coupled to estimated credit portfolio losses. The recently adopted IFRS9 and FASB's Current Expected Credit Loss (CECL) standards introduce revised expected credit loss or impairment calculation rules requiring financial institutions to calculate the expected loss for the banking book over the entire life of the exposures, conditional on macroeconomic factors, on a point-in-time basis, that is, recalibrating PDs where necessary to reflect the effects of the current economic conditions. Encouraged by regulators, banks devoted significant resources to develop an Internal Ratings Based approach (IRB) for the calculation of risk-weighted assets for credit risk to better support decisions when granting loans, to quantify expected credit losses, and to assign the mandatory economic capital (Chamboko & Bravo, 2020).⁴

To assess credit risk, in developed markets, lenders typically consider historical loan application and loan performance data collected regularly from a small number of sources on the basis of long-standing banking and credit

⁴ The recently approved BCBS (Basel IV) reforms of the standardized approach and of the CR-IRB approach for the calculation of risk-weighted assets for credit risk will limit the extent to which banks can reduce capital requirements through the use of internal models.

relationships to develop credit-scoring models to evaluate the ability to repay, the willingness to repay, and identify fraud. The Edward Altman Z-score model for bankruptcy prediction and the FICO score for retail credit scoring are some of the oldest industry standards, which loan providers still use because of their high interpretability (Baesens et al. 2016).

These methods are less effective in emerging economies and among low-income unbanked segments of the population, who often do not have access to formal financing and/or do not earn regular labor income. To cope with these constraints and to improve credit risk assessment, banks and loan providers are increasingly using non-traditional data sets (e.g., mobile operators, utilities, retailers, and direct-sales companies data) to sophisticate their credit bureaus and credit rating services. This factor poses new challenges to credit scoring modelers since non-traditional data must typically be collected from different sources, and its volume is several times that of traditional sources. By pursuing this approach, lenders seek to have more accurate information and incentive to grow the credit market under a robust credit control framework. By increasing their use of these new data sources, they try to provide more lending to their public customers and get to analyze loan requests better, ultimately increasing the loan ratio and decreasing the decision time. People will then have more monthly disposable money for spending, which will contribute to the economy, but it can also create risks for financial institutions. Therefore, as shown in Figure 3-1, non-traditional data sets provide the credit market a chance to manage different data sources to boost credit analysis outcomes and follow the stipulated recommendations of standards appropriately.

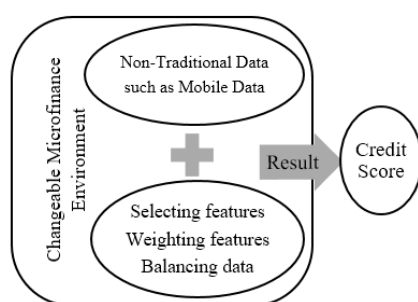


FIGURE 3-1- CHALLENGES IN CREDIT SCORING.

While this approach presents some opportunities, it also carries some challenges, which are represented in Figure 3-1. First, most of the mobile phone features are redundant and do not contribute to representing credit risk. Second, banks and loan providers should follow the regulators, and the critical change factor in banking is regulation. It means that fluctuations in the economy and regulations could change the behavior of both banks and customers. Non-traditional data sets can reveal these fluctuations, but current methods are computationally expensive with a high false-negative rate. The third risk factor is the smart behavior of fraudsters, and it means conscious changes may exist in non-traditional data, which influence the contribution level of features in credit scoring. However, the current methods are not wise enough to renew credit scores over time. As a result, we need highly effective and computationally less expensive solutions to calculate an informative credit score for satisfying the accuracy expectation of financial institutions. Although there are a large number of techniques employed in the development of credit-scoring, empirical studies show that the false-negative rate obtained is still not good enough for non-traditional data sets.

In this Chapter, we introduce a novel time-dependent credit scoring method to identify good loans with a low false-negative rate. The method uses a two-step approach based on an initial Kruskal-Wallis non-parametric statistic analysis to form a computationally efficient credit-scoring model based on an artificial neural network, Logistic regression with Ridge penalty, Random Forests (RF), and Support Vector Machine (SVM) to learn the model and to assess model performance. The Kruskal-Wallis statistic is used for selecting the most prominent features for assessing loan default in credit risk management over time. This statistic is sensitive to events that are far from the credit scores of good clients, computationally less expensive and very simple to implement. Additionally, we introduce a credit scoring index that uses the Kruskal-Wallis statistic as a weight of feature to decrease the false-negative rate which is able to purify features and decrease the dimensions in real-time. This new credit scoring index may be particularly interesting for loan providers assessing loan applications from individuals without any credit history and based exclusively on non-traditional data analysis. Illustrative empirical results on the use of this novel time-dependent credit scoring method are provided considering a sample credit dataset. The empirical findings show that the new credit scoring

methodology represents a reasonable coefficient of determination and a low false-negative rate. The accuracy is high and the model is computationally less expensive. This is the first study that develops a non-parametric credit scoring system that is able to reselect effective features for continued credit evaluation and weigh them out by their level of contribution with a good diagnostic ability.

The rest of this Chapter is structured as follows. In Section 3.2, we review credit-scoring models and new non-traditional data sources. Section 3.3 introduces the novel credit scoring method and discusses the calculation of the credit score. Then, using a credit risk data set, we compare the classification accuracy of credit scores with available features in section 3.4. Furthermore, an artificial neural network model is dedicated to the new method to show the accuracy of predicting the probability of default. In section 3.5, we discuss the main managerial and theoretical implications of this research. Finally, Section 3.6 contains some concluding remarks.

3.2 LITERATURE REVIEW

3.2.1 Credit scoring models using traditional data sets

Traditional credit-scoring models applying single-period classification techniques (e.g., logit, probit) to classify credit customers into different risk groups and to estimate the probability of default are still the most popular data mining techniques used in the industry (Chamboko & Bravo, 2019a, 2019b). Altman (1968) pioneered this area by developing the Z-score discriminant analysis model based on five financial ratios to predict corporate bankruptcy. Since then, several techniques have been developed to help decision-makers and analysts in predicting financial failure by considering both traditional statistical methods and more sophisticated (e.g., advanced machine learning) modeling approaches and alternative sets of predictor features. Standard models using external ratings provided by external credit assessment institutions have also been successfully applied. The set of classification algorithms used in credit scoring includes individual classifiers and homogenous and heterogeneous ensembles.

Individual classifiers employing single statistical or operational research methods include linear and multiple discriminate analysis (DA), logistic regression (LR), probit analysis, linear and quadratic programming, and data envelopment analysis (see, e.g., Altman et al. 1977; Zmijewski 1984; Jones and Hensher 2004; Premachandra et al., 2009; Kwak et al., 2012). Classifiers using machine learning methods such as neural network (NN), support vector machine (SVM), decision trees (DT), and genetic and evolutionary algorithms (GA) have also been investigated (see, e.g., Hand and Henley 1997; Arminger et al. 1977; B. Baesens et al. 2003; Shin et al. 2005; Lensberg et al. 2006; Erdogan 2013; Kruppa et al. 2013; Acosta-González and Fernández-Rodríguez 2014; Lessmann et al. 2015; Butaru et al. 2016; Abellán and Castellano 2017; Zhao et al. 2017). Homogenous ensembles typically employ one of the above individual classification methods with various samples or parameters to build base classifiers, which are subsequently combined using a majority voting rule or other frequentist or Bayesian integration methods (Feng et al. 2018). Recently, heterogeneous ensemble, which combines the prediction of base models created by alternative classification algorithms, often in a dynamic (adaptive or selective) way, has attracted much attention because of its superior predictive performance over homogeneous ensemble (see, e.g., Xia et al. 2018). Some approaches model default as a dynamic (sequential) process (see, e.g., Volkov et al. 2017). Recent proposals in the field of credit scoring focuses on three dimensions: novel classification algorithms using dynamic ensembles, deep learning methods, dissimilarity space, associative memories, and probabilistic rough sets, novel performance measures, and the minimization of the decision-relevant costs, and statistical hypothesis tests. Xia et al. (2018) propose a novel heterogeneous ensemble credit model (named bstacking) that integrates the bagging algorithm with the stacking method. Feng et al. (2018) develop a new dynamic ensemble classification method for credit scoring based on soft probability in which classifiers are first selected based on their classification ability and the relative costs of Type I error and Type II error in the validation set and then combined to get an interval probability of default by using soft probability. Luo et al. (2017) investigate and compare the classification performance of deep belief networks (DBN) with Restricted Boltzmann Machines with that of popular credit scoring models such as LR, multi-layer perceptron, and SVM using credit default swaps data. Cleofas-Sánchez et al.

(2016) propose an alternative technique for financial distress prediction based on a specific type of neural network called hybrid associative classifier with translation (HACT). The HACT neural network is an associative memory that merges the encoding phase of the linear associator with the decoding phase of the Steinbuch's lern matrix to improve the performance of the classifier. Pławiak et al. (2019) develop a new approach for credit scoring based on a deep genetic cascade ensemble of different SVM classifiers called Deep Genetic Cascade Ensembles of Classifiers (DGCEC) combining evolutionary computation, and ensemble and deep learning methods. García et al. (2019) address the problem of corporate bankruptcy prediction considering four linear classifiers (Fisher's linear discriminant, linear discriminant classifier, SVM, and LR) adopting a dissimilarity representation in which samples to be classified/predicted are derived from pairwise dissimilarities instead of being represented as usual by a set of features (explanatory variables), which defines a feature space. Maldonado et al. (2020) propose a methodology for credit scoring that minimizes the decision-relevant costs by classifying borrowers into three instead of two classes using the theory of three-way decisions with probabilistic rough sets. García et al. (2010) perform an experimental analysis to compare scorecard performance.

Despite their popularity, credit scoring models can only provide an estimate of the lifetime probability of default for a loan but cannot identify the existence of cures and/or other competing transitions and their relationship to loan-level and macro covariates, and do not provide insight on the timing of default, the cure from the default, the time since default, and time to collateral repossession (Chamboko & Bravo, 2020; Lessmann et al., 2015). Survival models incorporating time-varying covariates such as macroeconomic conditions which affect performance on loan payment over time and the ability to forecast event occurrence (default, recovery, prepayment, foreclosure) in the next instant of time, given that the event has not occurred until that time, have proven to overperform traditional methods in empirical studies (see, e.g., (Bellotti & Crook, 2013; Castro, 2013; Chamboko & Bravo, 2016, 2019a, 2019b; Noh, Roh, & Han, 2005; Sarlija et al. 2009; Tong et al. 2012). A handful of studies have also used the same to model foreclosure on mortgages (Gerardi, Shapiro, & Willen, 2011) and cure from delinquency to current (Ha, 2010; Ho Ha & Krishnan, 2012). The competing risks survival framework has also been used

to model the competing risks of early payment and default on loan contracts (Deng et al. 2000; Stepanova & Thomas, 2002).

3.2.2 Non-traditional data sets for credit scoring

Credit scoring models using non-traditional data sets are a cost-effective method of surveying personalities for risk management purposes of monetary institutions. It shifts credit scoring to high-tech to avoid the personal subjectivity of analysts or underwriters (Fensterstock, 2005). It also helps in increasing the speed and consistency of the application processes and allows financial firms to automate their processes (Rimmer, 2005).

Credit scoring and new technologies help loan providers shorten the processing time of loan applications and improve the allocation of resources (Jacobson & Roszbach, 2003). Additionally, it can aid insurance firms in making better predictions on claims and determining the interest rate which the firms should charge their consumers as well as the pricing of portfolios and products (Avery et al., 2000; Kellison & Wortham, 2003). Mobile phone data is a new Big Data source for smarter credit-scoring models, independent of the usual financial institutes databases. Mobile phones provide non-traditional data sources in the form of call detail records (CDR) and many other log files which can contribute to improving retail risk models not only for bank customers but also for the largely unbanked population who have no regular credit history. Figure 3-2 represents how non-traditional data is emerging in credit scoring by using parallel computing, distributed computing, and Big Data solutions. They represent how digitization in banking has gradually allowed financial institutions to use both Big Data and traditional credit records for managing risks. It shows how technology development is helping loan providers create value from several huge volumes of non-traditional data with increasing computation efficiency. Providing faster and consistent decisions for sub-prime customers with poor credit records, credit impairment, missing data in their credit histories, or difficulty in validating their income is another advantage of using non-traditional data in credit scoring modeling (Quittner, 2003).

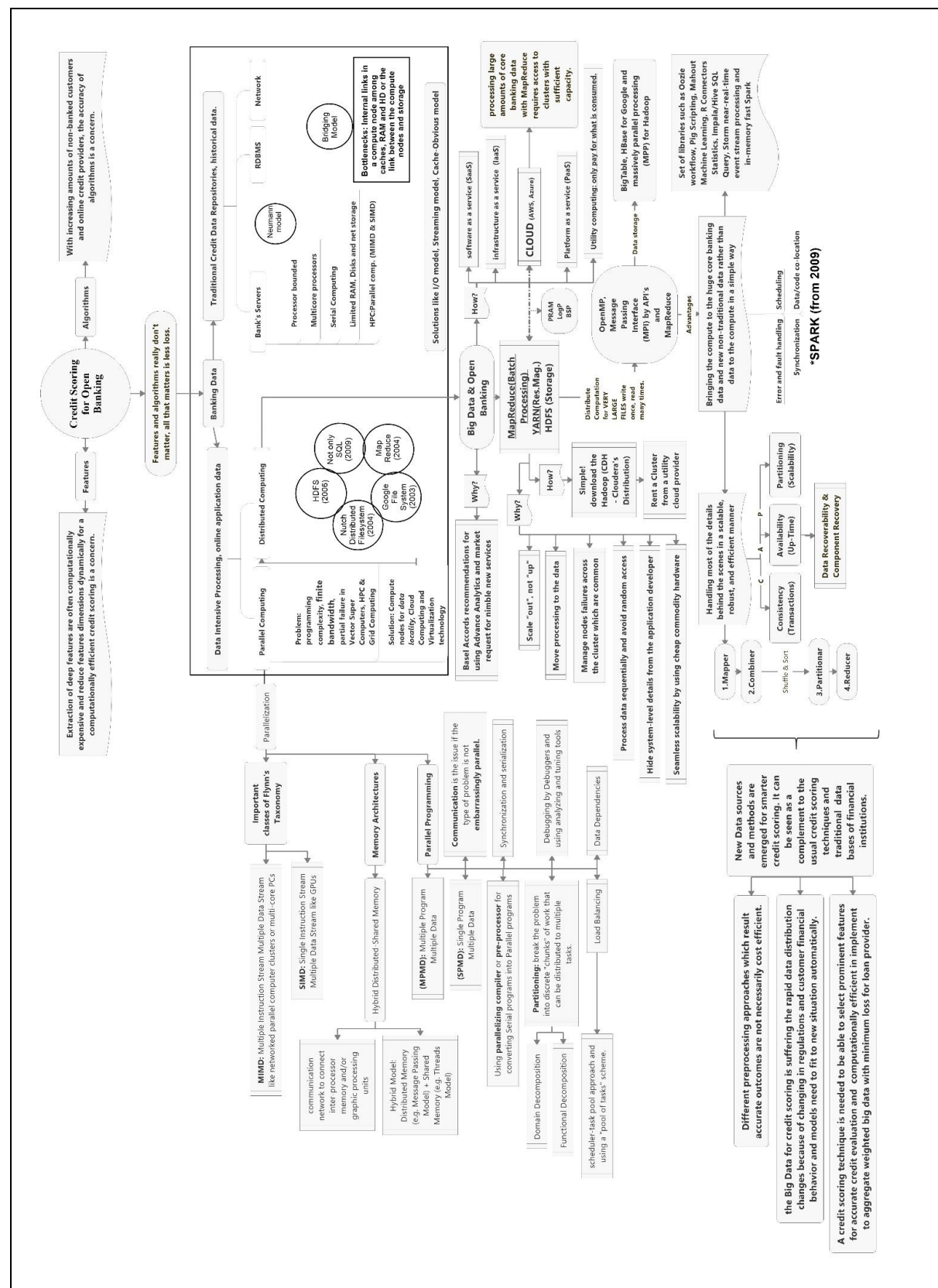


FIGURE 3-2 - EMERGENCE OF NON-TRADITIONAL DATA ANALYSIS IN CREDIT SCORING.
(HIGH QUALITY IMAGE: [HTTPS://WWW.LINKEDIN.COM/PULSE/BIG-DATA-A-ASHOFTE/](https://www.linkedin.com/pulse/big-data-a-ashofte/))

Despite its considerable benefits, lending to these groups is characterized as an inherent high risk due to a lack of collateral and information asymmetry. Some researchers use behavioral signatures in mobile phone data to predict default with an accuracy almost similar to that of credit-scoring methods that use financial history by Random Forest and Logistic regression (Bjorkegren & Grissen, 2018). Pedro et al. (2015) developed MobiScore, a methodology that models the user's financial risk using data collected from mobile usage using gradient boosting, support vector machine, and linear regression models. Other studies have used boosted decision trees and Logistic regression to create a credit score for underbanked populations considering information about people's usage of various mobile apps to make conclusions about their mood and personality traits (Chittaranjan et al. 2013; Do & Gatica-Perez, 2010; Skyler et al., 2017; Verkasalo et al., 2010). As a result, mobile phone data and social network analytics were used in credit scoring applications showing that incorporating telco data has the potential to increase the value of credit scoring (Óskarsdóttir et al, 2019).

Current credit scoring methods are computationally expensive and face critical challenges such as drift and class imbalance, reject inference, outliers, data set shift, irrelevant features, and missing and noisy data. As we discussed in Figure 3-1, the class imbalance problem and changes in the macro-finance environment and markets could potentially change the relationship between client characteristics and credit assessment results over time, causing concept drift in client credit assessments. Zhang and Liu (2019) proposed a novel multiple time scale ensemble classifier and a novel sample-based online class imbalance learning procedure to handle these two problems in the client credit assessment data streams. Because of the minority of delinquent customers, class distributions are highly imbalanced and represented skewed distributions. Although the topic of imbalanced classification has gathered the full attention of researchers during the last several years, such as the cost-sensitive learning technique by Douzas and Bação (2018), the emergence of mobile phone data brings new problems and challenges for the class imbalance problem in credit scoring, especially for unbanked individuals. The misclassification costs of false-negative cases are typically much higher than those associated with the non-default or non-bankrupt (negative) class. García et al. (2019b) investigate the potential links between the performance of several

classifier ensembles and the positive (default or bankrupt) sample types using 14 different real-life financial databases.

Credit scoring models are typically built on historical performances, which means that only accepted requests are used in estimating the probability of default, which may cause sample bias and reduce predictive performances. Recent proposals in the field of credit scoring adopt reject inference methods, i.e., use the information encompassed in the rejected samples in combination with accepted samples. Li et al. (2017) propose a new method in reject inference using the machine learning technique of Semi-supervised Support Vector Machines (SSVM) to classify the status of rejected borrowers and empirically investigate the performance of the new method. Xia (2019) propose a novel reject inference model (named OD-LightGBM) that combines a recent outlier detection algorithm (i.e., isolation forest) and state-of-the-art gradient boosting decision tree (GBDT) classifier (LightGBM). Nyitrai and Virág (2019) handle the problem of outliers in credit scoring and examine the impact of outliers winsorized at different levels. Gicić and Subasi (2019) address the problem of class imbalance in microcredit data and propose a new microcredit scoring model based on synthetic minority oversampling technique (SMOTE) for data preprocessing and ensemble classifiers. Óskarsdóttir et al. (2019) used the undersampling method to reduce class imbalance for training sets of data, which shows the intention to reduce the size of the majority class when applying these analytics techniques. Tian et al. (2018) propose a new method using the state-of-the-art kernel-free Fuzzy Quadratic Surface Support Vector Machine (FQSSVM) to infer the statuses of the rejected applicants and solve the outlier problem in credit assessment. Liu and Pan (2018) propose a new hybrid classifier based on fuzzy-rough instance selection to minimize the negative influence on the classification accuracy of using the wrong number of clusters or poor starting points of each cluster. García et al. (2012) empirically investigate whether the application of filtering algorithms leads to an increase in the accuracy of instance-based classifiers in the context of credit risk assessment. The authors consider 20 different algorithms and 8 credit databases and conclude that the filtered sets perform significantly better than the non-preprocessed training sets when using the nearest neighbor decision rule and that some techniques are most robust and accurate when confronted with noisy credit data. Recently, sound statistical and machine learning procedures that are computationally scalable to massive non-traditional datasets have been

proposed (Jordan, 2013). Examples are subsampling-based approaches (Kleiner et al., 2014; Kruppa et al., 2013; Liang et al., 2013; Ma et al., 2015; Maclaurin & Adams, 2015), divide and conquer approaches (Song & Liang, 2015), and online updating approaches (Schifano et al., 2016).

3.3 DYNAMIC NONPARAMETRIC CREDIT SCORE

The performance of a classification system can be improved by picking up the optimized features of mobile phones and decreasing the complexity of the model in the preprocessing stage. Many methods have been developed for choosing significant features with high information, such as the Kruskal-Wallis method (Saeys et al., 2007). The Kruskal-Wallis test as a nonparametric approach is useful to select informative features for loan default in credit risk management. Because it is sensitive to events that are far from the credit scores of good clients, we use the Kruskal-Wallis non-parametric statistic in our proposed method, which is computationally less expensive, and very simple to implement. Additionally, we introduce a new credit scoring approach that is able to purify features and decrease the dimensions in real-time. We will describe it in the following subsection 3.1. Next, we use the Kruskal-Wallis statistic measures as a weights of features to decrease the false-negative rate and improve the model's accuracy, which will be discussed in subsection 3.2, and finally in subsection 3.3, we combine these two steps to introduce a new credit scoring formula.

3.3.1 Kruskal-Wallis statistic for online features reduction

Let us define the null hypothesis that a feature does not contain discriminative information to detect default possibility in loan requests; otherwise, it is an informative feature and will be selected for contributing in the credit scoring. An assumption for this test is that the samples from the credit scores of good clients and credit scores of new clients are independent random samples from continuous distributions. In addition, we consider the time as an index of weights in our credit score method to ensure that the distributions of the training dataset of existing and new customers have the same shape at the time of analysis in this proposed online learning environment.

The computational procedure of the test can be considered as follows. Let X_{ijt} denote an observation of feature j from the client i at time t . If we let N_t be the total number of credit scores, which is equal to the total number of customers at time t , then by X_{ij} 's at time t , we will have a matrix $X_{N_t \times K_t}^t$ with N_t rows as the number of clients and K_t as the number of features at time t . Loans in banking credit risk literature usually divide into three categories: “Good” for good loans, “Medium” for substandard loans/doubtful loans, and “Bad” for loss loans. The Kruskal-Wallis test is appropriate for these kinds of categorical variables with three or more groups. However, most of the available datasets for credit purposes merge the first and second groups as bad loans to create a label attribute with two values, for instance, zero for the good loans and one for the bad loans.

We denote by S the number of groups in the Kruskal Wallis statistic, n_{ijt} the number of observations in group i at time t , and vector $Y_{N_t \times 1}$ the label vector considering, for instance, three possible labels for each customer (“Good,” “Substandard or Doubtful,” and “Loss”).

Computationally, if R_{ijt} is the rank assigned to the j^{th} feature of i^{th} client at time t , then the Kruskal-Wallis statistic for j^{th} feature at time t for N_t customers is

$$H_j^t = \frac{12}{N_t(N_t+1)} \left(\sum_{i=1}^S \frac{R_{ijt}^2}{n_{ijt}} \right) - 3(N_t + 1) \quad j = 1, 2, \dots, K_t. \quad (3-1)$$

Then

$$H_j^t \xrightarrow{d} \chi_1^2 \quad \text{in distribution,}$$

where χ_1^2 is the χ^2 distribution with one degree of freedom. The null hypothesis will be rejected if the computed value of H_j^t for each j from 1 to K_t exceeds the value of chi-square for reselected confidence level and 1 degree of freedom. Simply, in the credit scoring scenario, we define γ_j^t to find a balanced measure with less complexity to be followed with zero as a baseline and negative-positive values for making decisions about features. We make γ_j^t based on following τ_j^t .

Let

$$\tau_j^t = 1 - \frac{H_j^t}{H_j^t + \chi_1^2} = \frac{\chi_1^2}{H_j^t + \chi_1^2}, \quad (3-2)$$

$$0 \leq \tau_j^t \leq 1; \text{ for all } j, j = 1, 2, \dots, k.$$

Hence, we obtain that

$$-0.5 \leq \tau_j^t - 0.5 \leq 0.5.$$

Then we define a nonparametric measure γ for feature j at time t as $\gamma_j^t = \tau_j^t - 0.5$, therefore

$$\gamma_j^t = \frac{\chi_1^2 - H_j^t}{2 \times (\chi_1^2 + H_j^t)} \quad (3-3)$$

and

$$-0.5 \leq \gamma_j^t \leq 0.5. \quad (3-4)$$

If γ_j^t is positive, then the feature is not able to differentiate the classes, and if it is negative, then the feature could be used for the modeling phase to determine the creditworthiness of an applicant. It is clear that this change is only superficial to make it easier to understand, programing and represent it as a control chart in a dashboard. Now, we study the behavior of γ_j^t by looking at its distribution function.

Let γ_j^t be as defined in formula (3-3) and denote by $F_{\gamma_j^t}(y)$ be the corresponding cumulative distribution function (cdf)

$$\begin{aligned} F_{\gamma_j^t}(y) &= P(\gamma_j^t \leq y) = 1 - P\left(\frac{\chi_1^2 - H_j^t}{2 \times (\chi_1^2 + H_j^t)} > y\right) = 1 - P\left(H_j^t < \frac{(1-2y)\chi_1^2}{(1+2y)}\right) F_{\gamma_j^t}(y) = \\ P(\gamma_j^t \leq y) &= 1 - P\left(\frac{\chi_1^2 - H_j^t}{2 \times (\chi_1^2 + H_j^t)} > y\right) = 1 - P\left(H_j^t < \frac{(1-2y)\chi_1^2}{(1+2y)}\right), \end{aligned} \quad (3-5)$$

or, equivalently,

$$F_{\gamma_j^t}(y) = 1 - F_H \left[\left(\frac{0.5-y}{0.5+y} \right) \times \chi_1^2 \right] F_{\gamma_j^t}(y) = 1 - F_H \left[\left(\frac{0.5-y}{0.5+y} \right) \times \chi_1^2 \right], \quad (3-6)$$

where F_H is the cumulative distribution function of the Kruskal-Wallis statistic and $\left(\frac{0.5-y}{0.5+y} \right) \times \chi_1^2 > 0$ for any values of $y \in (-0.5, 0.5)$.

Furthermore, because $H \rightarrow \chi_1^2$ in distribution, the density function of γ_j^t will be

$$f_{\gamma_j^t}(y) = \frac{dF_{\gamma_j^t}(y)}{dy} = - \frac{d \left(\left(\frac{0.5-y}{0.5+y} \right) \times \chi_1^2 \right)}{dy} \times f_H \left(\left(\frac{0.5-y}{0.5+y} \right) \times \chi_1^2 \right) f_{\gamma_j^t}(y) = \frac{dF_{\gamma_j^t}(y)}{dy} = - \frac{d \left(\left(\frac{0.5-y}{0.5+y} \right) \times \chi_1^2 \right)}{dy} \times f_H \left(\left(\frac{0.5-y}{0.5+y} \right) \times \chi_1^2 \right); \quad (3-7)$$

or, equivalently,

$$f_{\gamma_j^t}(y) = \sqrt{\frac{\chi_1^2}{2\pi(0.25-y^2)(0.5+y)^2}} \times \text{Exp} \left\{ - \left[\left(\frac{0.5-y}{1+2y} \right) \times \chi_1^2 \right] \right\}, \quad -0.5 < y < 0.5 \quad (3-8)$$

With 95% confidence level $\chi_{1,0.05}^2 = 3.841$ and the γ_j^t density function is

$$f_{\gamma_j^t}(y) = 0.7818 \times \sqrt{\frac{1}{(0.25-y^2)(0.5+y)^2}} \times \text{Exp} \left[\frac{3.841 \times (y-0.5)}{1+2y} \right], \quad -0.5 < y < 0.5 \quad (3-9)$$

The density function of γ_j^t is shown below (Figure 3-3.).

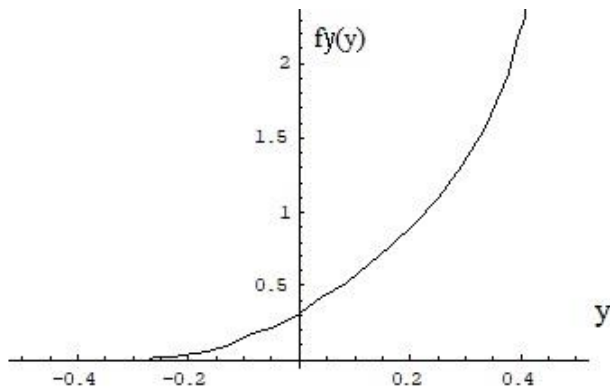


FIGURE 3-3 - THE DENSITY PLOT OF γ_j^t .

From Figure 3-3, it is clear that the majority of probability density based on the area under the density function curve for statistically significant features is between -0.3 and 0. We can expect that most of the magnificent fluctuations in

the effectiveness of features on our credit score model will happen in the small probability area under the density function between -0.3 and 0. Therefore, if we consider only the sign of γ_j^t in a flag attribute, instead of its value to select the informative features based on positive signs, we can manage memory better and eliminate unnecessary features in our computer program with only a Boolean type attribute, which only needs one byte of memory, the smallest unit addressable with the CPU in compare with, for instance, decimal which needs 12 bytes. It means less memory will be occupied and data transfer in the network will be optimized, especially for massive non-traditional datasets with numerous features.

3.3.2 Kruskal-Wallis statistic for empowering features

In this step, we use γ_j^t in our credit score model to boost the effective features. It was shown that γ_j^t signs can be a part of credit scoring as an indicator for feature selection. Now, we define a transformed γ_j^t to be used as a transformation weight for the k^{th} feature to improve the performance of classification and credit scoring.

Let us define w_k^t as the transformation weight of feature k at time t :

$$w_j^t = \begin{cases} 2 \times |\gamma_j^t| & \text{for } \gamma_j^t < 0 \\ 0 & \text{for } \gamma_j^t > 0 \end{cases} \quad j = 1, 2, \dots, N. \quad (3-10)$$

Then, φ as the impact factor of feature k at time t will be

$$\varphi_j^t = \begin{cases} \frac{w_j^t}{\sum_{j=1}^k w_j^t} & \text{for } w_j^t > 0 \\ 0 & \text{for } w_j^t = 0 \end{cases}, \quad \sum_{j=1}^N \varphi_j^t = 1. \quad (3-11)$$

As we discussed, for a 95% confidence level, if $\{H_j^t | H_j^t > \chi_{1,0.05}^2 = 3.84\}$ then we reject the null hypothesis. It means that the feature contains discriminative information to detect default possibility in loan requests. Furthermore, as shown in Figure 3-3, important fluctuations in the effectiveness of features in

credit scoring happens when γ_j^t is between -0.3 and 0. Now, the equivalent values of γ_j^t and w_j^t to values of H_j^t is shown below in Table 3.1.

TABLE 3.1 - . H_j^t VALUES AND ITS EQUIVALENT IN γ_j^t , w_j^t AND ϕ_j^t .

| H_j^t | 0 | 1 | 2 | 3 | 3.5 | 4 | 5 |
|--------------|-----|------|------|------|------|-------|-------|
| γ_j^t | 0.5 | 0.29 | 0.15 | 0.06 | 0.02 | -0.01 | -0.06 |
| w_j^t | 0 | 0 | 0 | 0 | 0 | 0.020 | 0.131 |
| ϕ_j^t | 0 | 0 | 0 | 0 | 0 | 0.004 | 0.026 |

| H_j^t | 10 | 20 | 40 | 100 | 400 | 700 |
|--------------|-------|-------|-------|-------|-------|--------|
| γ_j^t | -0.22 | -0.33 | -0.41 | -0.46 | -0.49 | -0.495 |
| w_j^t | 0.445 | 0.678 | 0.825 | 0.926 | 0.981 | 0.989 |
| ϕ_j^t | 0.089 | 0.136 | 0.165 | 0.185 | 0.196 | 0.198 |

Source: Author's preparation.

Table 3.1 illustrates that w_j^t and ϕ_j^t react impressively to the changes of γ_j^t in this interval. Therefore, w_j^t close to one shows the high capability of the feature to recognize the credit level of the customer. A trend towards zero means the probability of the feature's influence is decreasing. Monitoring the trend of w_j^t for different t could show the level of the model's reliability during the time. If it decreases to zero randomly and sequentially then it can be concluded that the reliability of the trained model based on those features is decreasing, and it is a warning sign to renew the model by using a new set of features. It helps to identify for how long our scoring model keeps up with the initial performance and to discover the right time of redoing the training step with a new set of features. It would also be useful to monitor w_j^t by considering zero for excluding the feature from the credit score model. In addition to the above mentioned advantages, we use w_j^t at time t as the power of attribute j to improve the accuracy of the model and to decrease the false-negative rate.

For this purpose, we consider all attributes for defaulted customers within the training dataset to the power of corresponding w_j^t in the training stage. The useful attributes with the higher H_j^t will have w_j^t closer to one. It means these optimal features will experience less change than features with a lower ability to determine the loan default. As a result, in the training stage, we can make distinguish between attributes of good and bad loans based on their power of

contribution in the model. In fact, we are saving the high-performance attributes, and we use the less important attributes as labels for bad loans. By using the features with high w_j^t for making the model and the features with low w_j^t to differentiate between good and bad loans the model's performance is expected to improve.

3.3.3 Credit score formulation

Now, we introduce our new credit risk index (CRI) for attributes with interval or ratio scale of the i^{th} client at time t by using the geometric mean as following:

$$CRI_{it} = \prod_{j=1}^k \left[\frac{(x_{ij}^t)^{w_j^t}}{(\bar{x}_j^t)} \times 100 \right]^{\varphi_j^t}, \quad i = 1, 2, \dots, N, \quad (3-12)$$

where \bar{x}_j^t is the standard profile of attribute j extracted from the data of good clients. If we investigate these credit score characteristics by using the most desirable axioms of the axiomatic approach to the index number theory, it satisfies most of them⁵. Thus, this credit index formula as a homogeneous symmetric average can be calculated as an accurate aggregate measure, and it is able to renew features dynamically and weighted out by φ_j^t as their impact factors. The pseudo-code of the proposed methodology is listed in Table 3.2.

In the training stage, obviously, we will have a very large area under the AUC curve because the values of bad loans in the training set have experienced the shift based on w_j^t . However, in the test stage, we are using an independent out-of-sample dataset, and we actually have no idea which loan is the default. As shown in Table 3.2, we first apply blindly the transformation for all attributes of all clients in the test dataset, including good and bad loans. This will shift the attributes of good loans to bad loans and potentially could decrease the true positive rate, but it will also decrease the false-negative rate dramatically.

⁵ Axiomatic approach to the index number theory such as the Positivity test, Continuity Test, Identity Test, Homogeneity Test for Period t , Homogeneity Test for Period zero, Commodity Reversal Test, Invariance to Changes in the Units of Measurement or the Commensurability Test, Time Reversal Test, Circularity or Transitivity Test, Mean Value Test, Monotonicity Test with Respect to Period t and Monotonicity Test with Respect to Period zero.

TABLE 3.2 - PSEUDO CODE OF PROPOSED METHODOLOGY.

INPUT attributes and Status variable;
 OUTPUT weighted attributes, CRI;

1. StatExplore attributes;
2. Change[outliers] = False; {outliers are important in credit scoring.}
3. Set $\chi^2_{1,0.05} = 3.84$;
4. For each attribute do
 5. The Kruskal-Wallis.test;
 6. If (Scale[attribute] is not Nominal) then
 7. If (KW.statistic > $\chi^2_{1,0.05}$) then
 8. If (data=train.set and Default=True) then
 9. Set Gama = $(\chi^2_{1,0.05} - \text{KW.statistic}) / (2 * (\chi^2_{1,0.05} + \text{KW.statistic}))$
 10. Set W = 2 * ABS(Gama);
 11. SET attribute.value = POWER [attribute.value , W]
 12. IF (MISSING(attribute)=TRUE) THEN attribute=attribute
 13. IF (MISSING(attribute)=TRUE) THEN attribute=AVERAGE(attribute)
 14. IF (attribute = 0) THEN attribute=AVERAGE(attribute)
 15. ElseIf (data=test.set for ALL) then
 16. Set Gama = $(\chi^2_{1,0.05} - \text{KW.statistic}) / (2 * (\chi^2_{1,0.05} + \text{KW.statistic}))$
 17. Set W = 2 * ABS(Gama);
 18. SET attribute.value = POWER [attribute.value , W]
 19. Else
 20. Set attribute = excluded; {equivalent to Set W = 0; Set Phi = 0;}
 21. Else
 22. Set attribute = unchanged; {equivalent to Set W = 1; Set Phi = 1}
 23. End do
 24. # computing CRI
 25. For each attribute do
 26. If (Default=False) then Mean_default_NO= Average(attribute)
 27. Set Phi = W / SUM(W's);
 28. Set attribute_CRI = POWER [(attribute.value.W / Mean_default_NO× 100) , Phi]
 29. End do
 30. For each client DO
 31. CRI = Multiply(ALL attribute_CRI's)
 32. END DO

Source: Author's preparation.

It could be interesting for loan providers, especially when they want to offer a loan to clients without any credit history and only based on Big Data analysis.

In this case, it is beneficial if we can detect the separated sections of good and bad customers and struggle to detect good customers from the muddy intersection of good and bad loans in the dataset, where there is high similarity in attributes of different categories. Additionally, using the CRI as an aggregate of features with an interval/ratio scale will significantly decrease the required computation for modeling.

3.4 EXPERIMENTAL DESIGN

In this section we illustrate the use of our online credit scoring method, using two public loan datasets. One, “German Credit Data” which is a small dataset with 6,377 observations and another “Lending club loan data” dataset with more than two million observations (2,260,668). It would be suitable to compare the performance of this credit scoring methodology in different situations.

Without loss of generality, we assume for simplicity that t equals one and compute γ_j^1 and w_j^1 . In this Chapter, we use receiver operating curves (ROC) to show the statistical performance of the models. In the ROC chart, the horizontal axis represents the specificity, and the vertical axis shows the sensitivity. The greater the area between the curve and the baseline, the better the feature performance in default prediction. After investigating the characteristics of the new credit score model, we employed the area ratio of ROC curves to compare the classification accuracy and evaluate how well this credit scoring model performs. The data sets will randomly be divided into two groups, 65% for model training and the other 35% to apply different algorithms to the novel credit scoring methodology.

3.4.1 Small data set

3.4.1.1 Data description

We obtained the data from ‘German Credit Data’ and removed unnecessary features from it. We consider the following 13 explanatory variables. “V1: Seniority” for Job seniority (year), “V2: Home” for type of homeownership, “V3: Time” for time of requested loan, “V4: Age” for client age, “V5: Marital” for marital status, “V6: Records” for existence of records, “V7: Job” for type of

job, “V8: Expenses” for amount of expenses, “V9: Income” for amount of income, “V10: Assets” for amount of assets, “V11: Debt” for amount of debt, “V12: Amount” for amount requested of loan and “V13: Price” for price of goods. Among the total 6,377 observations, 2,217 (34.8%) are loan requests with default payment according to the Basel accords definition, which have three or more late payments that imply default. The response variable is a binary variable named “Status,” which represents loan default (No=1, Yes=2). The dataset is anonymized, and does not contain personal information. Descriptive statistics of attributes are presented as follows.

Descriptive statistics of categorical variable V2 are shown in the Table 3.3.

TABLE 3.3 - - DESCRIPTIVE STATISTICS OF “TYPE OF HOME-OWNERSHIP.”

| LOAN DEFAULT | V2: TYPE OF HOMEOWNERSHIP | | | | | | |
|-----------------|---------------------------|-------|---------|--------|---------|-------|-------|
| | RENT | OWNER | PRIVATE | IGNORE | PARENTS | OTHER | MODE |
| NO | 580 | 1697 | 159 | 11 | 544 | 1169 | Owner |
| YES | 1383 | 368 | 82 | 9 | 230 | 145 | Rent |

Source: Author’s preparation.

Descriptive statistics of categorical variables V6 and V7 are shown in Table 3.4.

TABLE 3.4 - DESCRIPTIVE STATISTICS OF “EXISTENCE OF RECORDS” & “TYPE OF JOB.”

| LOAN DEFAULT | V6: EXISTENCE OF RECORDS | | | V7: TYPE OF JOB | | | | |
|-----------------|--------------------------|-------------|--------|-----------------|---------------|----------------|--------|--------|
| | NO- REC | YES- REC | MODE | FIXED | PART- TIME | FREE LANCER | OTHERS | MODE |
| NO | 2820 | 1340 | No-rec | 3206 | 180 | 673 | 101 | Fixed |
| YES | 1802 | 415 | No-rec | 577 | 269 | 307 | 1064 | Others |

Source: Author’s preparation.

and categorical variables V5 are described in Table 3.5.

TABLE 3.5 - DESCRIPTIVE STATISTICS OF “MARITAL STATUS.”

| LOAN DEFAULT | V5: MARITAL STATUS | | | | | |
|-----------------|--------------------|---------|-------|----------|----------|---------|
| | SINGLE | MARRIED | WIDOW | SEPARATE | DIVORCED | MODE |
| NO | 639 | 3383 | 47 | 67 | 24 | MARRIED |
| YES | 1318 | 806 | 19 | 60 | 14 | SINGLE |

Source: Author’s preparation.

In addition, descriptive statistics of other variables are depicted in Table 3.6.

TABLE 3.6 - DESCRIPTIVE STATISTICS OF SCALE VARIABLES.

| JOB SENIORITY (YEAR) | LOAN DEFAULT | N | MEAN | MIN. | MAX. | STD. EV. | KURTOSIS | SKEWNESS |
|----------------------------|-----------------|------|-------|------|------|----------|----------|----------|
| | NO | 4160 | 16.06 | 0 | 48 | 14.26 | -1.01 | 0.643 |
| | YES | 2217 | 3.82 | 0 | 43 | 4.62 | 13.57 | 3.15 |

| | | | | | | | | |
|---------------------------------|--------------|------|----------|-----|--------|-----------|--------|-------|
| | TOTAL | 6377 | 11.81 | 0 | 48 | 13.20 | 0.05 | 1.18 |
| TIME OF REQUESTED LOAN | NO | 4160 | 36.05 | 6 | 60 | 21.47 | -1.48 | -0.28 |
| | YES | 2217 | 59.39 | 6 | 72 | 14.81 | 0.29 | -1.06 |
| | TOTAL | 6377 | 44.16 | 6 | 72 | 22.37 | -1.01 | -0.53 |
| CLIENT AGE | NO | 4160 | 43.13 | 18 | 70 | 14.00 | -1.16 | 0.179 |
| | YES | 2217 | 29.06 | 18 | 65 | 10.37 | 0.59 | 1.18 |
| | TOTAL | 6377 | 38.23 | 18 | 70 | 14.50 | -0.96 | 0.47 |
| AMOUNT OF EXPENSES | NO | 4160 | 80.71 | 35 | 173 | 48.22 | -0.82 | 0.89 |
| | YES | 2217 | 50.16 | 35 | 165 | 17.66 | 3.70 | 1.80 |
| | TOTAL | 6377 | 70.09 | 35 | 173 | 42.86 | 0.49 | 1.38 |
| AMOUNT OF INCOME | NO | 4160 | 325.26 | 0 | 959 | 334.9 | -0.66 | 1.07 |
| | YES | 2217 | 99.38 | 0 | 959 | 71.22 | 11.49 | 1.69 |
| | TOTAL | 6377 | 246.73 | 0 | 959 | 294.11 | 1.09 | 1.66 |
| AMOUNT OF ASSETS | NO | 4160 | 62212.56 | 0 | 250000 | 100457.06 | -0.55 | 1.19 |
| | YES | 2217 | 6474.55 | 0 | 100000 | 7860.78 | 28.93 | 3.58 |
| | TOTAL | 6377 | 42834.27 | 0 | 250000 | 85491.53 | 1.49 | 1.85 |
| AMOUNT OF DEBT | NO | 4160 | 269.05 | 0 | 23500 | 1008.43 | 143.02 | 9.07 |
| | YES | 2217 | 9164.08 | 0 | 30000 | 10492.02 | -1.3 | 0.555 |
| | TOTAL | 6377 | 33.61.45 | 0 | 30000 | 7541.20 | 3.37 | 2.18 |
| AMOUNT REQUESTED OF LOAN | NO | 4160 | 1776.21 | 100 | 4500 | 1446.13 | -0.71 | 1.00 |
| | YES | 2217 | 951.72 | 105 | 4500 | 519.94 | 5.44 | 1.63 |
| | TOTAL | 6377 | 1489.57 | 100 | 4500 | 1261.64 | 0.81 | 1.52 |
| PRICE OF FINANCE GOODS | NO | 4160 | 3632.35 | 125 | 11140 | 3903.34 | -0.56 | 1.17 |
| | YES | 2217 | 1163.38 | 105 | 6802 | 608.47 | 11.86 | 2.36 |
| | TOTAL | 6377 | 2774.00 | 105 | 11140 | 3383.74 | 1.37 | 1.79 |

Source: Author's preparation.

3.4.1.2 Results of small data-set

The results are organized in two parts starting with the Kruskal-Wallis statistics, w_k^1 and ϕ_i^1 calculation to establish the artificial neural network model

based on propose methodology. Subsequently, the results are detailed, first in terms of model comparison and then in terms of computation performance.

The ϕ gives positive values to statistically significant features based on their detection power of default and gives others zero value, which means excluding those features from the model. The ϕ_i^1 based on w_k^1 is shown below (Table 3.7). In this dataset, all features are statistically significant.

TABLE 3.7 - K-W, w_k^1 AND ϕ_i^1

| | V1 | V2 | V3 | V4 | V5 | V6 | V7 |
|------------|--------|--------|--------|--------|--------|--------|--------|
| K-W | 1361 | 1191 | 1795 | 1539 | 1103 | 131 | 1923 |
| w_k^1 | 0.994 | 1 | 0.995 | 0.995 | 1 | 1 | 1 |
| ϕ_i^1 | 0.0772 | 0.0776 | 0.0773 | 0.0772 | 0.0776 | 0.0776 | 0.0776 |
| | V8 | V9 | V10 | V11 | V12 | V13 | |
| K-W | 556 | 983 | 136 | 669 | 419 | 1175 | |
| w_k^1 | 0.986 | 0.992 | 0.945 | 0.988 | 0.981 | 0.993 | |
| ϕ_i^1 | 0.0766 | 0.0770 | 0.0734 | 0.0767 | 0.0762 | 0.0771 | |

Source: Author's preparation.

From Table 3.8, The Kruskal-Wallis test for credit risk index (CRI) resulted in a p-value of less than 0.0001, which means that CRI is significantly able to differentiate the categories of good and default loans.

TABLE 3.8 - CREDIT SCORE VALIDATION.

| Credit Risk Index | N | Mean | Median | Mean Rank | K-W |
|-------------------|------|------|--------|-----------|------------|
| Good loans | 4160 | 1.64 | 1.28 | 3301.09 | 44.35 |
| Default loans | 2217 | 1.27 | 0.29 | 2978.67 | P-value=.0 |

Source: Author's preparation.

First, we use Logistic regression, which is a simple classifier and performs very well for credit scoring as a benchmark. It can produce a probabilistic estimation of the binary response variable, and it is a prevalent method for credit scoring. Thus, model A1 in Table 3.9 shows the results of Logistic regression for the original variables. Similarly, the artificial neural network for the original variables is represented by A2 in Table 3.9. We use these two models to estimate the improvement of predictive accuracy and the performance of the classification of proposed models, shown in the same Table by A3 and C.

Therefore, model A3 is built with the weighted explanatory variables (see Section 4.1.1), as well as model A4, with a combination of CRI and nominal variables. We propose CRI in model A4 as a candidate for representing

variables with a ratio scale, that is, the reduction in the features dimension and efficient computation resource management. Models A3 and A4 show the main effects of each transformation on variables in accuracy and computation efficiency of the proposed credit scoring algorithms and models A1 and A2 combine unchanged variables to categorize loans into groups of default and non-default to show how much the new models are able to improve the accuracy and performance of classification.

As is common practice, in credit scoring, statistical model performance is measured by the area under the receiver operating curve (AUC), and it is represented in Table 3.9 and Table 3.10.

TABLE 3.9 - CLASSIFICATION TABLE AND STATISTICAL MODEL PERFORMANCE (AUC) FOR ORIGINAL VARIABLES.

| Original Variables | | Predicted negative | Predicted positive | Percent correct | AUC | Time (Millisecond) |
|-------------------------|-------------------|--------------------|--------------------|-----------------|-------|--------------------|
| A1: Logistic Regression | Training negative | 2757 | 161 | 94.5% | 0.925 | - |
| | positive | 403 | 1120 | 73.5% | | |
| | Overall percent | 71.1% | 28.8% | 87.3% | | |
| | Holdout negative | 1159 | 83 | 93.3% | | |
| | positive | 194 | 500 | 70.1% | | |
| | Overall percent | 69.9% | 30.1% | 85.69% | | |
| A2: Neural network | Training negative | 2540 | 174 | 93.6% | 0.926 | 840 |
| | positive | 401 | 938 | 70.1% | | |
| | Overall percent | 72.6% | 27.4% | 85.8% | | |
| | Holdout negative | 1370 | 76 | 94.7% | | |
| | positive | 197 | 681 | 77.6% | | |
| | Overall percent | 67.4% | 32.6% | 88.3% | | |

Source: Author's preparation.

In Table 3.10, the model A3 includes all variables belonging to bad loans to the power of w_k^1 except for the nominal features. We consider all nominal variables and CRI in model A4 as features of the model. As is evident in Table 3.10, model A4 did not provide more significant results than model A3 except for less

computation. It could be important to deal with non-traditional datasets such as mobile data and Big Data sources of credit scoring.

TABLE 3.10 - CLASSIFICATION TABLE AND STATISTICAL MODEL PERFORMANCE (AUC) FOR WEIGHTED VARIABLES AND CRI.

| Artificial Neural Network | | Predicted negative | Predicted positive | Percent correct | AUC | Time (Millisecond) |
|-----------------------------------|-------------------|--------------------|--------------------|-----------------|-------|--------------------|
| A3: weighted variables | Training negative | 2715 | 0 | 100% | 0.999 | 320 |
| | positive | 0 | 1339 | 100% | | |
| | Overall percent | 67.0% | 33.0% | 100% | | |
| | Holdout negative | 514 | 932 | 35.5% | | |
| | positive | 2 | 876 | 99.8% | | |
| | Overall percent | 22.2% | 77.8% | 59.8% | | |
| A4: CRI and categorical variables | Training negative | 2714 | 0 | 100% | 0.999 | 80 |
| | positive | 0 | 1339 | 100% | | |
| | Overall percent | 67.0% | 33.0% | 100% | | |
| | Holdout negative | 504 | 942 | 34.9% | | |
| | positive | 1 | 877 | 99.9% | | |
| | Overall percent | 21.7% | 78.3% | 59.4% | | |

Source: Author's preparation.

In this credit scoring case, error rates are not the appropriate criteria to evaluate the performance of the credit score model because most clients are classified into creditable customers (93.6%). From Tables 3.9 and 3.10, it is clear that there is not a significant difference in the performance of the same models. However, by considering the area under the ROC curve as an essential factor of credit risk cost, models A1 and A2 perform the worst, of which models including the proposed methodology (models A3, A4) perform the best, not only in accuracy but also in computational efficiency. To offer a loan based on non-traditional data analysis, the benefit of correctly identifying a defaulter plays a prominent

role, and it is interesting to see that having only CRI and the pre-calculation of weighted features on the bad loans section of the dataset allows discriminating potentially better clients.

Furthermore, by using the area ratio in the test data, the classification result shows almost the same performance of models A3 and A4; however model A4 yields a better performance in computation time that is a critical factor in the performance of parallel and distributed computing for non-traditional datasets.

3.4.2 Big Data set

3.4.2.1 Data description

Lending club loan data contains complete loan data for all loans through 2007-2018, each loan includes applicant information provided by the applicant as well as the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. We found two versions of this dataset; one contains loans issued through the 2007-2015 and another version through 2012-2018. As a result, we combined these two datasets and removed the duplicates to obtain a complete dataset from 2007 to 2018 with maximum possible cases.

3.4.2.2 Application data

We consider the following application data as the following numeric attributes.

The "loan_amnt" for the listed amount of the loan applied for by the borrower with any possible reductions in the loan amount with the credit department by the time. The "emp_length" for employment length in years with possible values between 0 and 10, where 0 means less than one year and 10 means ten or more years. In the original dataset, the employment length is a combination of numbers, characters, plus signs that are converted to the numbers by the codes available in appendix 1. The "annual_inc", for the self-reported annual income provided by the borrower during registration, "dti" as a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income. The "delinq_2yrs" for the number of 30+ days past-

due incidences of delinquency in the borrower's credit file for the past 2 years, the "revol_util" for revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit, the "total_acc" for the total number of credit lines currently in the borrower's credit file, the "int_rate" for the interest rate on the loan, and finally, one categorical attribute "term_month", which stands for the number of payments on the loan with values in months, and can be either 36 or 60. We changed all the numeric attributes to float and years to double format and all rates in percentage in PySpark as in appendix 1.

3.4.2.3 *Behavioral data*

The selected behavioral data are all categorical, and we convert them into dummy variables in PySpark to be able to contribute to the modeling stage. These attributes are "home_ownership" which is the homeownership status provided by the borrower during registration or obtained from the credit report with values: RENT, OWN, MORTGAGE, and OTHER. "Purpose" as a category provided by the borrower for the loan request, "addr_state" for the state provided by the borrower in the loan application, "verification_status" for indicates if income was verified by LC, not verified, or if the income source was verified. We mapped multiple levels if verification_status attribute into the one-factor level as is shown in appendix PySpark codes. Finally, "application_type" which indicates whether the loan is an individual application or a joint application with two co-borrowers.

The label variable is "default_loan" with TRUE value (code 1) for default loans with values of "Default", "Charged Off", "Late (31-120 days)", "Late (16-30 days)", and FALSE (code 0) for non-default loans for "Fully Paid" loans.

3.4.2.4 *New measures*

We created two new measures to be considered in credit scoring models. For length of credit in years, "credit_length_in_years" is computed by subtracting the issue year from the earliest year. The issue year is extracted from the issue date, and the earliest year is also substring of the date that the borrower's earliest reported credit line was opened. Additionally, we want to know the fraction of the initial loan amount that has been reimbursed and to evaluate the

loan provider profit and loss according to model results. Therefore, we created a new column named "remain" by subtracting "loan payments" from the "total loan amount". This will represent the amount of money earned or lost per loan and the outstanding loan balance.

3.4.2.5 *Model and train and test datasets*

With the datasets featurized, credit scoring models are built using binary classifiers with a k-fold cross-validation on a dataset with 1,048,575 observations, which contains 217,930 default loans (True or 1), and 830,645 good loans (False or 0). The method of leave-one-out cross-validation is used to examine the between-sample variation of default prediction. The available data are divided into 10 disjoint subsets, and the models are trained on 9 of these subsets and the models selection criterion evaluated on the unused subset. This procedure is then repeated for all combinations of subsets by Python API of Apache Spark, as is presented in appendix 1. Leave-one-out cross-validation helps the algorithm to use all data as both training and validation, and consider the mean of the model selection criterion computed over the unused subset in each fold for better accuracy estimation.

Furthermore, Logistic regression works well for many business applications, which often have a simple decision boundary. Moreover, because of its simplicity, it is less prone to overfitting than flexible methods such as decision trees. Further, as we will show, variables that contribute to overfitting might be eliminated using Lasso or Ridge regularisation, without compromising out-of-sample accuracy. In this case, the Ridge method presented better performance than Lasso, and the following Logistic regressions are based on the Ridge penalty with elastic net regularization zero and regparam 0.3 as the best hyperparameters. In addition to the Logistic regression classifier as an industry standard for building credit scoring models, other binary classifiers such as random forests and linear support vector machines are used for the empirical analysis. Although they are more complex and powerful than Logistic regression in an application, the interpretability of these models could not be guaranteed as well as Logistic regression outputs.

3.4.2.6 Results of big data-set

The results of the Kruskal-Wallis statistics, w_k^1 and ϕ_i^1 is presented in Table 3.11, based on the propose methodology and the CRI is calculated and checked for performance as is shown in Figure 3-4. As we can see, all numerical attributes have the K-W statistic greater than $\chi_{1,0.05}^2$ and they are statistically significant. Therefore, we consider all of them for the next step, otherwise, the program specifies the power of zero for insignificant attributes, that is not the case here.

TABLE 3.11 - K-W, w_k^1 AND ϕ_i^1 .

| | Term months | Loan amnt | Emp length | Annual inc | dti | Delinq 2yrs |
|-----------------------------|----------------|--------------|---------------------------|---------------|----------|----------------|
| K-W | 39784 | 4870 | 134 | 4718 | 13055 | 342 |
| w _k ¹ | 1 | 0.998 | 0.944 | 0.998 | 0.999 | 0.978 |
| φ _i ¹ | 0.0926 | 0.0924 | 0.0874 | 0.0924 | 0.0925 | 0.0906 |
| | Revol util | Total acc | Credit length in years | | Int rate | Remain |
| K-W | 2760 | 63 | 1685 | | 69573 | 449682 |
| w _k ¹ | 0.997 | 0.887 | 0.995 | | 1 | 1 |
| φ _i ¹ | 0.0923 | 0.0822 | 0.0922 | | 0.0926 | 0.0926 |

Source: Author's preparation.

We start with the default flag rows of the training set and calculate the new numerical attribute k with powering to w_k^1 . For non-default loans, we consider the same value in the original attribute for the new one and impute the missing values with the average. For computing the CRI, we have to impute the zero values in each numeric attribute with average to avoid a null result when multiplying by the CRI. By using the average, we nullify the effects of those specific zero values and extract the information of the other attributes in the benefit of CRI. Moreover, as we discussed in section 3, we need to hold out an unseen portion of the dataset to apply blindly the transformation to all attributes of all clients, including good and bad loans, and use it as the main part of our proposed algorithm. The model will apply this transformation for every new customer after deploying. As we do not have the label for this test set, therefore we apply the mentioned transformations to all attributes and

create the new features. The test dataset featurized obtains 223,722 observations.

Now, the new attributes in the train and test datasets are ready to compute the CRI based on Formula 3-12. We need to calculate the average of each attribute for the good loans in the training dataset and use φ_i^1 from the Table 3.11.

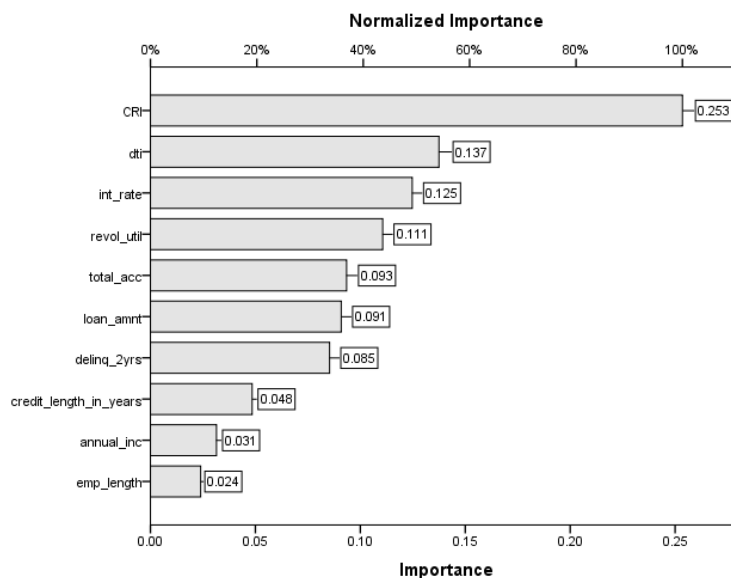
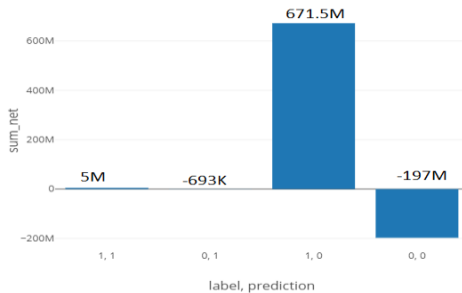
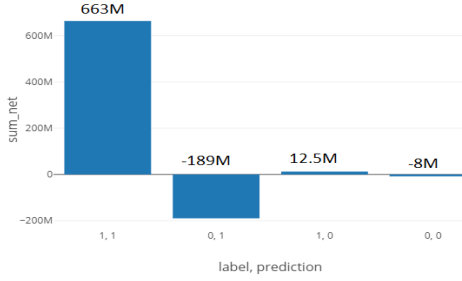
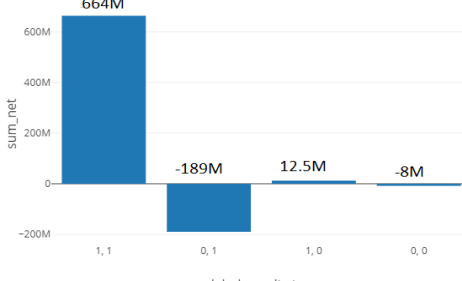


FIGURE 3-4 - NORMALIZED IMPORTANCE OF ATTRIBUTES IN MODELING STAGE.

From Figure 3-4, it is clear that CRI shows an appropriate performance to differentiate the categories of good and bad loans and it could be considered as a candidate to contribute to the modeling phase for improving the accuracy. The resulted dataset consists of categorical and numerical attributes, for both original and transformed values, label variable for the train dataset, and an indicator variable to divide the dataset into test and train. As we discussed, the selected algorithms will be implemented using recursive partitioning with ten-fold cross-validation on the featurized training set to tune the models. The area under the ROC curve (AUC) and recall/sensitivity are computed to evaluate the model's performance in each scenario. The loss amount is also calculated in each scenario to evaluate the model's ability to reduce credit losses. Subsequently, the results of the proposed methodology are detailed, first for each algorithm in Tables 3.12, 3.13, and 3.14, then a summary is represented in Table 3.15 for both statistical and financial performance.

Chapter III – Novel Machine Learning Approach for Online Credit Scoring

TABLE 3.12 - CLASSIFICATION TABLE FOR LOGISTIC REGRESSION WITH RIDGE PENALTY. IN THE TABLE 0, 1 INDICATE NON-DEFAULT LOAN AND DEFAULT LOAN/PAYMENT ARRIARS, RESPECTIVELY, AND LOSS STANDS FOR THE SUBTRACTION OF NET REMAIN OF (0,0) FROM (1,0) COMBINATIONS.

| Logistic regression with the Ridge penalty | | | | | | |
|--|-------|-------------------|-------------------|-----------|---|--|
| | | Pred. negative | Pred. positive | % correct | Metrics | Sum Net (remain) |
| B1: Normal Model | neg | 162516 | 245 | 99.85% | AUC = 0.69 Recall = 0.73 Loss = (474.5) M |  |
| | pos | 60649 | 312 | 0.51% | | |
| | ov. % | 99.75% | 0.25% | 72.78% | | |
| B2: phi model | neg | 12296 | 150465 | 7.55% | AUC = 0.67 Recall = 0.92 loss = (4.5) M |  |
| | pos | 1124 | 59837 | 98.16% | | |
| | ov. % | 6.00% | 94.00% | 32.24% | | |
| B3: phi plus CRI model | neg | 12530 | 150231 | 7.70% | AUC = 0.67 Recall = 0.92 loss = (4.5) M |  |
| | pos | 1126 | 59835 | 98.15% | | |
| | ov. % | 6.10% | 93.90% | 32.35% | | |

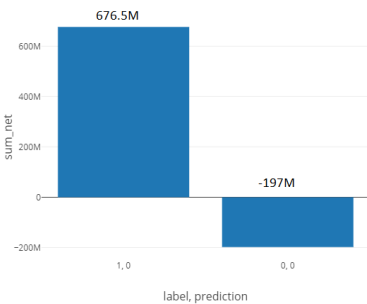
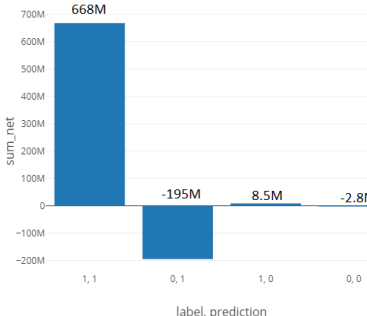
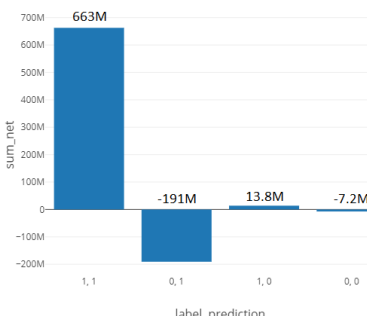
Source: Author's preparation.

We start with the Logistic regression with the Ridge penalty on the original dataset as a benchmark. The model B1 in Table 3.12, represents the results of Logistic regression for the original variables to investigate the possible improvements in the performance of the classification of proposed models, shown in the same Table by B2 and B3. The run time of all three models are very close and around 6.6 minutes by PySpark.

This dataset is representing a very high-risk scenario with a high false-negative rate. This situation is riskier in comparison with the case of the small dataset in the previous section. The amount of loss shows the benefit of correctly identifying a defaulter by the proposed algorithm, and the model with

weighted attributes and CRI allows discriminating the bad loans and minimizing the loss.

TABLE 3.13 - CLASSIFICATION TABLE FOR RANDOM FOREST CLASSIFIER. IN THE TABLE 0, 1 INDICATE NON-DEFAULT LOAN AND DEFAULT LOAN/PAYMENT ARRIARS, RESPECTIVELY, AND LOSS STANDS FOR THE SUBTRACTION OF NET REMAIN OF (0,0) FROM (1,0) COMBINATIONS.

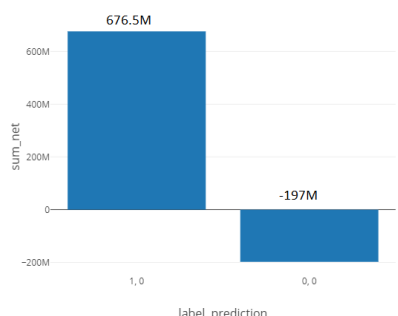
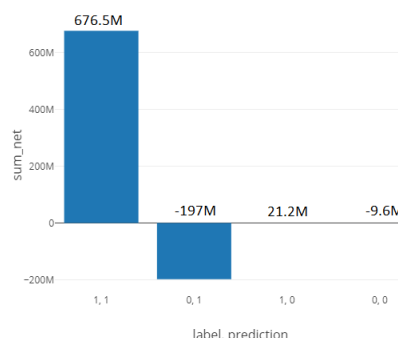
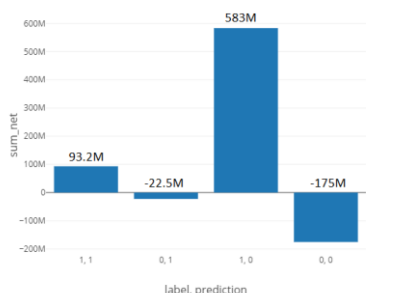
| RANDOM FORESTs CLASSIFIER | | | | | | |
|---------------------------|-------|-------------------|-------------------|-----------|---|---|
| | | Pred. negative | Pred. positive | % correct | Metrics | Sum Net (remain) |
| B4: Normal Model | neg | 162761 | 0 | 100% | AUC = 0.61 Recall = 0.72 Loss = (479) M |  |
| | pos | 60961 | 0 | 0% | | |
| | ov. % | 100% | 0% | 72.75% | | |
| B5: phi model | neg | 1999 | 160762 | 1.23% | AUC = 0.48 RECALL = 0.75 LOSS = (5.7) M |  |
| | pos | 666 | 60295 | 98.91% | | |
| | ov. % | 1.19% | 98.81% | 27.84% | | |
| B6: phi plus CRI model | neg | 6841 | 155920 | 4.20% | AUC = 0.62 RECALL = 0.87 LOSS = (6.6) M |  |
| | pos | 1002 | 59959 | 98.36% | | |
| | ov. % | 3.51% | 96.49% | 29.86% | | |

Source: Author's preparation.

The Logistic regression model with new features could migrate the delinquent customers to the reject area and reduce the loan delinquency rate and subsequently the loss amount. As a trade-off between model sensitivity and specificity, AUC shows almost the same performance among the various scenarios. However, our new Logistic regression obtained a higher sensitivity

(i.e. 0.92) in comparison with the normal model with a sensitivity of 0.73. Overall, the model B1 performs the worst, of which B2 and B3 including phi and CRI features perform best.

TABLE 3.14 - CLASSIFICATION TABLE FOR LINEAR SUPPORT VECTOR MACHINE IN THE TABLE 0, 1 INDICATE NON-DEFAULT LOAN AND DEFAULT LOAN/PAYMENT ARRIARS, RESPECTIVELY, AND LOSS STANDS FOR THE SUBTRACTION OF NET REMAIN OF (0,0) FROM (1,0) COMBINATIONS.

| LINEAR SUPPORT VECTOR MACHINE | | | | | | Sum Net (remain) | |
|-------------------------------|-------|----------------|----------------|-----------|---|---|--|
| | | Pred. negative | Pred. positive | % correct | Metrics | | |
| B7: Normal Model | neg | 162761 | 0 | 100.00% | AUC = 0.50 Recall = 0.72 Loss = (479.5) M |  | |
| | pos | 60961 | 0 | 0.00% | | | |
| | ov. % | 100.00% | 0.00% | 72.75% | | | |
| B8: phi model | neg | 11 | 162750 | 0.01% | AUC = 0.50 RECALL = 0.91 LOSS = (11.6) M |  | |
| | pos | 1 | 60960 | 100.00% | | | |
| | ov. % | 0.01% | 99.99% | 27.25% | | | |
| B9: phi plus CRI | neg | 149196 | 13565 | 91.67% | AUC = 0.52 RECALL = 0.73 LOSS = (408) M |  | |
| | pos | 54141 | 6820 | 11.19% | | | |
| | ov. % | 90.89% | 9.11% | 69.74% | | | |

Source: Author's preparation.

For the second algorithm, we use random forests classifier as an ensemble method. We consider 3, 5, and 10 decision trees to construct each forest and jointly decide upon the credit score. The model B4 in Table 3.13, represents the results of random forests classifier for the original variables to investigate the possible improvements in the performance of the classification of proposed

models, shown in the same Table by B5 and B6. The run time of all three models are very close and around 30 minutes by PySpark.

Table 3.13 demonstrates that there is a significant difference in the performance of the three scenarios. The normal random forests model is not able to predict any unseen sample from samples it has seen during training and it was not able to model the distribution of loan situations, whereas the model B6, which is a random forests classifier with a combination of phi features and CRI, perform significantly better. This shows the usefulness of the new algorithm to improve the performance of the models. Moreover, it successfully maximize the profit and minimize the operational costs and risks as another benefit of this new approach by focusing on correctly identifying a defaulter.

The Linear support vector machine (SVM) algorithm is considered as another sophisticated supervised machine learning technique with expected higher accuracy than Logistic regression. As our data set is large, SVM has high training time (1.05 hours) compare to other algorithms. The results are very similar to the Random forest, however, the amount of loss is dramatically higher than the two last techniques. Regardless of improvement in the performance of the model, with the help of our proposed algorithm, the amount of loss is not improved as much as other classifiers.

Finally, Table 3.15 represents the best model of each classifier, and again, the new Logistic regression B3 produces the best-performing model. It has also high interpretability and has been applying for credit scoring as an industry standard for many years. Model B3 has the highest profit, followed by models B6 and B8. It is interesting to see that having new features in all of these models produce decent profits, whereas the Normal models of the same algorithms do not, at least not when compared to the proposed algorithm.

TABLE 3.15 - THE BEST MODEL OF EACH CLASSIFIER.

| Classifier | Model ID | Feature group | AUC | Recall / Sensitivity | Loss |
|------------------------|----------|---------------|------|----------------------|-------|
| Logistic regression | B3 | PHI, CRI | 0.67 | 0.92 | 4.5M |
| Random forests | B6 | PHI, CRI | 0.62 | 0.87 | 6.6M |
| Support vector machine | B8 | PHI | 0.50 | 0.91 | 11.6M |

Source: Author's preparation.

3.5 IMPACT OF RESEARCH

This section identifies various levels of impact based on the research findings.

3.5.1 Confidentiality and privacy

Transferring the sensitive data from data warehouses of financial institutions to different machines and nodes for parallel or distributed computation is always affected by privacy concerns. Financial service providers try to enhance trust in their systems as a fundamental policy of client rights and maintaining the confidentiality of personally identifiable information is crucial. Additionally, there are some standards and regulations, such as the General Data Protection Regulation (GDPR) in the European Union. The result of this study shows that an index of features can be calculated as an aggregation before data distributing for the mapping stage of the MapReduce algorithm. There is an ethical concern in data anonymization as well because of outliers, which mostly belong to well-known special customers. This indexing can guarantee the confidentiality of sensitive data to provide easier access to parallel computing tasks.

3.5.2 Financial inclusion

Models based on non-traditional data sources such as mobile phone data in the form of call detail records or mobile phone log files, which are examples of Big Data sources, typically suffer from complexity and time-consuming sophisticated algorithms. Despite facilitating credit access to people without historical financial data, the models should be highly accurate to fulfill the expectations of loan providers. Using conservative models for these new sources of data or high-risk situations such as pandemics or economic crisis can help loan providers offer even small credits to underbanked populations, young people, patients, and immigrants, enhancing the assessment of whether the new clients are creditworthy.

3.5.3 Compliance risk impact

Committee on payment and settlement systems in key consideration 3-4-7 of principles for financial market infrastructures⁶ and its explanatory notes declares that the financial systems should have clearly defined procedures for the management of credit and liquidity risks. It should specify the respective responsibilities of the system operator and the participants and provide appropriate incentives to manage and contain those risks. Credit-scoring models should also be considered in provisions and capital buffers calculation by financial institutions according to standards such as the Basel Accords and IFRS9. In this research, we illustrated how credit scoring has to be conservatively formulated to propagate in non-traditional datasets with the potential high-risk of the false-negative rate to detect default. This insight paves the way for loan providers to be able to use new sources of data more soundly and solidly and try to adapt to new emerged technologies without risk management concerns.

3.6 CONCLUSION

This study described a non-parametric statistics approach to assess credit candidate applicants' profiles and continued credit scoring based on non-traditional data. The approach uses a two-step approach based on an initial Kruskal-Wallis analysis and a neural network to learn the model. It introduced a novel credit scoring methodology that reselects significant highly informative features and weighted out by their level of contribution in predicting credit categories of loans to be used in modeling phase. This new credit scoring uses the Kruskal-Wallis non-parametric test, which enables it to be used for two or more categories. Therefore, categories could be "good loans" and "default loans" or even more than two categories such as "good," "doubtful," and "bad" loans as recommended by Basel Accords. The proposed credit risk index is computationally less expensive with reasonable accuracy in comparison with current computationally expensive hybrid algorithms in credit scoring or fixed-

⁶ An FMI should establish explicit rules and procedures that fully address any credit losses it may face as a result of any individual or combined default among its participants with respect to any of their obligations to the FMI.(2012).

weight models in scorecards. The advantages of this approach could be summarized as following:

- Occupying less memory and transferring optimized data in the network by using only the sign of γ_j^t as a flag attribute.
- Having a warning sign to renew the model by a new set of features based on the values and trend of w_j^t .
- Improving the performance of the model and decreasing the false-negative rate.
- Using the CRI as a complementary feature with an interval/ratio scale.

In the classification accuracy, the results showed that this credit scoring method is more informative and conservative. It is able to predict default probability showing good performance with AUC = 0.99 for small dataset and unchanged AUC = 0.67 for big dataset with 18% improvement in Recall and Sensitivity. Thus, this credit index formula as a homogeneous symmetric average is an accurate aggregate measure, able to renew features dynamically and weighted out the attributes as their impact factors. It is suitable for traditional and non-traditional data sets such as regular loan data repositories or new mobile phone data-sets, especially where selecting and extracting the information of features in one aggregated measure is needed for online credit scoring. ■

CHAPTER FOUR - A NEW ENSEMBLE LEARNING STRATEGY FOR PANEL TIME-SERIES FORECASTING WITH APPLICATIONS TO TRACKING RESPIRATORY DISEASE EXCESS MORTALITY DURING THE COVID-19 PANDEMIC

In the previous chapter, we discussed how data science could help banking risk management in a high-risk situation. Although monetary institutions and mainly Central Banks are domain players of financial stability, however, insurance companies are also other important players with different characteristics and risk management approaches in comparison with banks. In this chapter, we are going to focus on this sector of economy and try to develop a new ensemble time-series modelling based on layered learning approach for mortality, longevity, and health risk management in insurance companies. Quantifying and analyzing excess mortality in crises such as the ongoing COVID-19 pandemic is crucial for policymakers, public health officials, and epidemiologists. The traditional way it is measured does not account for differences in the level, long-term secular trends, and seasonal patterns in all-cause mortality across countries and regions. This chapter develops and empirically investigates the forecasting performance of a novel flexible and dynamic ensemble learning strategy for seasonal time series forecasting of monthly respiratory diseases deaths data across a pool of 61 heterogeneous countries. The strategy is based on a Bayesian Model Ensemble (BME) of heterogeneous time series methods involving both the selection of the subset of best forecasters (model confidence set), the identification of the best holdout period for each contributed model, and the determination of optimal weights using the out-of-sample predictive accuracy. A model selection strategy is also developed to remove the outlier models and to combine the models with reasonable accuracy in the ensemble. The empirical results of this large set of experiments show that the accuracy of the BME approach improves noticeably by using a flexible and dynamic holdout period selection. Additionally, that

the BME forecasts of respiratory disease deaths for each country are highly accurate and exhibit a high correlation (0.94) with COVID-19 deaths in 2020⁷.

4.1 INTRODUCTION

Quantifying and analyzing excess mortality - the count of fatalities from all causes above and beyond what would have been predicted under normal (baseline) circumstances for a given period in a population - in crises (pandemic or epidemic disease, natural disasters, military conflicts, displacement situations, political repression, hunger) is highly relevant for policymakers, public health officials and epidemiologists (Checchi & Roberts, 2005).

The impact of the ongoing coronavirus 2 (SARS-CoV-2) pandemic on a given country is typically measured by the number of confirmed cases and the death toll, two statistics that have been collected and reported daily by institutional-based repositories of public health data such as the one maintained by the World Health Organization (WHO), the European Centre for Disease Prevention and Control (ECDC) or the Johns Hopkins University (Dong et al., 2020). However, it is well-known that both metrics are severely affected by limited testing capacity across countries and within countries over time, different standards regarding the classification of COVID-19 related deaths, systematic measurement errors, and data completeness and accuracy problems. Data quality problems produce biased and inconsistent parameter estimates and lead to flawed conclusions in epidemiological analysis (Ashofteh & Bravo, 2020). This is a matter of countless concern since epidemiological models have been used and will continue to be used worldwide to inform national and local authorities on topics such as forecasts of the numbers of deaths, hospital utilization rates, the impact of quarantine, stay-at-home orders, or physical distancing measures, the adoption of travel restrictions, when to re-open the economy, or the impact of vaccination campaigns (Cui, et. al., 2021).

⁷ Please cite this chapter as: Ashofteh, A., Bravo, J. M., Ayuso, M. (2021). A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic. In Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao 2021- Proceeding of 21th Conference of the Portuguese Association for Information Systems, CAPSI 2021.

Because of that, estimates of the excess mortality are considered a better and more robust indicator for monitoring the dynamics and consequences of the ongoing SARS-CoV-2 pandemic and for comparing the experience of different countries or regions where either the degree of misdiagnosis or underreporting or data quality problems may differ (Leon et al., 2020; Banerjee et al., 2020; Kontis et al., 2020; Islam et al., 2021). The use of excess mortality data can also capture mortality variation indirectly related to COVID-19 infection, for instance, the increase in mortality due to delayed or deferred health care during the pandemic or due to increases in mental health disorders (e.g., depression, suicide, increased alcohol or opioid use, domestic violence), or mortality declines due to reduced traffic accidents or occupational injuries (because of general lockdown or travel restrictions), or due to reduced transmission of other viruses (e.g., influenza, AIDS).

Excess mortality is typically measured by national or supranational statistical agencies using the absolute, relative (P-score) or standardized (Z-score) number of “excess” deaths, where the benchmark is often computed in a very naïve way, for instance using the simple average of the previous year’s deaths. The EuroMOMO project (<https://www.euromomo.eu>) is a notable exception, with baseline mortality modelled using a generalized linear model corrected for overdispersion assuming the number of deaths follows a Poisson distribution. This approach does not account, for instance, for differences in the level, long-term secular trends, and seasonal patterns in all-cause mortality across countries and regions.

Against this background, this chapter develops and empirically investigates the forecasting performance of a novel flexible and dynamic ensemble learning strategy for seasonal time series forecasting. The strategy is based on a Bayesian Model Ensemble (BME) of heterogeneous models involving both the selection of the subset of best forecasters (model confidence set) to be included in the forecast combination, the identification of the best holdout period for each individual contributed model, and the determination of optimal weights using the out-of-sample predictive accuracy. A model selection strategy is also developed to remove the outlier models and combine the models with reasonable accuracy in the ensemble. The novel approach is empirically investigated using monthly respiratory diseases deaths data for 61 heterogeneous countries.

This new approach is based on the forecasting model selection and model combination, which are the two contending approaches in the time series forecasting literature. The customary approach to seasonal and non-seasonal time series forecasting is to adopt a single believed to be the best model for each series chosen from the set of candidate models using some criteria or procedure (e.g., information criteria, forecasting accuracy measure, cross-validation, bootstrapping, construction of confidence intervals, hypothesis testing for nested models), often neglecting model and parameter risk for statistical inference purposes. To this end, a growing number of linear and non-linear univariate and multivariate times series methods and statistical machine learning techniques are proposed to increase the short- and long-term predictive accuracy on a wide range of problems, including stochastic population – mortality, fertility, net migration - forecasting (Hyndman and Ullah, 2013; Bravo and Coelho, 2019), epidemiological and excess mortality forecasting (Scortichini et al., 2020) and the pricing of longevity-linked securities (Bravo and Nunes, 2021).

Empirical studies in multiple areas show that it is hard to find (if exists) a single widely accepted forecasting method that performs consistently well across all data sets and time horizons (Aiolfi and Timmermann, 2006; Chatfield, 2016). The use of different selection methods, different fitting periods, alternative accuracy measures, structural breaks in the data generating process, and misspecification problems can lead to different model choices and time series forecasts. To tackle the model risk problem, i.e., the uncertainty regarding the identification of the true data generating process and the best fitting or forecasting method, to improve the forecasting accuracy, to tackle the limitations of some methods, and to generate comparable cross-country and/or sub-national forecasts, an alternative approach is to use an ensemble of heterogeneous time series methods and algorithms, each one carrying new information that is not encompassed in other forecasting techniques social policy design and reform.

Since the original work of Bates and Granger (1969), several comprehensive theoretical and empirical studies have confirmed the superior predictive performance of ensemble methods using different approaches (Makridakis and Winkler, 1983; Breiman, 1996; Ueda and Nakano, 1996), including stacking and

blending to improve-predictions, bagging to decrease variance or boosting to decrease bias (Akyuz et al. 2017) and Bayesian Model Ensemble (Raftery et al. 2005; Bravo et al. 2021a,b; Ayuso et al. 2021; Bravo, 2021). When adopting this empirical strategy, choices must be made with regards to which models to include in the combined pool and with regards to each model contribution (weight) in the final prediction. A significant body of literature has examined optimal model combination weights (see, e.g., Capistrán et al., 2010), focusing either on the selection of optimal combination schemes and weights (De Menezes et al., 2000; Jose and Winkler, 2008; Andrawis et al., 2011; Hsiao and Wan, 2014), assigning equal weights to the set of superior models (Stock and Watson, 2004; Samuels and Sekkel, 2017), selecting a subset of best models among the set of candidates (model confidence set) using a dynamic trimming scheme and considering the model's out-of-sample forecasting performance in the validation period (Bravo et al. 2021a; Bravo and Ayuso, 2020, 2021), or using meta-learning (Brazdil et al. 2009) or regret minimization (Cesa-Bianchi and Lugosi 2006) approaches to choose the best models for contributing to the ensemble model. To cope with concept drift, memory, change detection, learning, and loss estimation adaptive algorithms have been proposed (Gama et al. 2014).

Theoretically, any potential model carrying useful information may be considered in the model space. In real-world applications, the marginal benefit of adding forecasts to the model confidence set may be small if a sufficient number of models capturing the data generating process has been included, if (especially in small samples) the cost in terms of extra parameter risk overcomes the gains in terms of forecasting accuracy, and if the diversity of models within the pool of heterogeneous models is such that includes a significant percentage of the worst forecasters in past data and/or the validation dataset and the degree of disagreement within the ensemble is substantial. In these cases, model combinations tend to retain some bias in their joint predictions. In such circumstances, the use of windowing approaches including trimming models could lead to better estimates of each model's weight in the combined forecast (Aiolfi and Timmermann, 2006). The performance of model combinations is high when the individual models included in the pool exhibit a consistent forecasting performance. Excluding worst forecasters from the pool or assigning them a very low weight minimizes the impact of parameter risk and is likely to achieve a better bias-variance

trade-off. Building better model combinations to solve real-world time series problems has become a critical and active research area in recent years (Khairalla et al. 2018).

In this Chapter we develop and empirically investigate the forecasting performance of a novel flexible and dynamic ensemble learning strategy for seasonal time series forecasting of monthly respiratory diseases deaths data across a pool of 61 heterogeneous countries. The strategy is based on a Bayesian Model Ensemble (BME) of heterogeneous time series methods involving both the selection of the subset of best forecasters (model confidence set), the identification of the best holdout period for each contributed model, and the determination of optimal weights using the out-of-sample predictive accuracy. A model selection strategy is also developed to remove the outlier models and to combine the models with reasonable accuracy in the ensemble. The novel approach is empirically investigated using monthly respiratory diseases deaths data for 61 heterogeneous countries.

The pool of candidate models considered in this study includes traditional linear and non-linear univariate time series methods and novel statistical machine learning techniques. We consider a range of both existing and new models that have proven to perform well in fitting and forecasting empirical studies. The set of candidate models includes Seasonal Trend Decomposition using Loess for estimating nonlinear relationships (STL), a Seasonal Naïve random walk forecast model (SNAÏVE), the classical Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model, the Exponential Smoothing State Space Model (ETS), the Holt-Winters' multiplicative method (HWM), the Holt-Winters' additive method (HWA), Random Walk with drift model (RWF), Extreme Learning Machine methods (ELM), Multilayer Perceptron for time series (MLP), a Neural network autoregression model (NNAR), the TBATS model and Singular spectrum analysis (SSA). We examine and compare run times, accuracy, level of contribution, and error metric of the proposed ensemble techniques in comparison with traditional ensemble model and individual forecasting models.

The proposed ensemble learning procedure involves: (i) setting the different holdouts to be checked for each contributed model; (ii) choosing the best

holdout for each model based on the out-of-sample forecasting accuracy; (iii) Selecting the subset of best forecasters (model confidence set) using a variable trimming scheme in which a multiple of the forecasting accuracy metric range obtained across all candidate models is used as the threshold for model exclusion; (iv) the determination of each model posterior probabilities (model weights) using the normalized exponential (Softmax) function; and (v) finally, ensemble forecasts are obtained based on the law of total probability considering the model confidence set and the corresponding model weights. Contrary to previous approaches focusing either on the selection of optimal combination schemes and weights or equally weighting a subset of best forecasters, our ensemble procedure involves, for each dataset, both the identification of the best holdout period for each model, the selection of the best forecasting models and the determination of optimal weights based on the out-of-sample forecasting performance.

Our empirical results show proposed approach leads to a decrease in the individual error of ensemble members in comparison with normal model selection with equal holdouts for selected models, and without overly decreasing the diversity among them. Hopefully, this article brings more clarity on which time series techniques contribute better to ensembles, and presents a suitable ensemble time series with improved predictive accuracy. All illustrated under the empirical application of predicting the excess mortality produced in the year 2020.

The remaining sections of the Chapter are organized as follows. In section 2, we provide the materials, methods, and related works considered in this research. Section 3 describes our proposed method. The results of an extensive set of experiments on respiratory disease deaths of 61 countries are given and discussed in Section 4. Finally, the main discussion and conclusion are presented in section 5.

4.2 MATERIALS AND METHODS

The proposed method is based on a meta-learning approach to adopt the ensemble to the best combination of forecasting models. The candidate models are extracted from different layers with the best holdout for each contributed model and each panel member. The Figure 4-1 shows a graphical abstract of

the materials and methods. We use multiple learning processes to improve the predictive performance of the ensemble. It is built by an ensemble learning approach from the addressed candidates with the last layer. In this section, we discuss these techniques in brief and highlight their contributions as well.

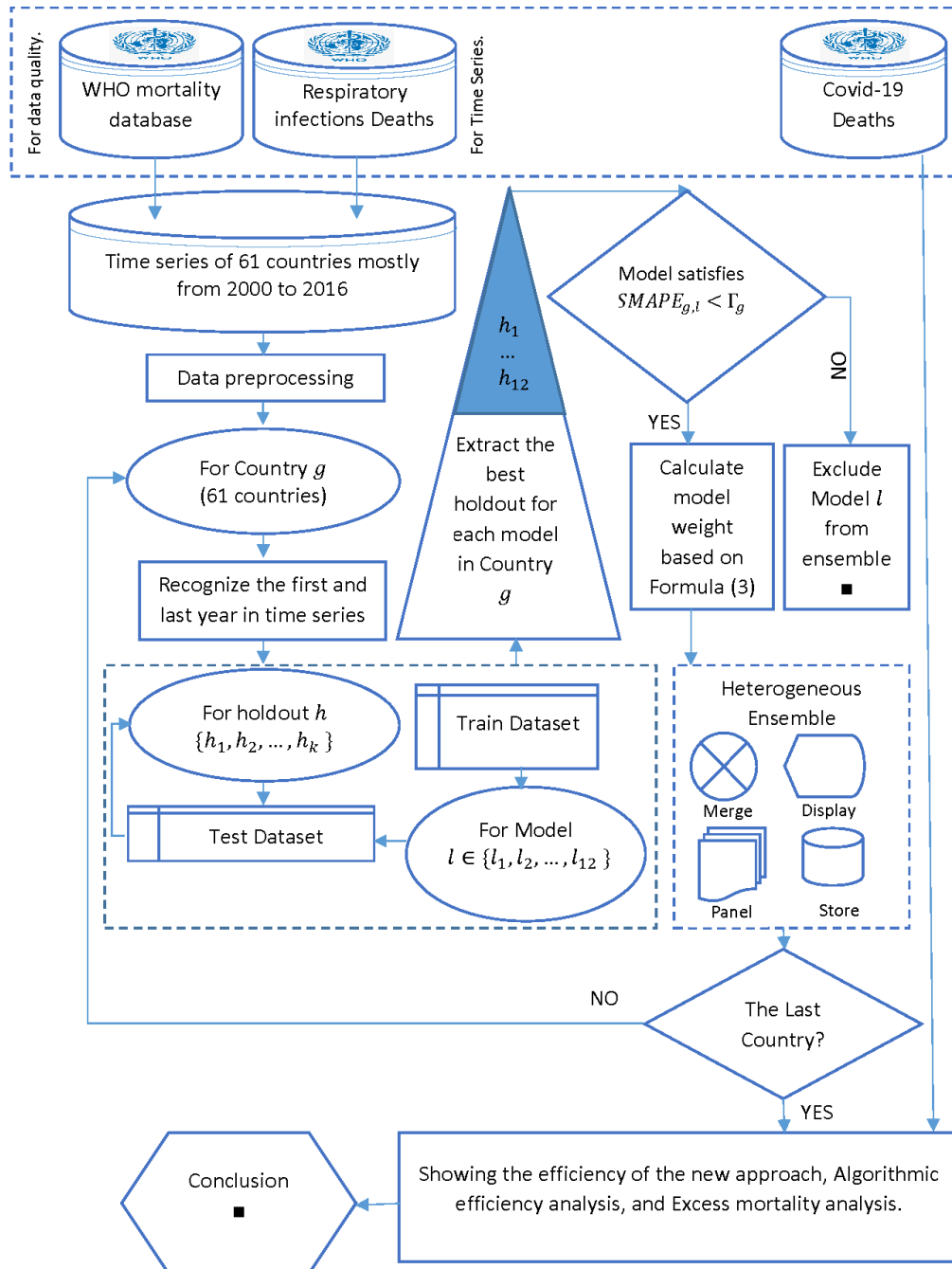


FIGURE 4-1 - GRAPHICAL ABSTRACT OF THE PROPOSED DYNAMIC ENSEMBLE LEARNING STRATEGY.

4.2.1 Layered learning and the proposed ensemble learning strategy

The layered learning approach in time series consists of breaking a forecasting problem down into simpler subtasks in several layers. Each layer addresses a different predictive task and the output of one layer could be used as the input of the next layer (Cerqueira et al. 2020). In this research, the first task is to obtain a direct mapping from the time series of different countries, combining the intractable time series algorithms, and predicting the ensemble model as the final output. Therefore, the task of the first layer is finding the best holdout for each panel member and for each time series algorithm. It facilitates the task of model selection in the second layer, which facilitates the identification of the model confidence set of best forecasters in the last layer. It is useful to maximize the forecasting accuracy in panel time series, which is done dynamically, and adapt the learning process of the model to possible unexpected shocks.

Along with the Layered Learning approach, our ensemble method runs multiple learning algorithms to employ adaptive heuristics to combine forecasters. As a result, it obtains better predictive performance than could be obtained from any of the constituent learning algorithms. It consists of several selected models with the best performance based on minimum error measures. Each model considers different holdouts to solve the problem at hand and then the best holdout will be chosen for each model. This leads to a more robust overall performance of the ensemble by increasing the diversity in the holdouts; however, the length of time series would be different according to their different holdouts. It could be problematic for the ensemble layer to merge the models with different lengths. It is necessary to force all selected models to have equal length and finally the length of the ensemble would be equal to the minimum length time series in our time series set. Although this windowing strategy can offer the best prediction of each forecaster and as a result, the best performance for the ensemble, but it is clear that for the best results, the length of all time series should be enough large and almost the same.

Let each candidate model be denoted by M_l , $l = 1, \dots, L$ representing a set of probability distributions in which the "true" data-generating process is assumed to be included, comprehending the likelihood function $L(y|\theta_l, M_l)$ of the observed data y in terms of model-specific parameters θ_l and a set of prior probability densities for said parameters $p(\theta_l|M_l)$. Consider a quantity of

interest Δ present in all models, such as the future observation of y . The marginal posterior distribution across all models is

$$p(\Delta|y) = \sum_{l=1}^L p(\Delta|y, M_l) p(M_l|y) \quad (4-1)$$

where $p(\Delta|y, M_l)$ denotes the forecast PDF based on model M_l , and $p(M_l|y)$ is the posterior probability of the model M_l given the observed data. The posterior probability for the model M_l is denoted by $p(M_l|y)$ with $\sum_{l=1}^L p(M_l|y) = 1$. The weight assigned to each model M_l is given by its posterior probability

$$p(M_l|y) = \frac{p(y|M_l)p(M_l)}{\sum_{l=1}^L p(y|M_l)p(M_l)}. \quad (4-2)$$

The workflow of our proposed method is presented in Figure 4-2. To identify the model confidence set and compute model weights, for each dataset we first set the different holdouts to be checked for each contributed model. Let $H = \{h_1, h_2, \dots, h_k\}$ represent the set of holdout periods to be considered in the estimation procedure.

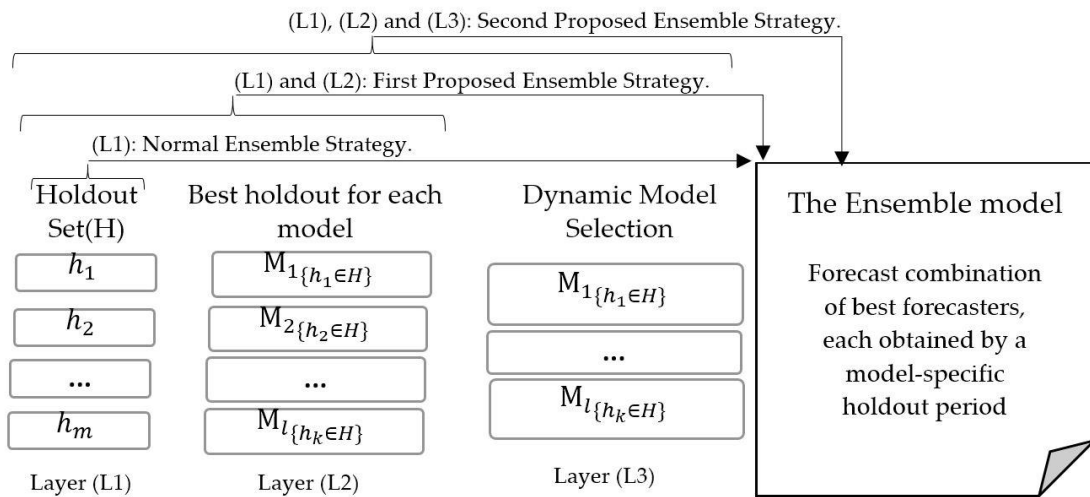


FIGURE 4-2 - PROPOSED STRATEGY OF ENSEMBLE LEARNING.

The second step is about choosing the best holdout for each candidate model based on the out-of-sample forecasting accuracy measure. We use the symmetric mean absolute percentage error (SMAPE) as the forecasting

accuracy measure.⁸ For choosing the best holdout for each model, we tested the different values of the holdouts from three to ten and consider the holdout set ($H = \{3,5,7\}$ years⁹) as representatives of short, medium, and long term, and compared the SMAPE's values at each iteration, keeping the model with the lowest SMAPE as the candidate for the model confidence set selection step. This provides an opportunity to cover different parts of the data space and to handle different dynamic regimes in different candidate time series. Additionally, it will empower the final ensemble model with managing the limitations of each one in the others.

Third, the subset of best forecasters is selected using the best holdout period and a variable trimming scheme in which a multiple θ (pre-set at 0.5) of the distance between maximum and minimum of the forecasting error metric is used as the threshold for model exclusion, i.e., using

$$\Gamma_g = \frac{\max\{SMAPE_{g,l}\}_{l=1,\dots,K} - \min\{SMAPE_{g,l}\}_{l=1,\dots,K}}{2}. \quad (4-3)$$

where $SMAPE_{g,l}$ is the SMAPE value for model l in the dataset (country) g . For each dataset, if the forecasting accuracy of a candidate model is greater than the Γ_g indicator, (i.e., $SMAPE_{g,l} > \Gamma_g$) the model is excluded from the model confidence set and of the ensemble forecast computation, i.e., it receives a zero weight in (4-1).

Depending on the distribution of the SMAPE values, the number of models excluded from the model confidence set can be high or small. From a frequentist point of view, building up a model confidence set is a way of summarizing the relative forecasting performances of the entire set of candidate models and identifying the set of statistically best forecasters. The advantage of this statistic defined in (4-2) is its simplicity, easiness of application, and interoperability. Additionally, it falls between the close models to the time series and extremely far models. In this case, the far forecasting

⁸ We avoid using the AIC or BIC criteria because the candidate models are in different model classes, and the likelihood is computed in different ways. For selected models in the same class the BIC is useful, and it is used automatically by the algorithm to select, for instance, an SARIMA model between candidate SARIMA models. Another caution of the error term in ensemble modelling could be avoided using the accuracy measures with logarithm in their formula such as MSLE, RMSLE, and SLE. According to our experiment, the program would be interrupted because of some possible negative values presented to these measures by some algorithms.

⁹ The results for the other holdout periods are consistent with the ones, that we report in this paper.

models will be removed from the ensemble. This could be magnificent to avoid overfitting and control the redundancy in the output of the ensemble model. The intuition is that the models with a minimum error are the closest to the actual data generating process. Comparing the error measure with the mean of the errors removes only models which are extremely far away from other candidate models. It will save the diversity of the selected models and prevent the overfitting problem.

Fourth, the best forecasters model posterior probabilities (model weights) are computed using the normalized exponential (Softmax) function

$$p(M_l|y) = \frac{\exp(-|\xi_l|)}{\sum_{l=1}^L \exp(-|\xi_l|)}, \quad k = 1, \dots, K., \quad (4-4)$$

with $\xi_l = S_l / \max\{S_l\}_{l=1, \dots, L}$ and $S_l := SMAPE_{g,l}$. The Softmax function is a generalization of the logistic function often used in classification and forecasting exercises using traditional, machine learning, and deep learning methods as a combiner or an activation function (Sergio et al., 2016). The function assigns larger weights to models with smaller forecasting errors, with the weights decaying exponentially the larger the error. Fifth, the Bayesian model ensemble forecasts are obtained based on the law of total probability (4-1) considering the model confidence set and the corresponding model weights (4-3). The sampling distribution of the ensemble forecast of the quantity of interest is a mixture of the individual model sampling distributions.

The pseudo-code of the proposed methodology is listed in Table 4.1.

TABLE 4.1. PSEUDO CODE OF THE PROPOSED ENSEMBLE STRATEGY.

| |
|---|
| INPUT panel time series (panel members = countries); OUTPUT ensemble model |
| 1. STATEXPLORE time series decomposition |
| 2. IMPUTE[missing] = TRUE |
| 3. First_year = 2000 (for most of time series but some of them start later) |
| 4. Last_year = 2016 |
| 5. Target_year = 2020 |
| 6. Confidence_level = 0.95 |
| 7. Holdout_set={3, 5, 7} and SET Tetra = 0.5 |
| 8. Ensemble_criteria_for_computing_weights = "Symmetric Mean Absolute Percentage Error (SMAPE)" |
| 9. Set.seed() |

```

10. Model_list = {SNAIVE, RWF, HWA, HWM, ETS, SARIMA, TBATS, STL, NNAR, MLP,
    ELM, SSA, ENS}
11. FUNCTION model_weights (error)
12.   Pr = error/max(error)
13.   exp(-abs(pr))/sum(exp(-abs(pr)))
14. # First loop for selecting country
15. FOR each panel in list of countries DO
16. {   SET panel.data = SUBSET dataset(country = panel & Year > First_year &
    Months="Jan-Dec"
17.   SET Year_min = min(Year of panel.data)
18.   panel_data = MISSING VALUE IMPUTATION by na_seasplit
19.   SET (START of the run-time calculation)
20. # Second loop for selecting holdouts
21.   FOR each holdout in Houldout_set DO
22.   {
23.     IF ( ymax-ho+1 < ymin+3 ) { break }
24.     ELSE
25.       SET train_dataset WINDOW (START = Year_min , END = Last_year -
    holdout)
26.       SET test_dataset WINDOW (START = Last_year - holdout + 1)
27.       FIT models in Model_list
28.       CALCULATE accuracy (model , holdout)
29.       IF accuracy (model[holdout]) > last_accuracy (model[holdout - 1])
    THEN
30.         SET model = model[holdout]
31.         ELSE
32.         SET model = model[holdout -1]
33.   }
34.   CALCULATE error(ALL models), min_error(ALL models), max_error(ALL
    models)
35.   CALCULATE id_error = Teta × (min_error + max_error)
36.   FOR model in Model_list
37.   {
38.     IF ( error_model > id_error) THEN
39.       PRINT ("Model is excluded!")
40.     ELSE
41.       ADD model into selected_model_list
42.   }
43. # The model ensemble
44. IF selected_model_list = NULL {next country}
45. ELSE
46.   {
47.     CALCULATE model_weights for ensemble
48.     SET First_year based on the model with min_holdouts
49.     SET First_month based on the model with min_holdouts
50.     CALCULATE ensemble
51.     SET (END of the run-time calculation)
52.   }
53. # The outputs
54. PRINT GRAPHS
55. SAVE OUTPUTS }

```

Source: Author's preparation.

In the interest of reproducible science, the dataset and all methods will be publicly available (Ashofteh et al. 2021c).

4.2.2 The learning algorithms

This section summarizes the characteristics of the individual candidate learning algorithms (times series methods) used in this study. For a detailed presentation and discussion of the methods see, for instance, Hyndman and Athanasopoulos (2021). The Seasonal Trend Decomposition uses Loess (STL) is a filtering procedure for decomposing a time series into the trend and seasonal components. It is based on the Loess smoother and offers a simple, versatile, and robust method for decomposing a time series and estimating nonlinear relationships (Cleveland et al. 1990). The models should be robust against the outliers detected in the multiple panel member's (countries) datasets. In specifying the STL, we use a robust decomposition such that sporadic abnormal observations do not affect the estimates of the trend-cycle and seasonal components. The time series are tested for autocorrelation using the Ljung-Box test, considering the null hypothesis that the model exhibits appropriate goodness-of-fit. The method does not handle the calendar variation automatically, and it only provides facilities for additive decompositions, which could be considered as a limitation of this approach. We use the two parameters *t.window* and *s.window* to control how rapidly the trend-cycle and seasonal components can change. Smaller values allow for more rapid changes that we need strongly for some time series with strong turning points. As a result, the number six was chosen for *s.window* and *t.window* by looking at the results of the check residuals and Ljung-Box Test statistics.

The seasonal naive (SNAIVE) method sets the forecast to be equal to the last observed value from the same season of the year (i.e., the same month of the previous year). It is a useful benchmark for other forecasting methods, and we found out that it is very helpful to show the recent trend of time series and to adjust the ensemble model for the trend component.

Similarly, the SARIMA and the Random Walk Forecasts (RWF) as an SARIMA(0,0,0)(0,1,0)*m* model, where *m* is the seasonal period, are used as

state-of-the-art methods to memorize the repeating monthly patterns. However, many SARIMA models have no exponential smoothing counterparts, and the robust univariate forecasting models such as Holt-Winters' multiplicative method (HWM) and the Exponential Smoothing State Space Model (ETS) could be considered as a good complimentary for SARIMA models in our final Ensemble. All ETS models are non-stationary, while some SARIMA models are stationary. ETS follows the last trend of the time series and it is appropriate for the Ensemble model to empower the trend parameter in the final predictions. ETS point forecasts are equal to the medians of the forecast distributions. For models with only additive components, the forecast distributions are normal, so the medians and means are equal. For multiplicative errors, or multiplicative seasonality, which perform similarly in most time series analyzed in this study, the point ETS forecasts will not be equal to the means of the forecast distributions. In these cases, SARIMA is a better choice. On the other hand, ETS is a non-linear exponential smoothing model with no equivalent SARIMA counterpart. Therefore, we propose the ETS model to be selected automatically and the type of trend and seasonal component to be additive with the restriction of finite variance. The bootstrapping method for resampled errors was used rather than distributed errors and simulation was used rather than algebraic formulas for calculating prediction intervals. The other options for the ETS model can be seen in Table 2. The TBATS are also used to adopt the ensemble model with multiple seasonality of some time series. The TBATS stands for (T)rigonometric terms for seasonality, (B)ox-Cox transformations for heterogeneity, (A)RMA errors for short-term dynamics, (T)rend, and (S)easonal.

Regarding the neural network time series algorithms, the Extreme Learning Machines (ELM) was used with the Lasso penalty. ELM theory assumes that the randomness in the determination of coefficients of neural network predictors (input weights) can feed the learning models with no particular iterative tuning for any distribution as is the case in gradient-based learning algorithms. The model entails randomly defined hidden nodes and input weights without any optimization such that only output weights need to be calibrated during the training of the ELM (Huang et al., 2004). In the hyperparameter calibration of the ELM, we consider the maximum 500 hidden layers for 200 networks to be trained and sum up in ELM's final ensemble forecast model.

The Neural network autoregression (NNAR) refers to single hidden layer networks using the lagged values of the time series as inputs and automatic selection of parameters and lags according to the Akaike information criterion. In the NNAR model specification, we considered the last observed values from the same season as inputs to capture the seasonality patterns and use size equal to one, because we have one attribute without regressor, and for improvement, we use one hundred networks to fit with the different random starting weights and then averaged for producing forecasts. Additionally, we consider the Multilayer Perceptron (MLP) as a kind of NNAR-Neural Network Autoregression Model. It is more complicated and advanced than NNAR with three components in the form of $NNAR(p, P, k)$, in which p denotes the number of lagged values that are used as inputs and usually is chosen based on an information criterion, like AIC, P denotes the number of seasonal lags, and k denotes the number of hidden nodes.

Finally, Singular spectrum analysis (SSA) is used as one of the high-quality modeling approaches. The calibration of the SSA is an important but not easy task in a standalone modeling approach. It depends upon two basic parameters: the window length, and the number of eigentriples used for reconstruction. The choice of improper values for these parameters yields incomplete reconstruction, and the forecasting results might be misleading. In this study, we set length equal to 12 and eigentriples equal to NULL. Table 2 summarizes the hyper-parameters of the algorithms used in this study.

TABLE 4.2 - ALGORITHMS AND HYPER-PARAMETERS CHOICES.

| ID | Algorithm | Parameters | Value |
|--------|---|------------|--------|
| STL | Seasonal Trend Decomposition using Loess | lambda | "auto" |
| | | t.window | 6 |
| | | s.window | 6 |
| | | biasadj | TRUE |
| SNAIVE | Seasonal naive | drift | F |
| | | lambda | 0 |
| | | level | clevel |
| | | biasadj | TRUE |
| ARIMA | The Auto-Regressive Integrated Moving Average | Auto | |

| | | | |
|---------------|---|--|--|
| ETS | The Exponential Smoothing State Space Model | Model Box-Cox tran. Multiplicative trend restricted for the models with infinite variance | {ETS, TBATS} ¹⁰ ZZA TRUE Allow TRUE |
| HWM | Holt-Winters' multiplicative method | Seasonal level | Multiplicative clevel |
| HWA | Holt-Winters' additive method | Seasonal level | Additive clevel |
| RWF | Random Walk Forecasts | Drift Lambda Level biasadj | F "auto" clevel TRUE |
| ELM | Extreme Learning Machines | type hd comb reps difforder | Lasso 500 mean 200 NULL |
| MLP | Multilayer Perceptron for time series | Comb hd.auto.type hd.max | Mode Valid 5 |
| NNETAR | Neural network model to a time series | P size decay lambda repeats MaxNWts | 2 1 0.001 Auto 100 2000 |
| SSA | Singular spectrum analysis | Kind svd.method L neig force.decompose mask | 1d-ssa Auto 12 NULL TRUE NULL |

Source: Author's preparation.

The model fitting, forecasting, and simulation procedures have been implemented using R statistical software considering libraries such as the TSA,

¹⁰ The ETS method with automatic and ZZA parameter setting from the forecast statistical software R package (R. Hyndman et al., 2020), and the TBATS method, which includes Box-Cox transformation, ARMA errors, trend and seasonal components (de Livera, Hyndman, & Snyder, 2011).

Metrics, nnfor, tsfknn, Rssa, rpatrec, and forecast (see, e.g., R. Hyndman et al., 2020).

4.3 EMPIRICAL EXPERIMENTS

4.3.1 Data selection and cleansing

In this study, we use cause-of-death data from the World Health Organization (WHO) mortality database (World Health Organization, 2018). The database collects cause-of-death statistics from country civil registration systems and estimates from the United Nations Population Division for countries that do not regularly report population data.

We use an Excel file¹¹ of this database to evaluate the data quality of different countries and a CSV file that includes the death time series of different countries for all genders.

TABLE 4.3. DIFFERENT LEVELS OF QUALITY ALLOCATED FOR THE REPORTED RESPIRATORY DISEASE DEATHS BY COUNTRIES.

| Rank | Evaluation | Description |
|------|-------------------|---|
| 1 | Excellent quality | These countries may be compared, and time series may be used for priority setting and policy evaluation. |
| 2 | Moderate quality | Data have low completeness and/or issues with cause-of-death assignment, which likely affect estimated deaths by cause and time trends. Comparisons among countries should be interpreted with caution. |
| 3 | Low quality | Data have severe quality issues. Comparisons among countries should be interpreted with caution. |
| 4 | Unacceptable | Death registration data are unavailable or unusable due to quality issues. Estimates may be used for priority setting; however, they are not likely to be informative for policy evaluation or comparisons among countries. |
| 5 | Ignorable | Data should be ignored |

Source:(World Health Organization, 2018).

¹¹ https://www.who.int/healthinfo/global_burden_disease/GHE2016_Deaths_2016-country.xls?ua=1

The first mentioned file distinguishes the quality of data for each country by using green, yellow, and red colors. Green countries have multiple years of national death registration data with high completeness and quality of cause-of-death assignment. Estimates for these countries may be compared and time series may be used for priority setting and policy evaluation. However, this dataset only includes data for 2000, 2010, 2015, and 2016 and it is not complete for the time series. As a result, we used this dataset to identify the countries with high-quality reported data to the WHO and rank them according to the data quality. According to the Metadata of the dataset, we ranked the data quality of countries as is shown in Table 4.3.

We considered only countries with data quality ranked in the first three categories and removed some islands because of the lack of data (e.g., Aland Island). We have also cleaned the dataset by removing the total column, and some rows with an unknown month and zero deaths. Some countries reported the total death for three months in one row for some years. We divided this aggregate value into three equal values for each corresponding month. We filtered the datasets for respiratory diseases and considered the death variable as a univariate time series with monthly sampling frequency. Table 4.4 shows the codes that were classified as respiratory infections.

TABLE 4.4. METADATA OF CODE OF DISEASES CATEGORIZED AS RESPIRATORY DISEASE.

| Code | Description |
|------|---|
| 380 | For Respiratory infections (This code is the aggregate of 390 and 400) |
| 390 | For Lower respiratory infections |
| 400 | For Upper respiratory infections |
| 410 | Otitis media: Acute otitis media (AOM) is a common complication of upper respiratory tract infection whose pathogenesis involves both viruses and bacteria. |

Source: (World Health Organization, 2018)

For obtaining the total deaths caused by respiratory diseases, we had to aggregate either the codes 380 and 410 or equivalently the codes 390, 400, and 410. We also made some corrections in the name of countries (Appendix 2). From this, we calculated the proportion of deaths caused by respiratory diseases. To estimate the number of monthly deaths caused by respiratory diseases, we multiply the annual proportion by the total forecasted deaths each

month. We used the fraction of annual deaths of respiratory diseases over total deaths as a proportion of deaths in each month. The procedure provided us with a dataset with more than twelve thousand observations in a pool of 61-panel members' time series (countries) from 2000 to 2016. These panel time series cover the different possible situations of stationarity, non-stationarity, increasing trends, seasonality, and structural breaks to evaluate the accuracy improvement of candidate and ensemble models in different scenarios comprehensively.

According to the different data quality of countries/territories/areas regarding case detection, definitions, testing strategies, reporting practice, and lag times, it is normal to have missing values in the time series dataset. To deal with this problem, we tested the *Kalman*, *seasplit* and *seadec* algorithms to impute the missing values. From these algorithms, the *seasplit* shows the best performance both for saving the trend and the seasonality for our dataset. We impute only missing values within the time series and not at the beginning of the time series with a start date after 2000. As a result, instead of changing the first year of the time series to our base year 2000, we use the latest year available. To avoid the error caused by combining time series with different lengths in an ensemble model, we adapted the R code (Appendix 3) to handle different start years. The same problem occurs as a result of the procedure adopted to select the best holdout for each model, which may ultimately lead to model combinations considering forecasts based on different holdouts, i.e., different time series lengths. Therefore, adoption of the R software's code seems necessary to combine the models correctly. We considered at least three years of data remaining into the time series dataset to candidate a holdout to be tested for performance improvement.

4.3.2 Results

4.3.2.1 Forecasting accuracy comparison

Table 4.5 reports, for the three alternative holdout periods investigated, the predictive accuracy metrics obtained using three alternative backtesting procedures.

In the first approach entitled “Fixed holdout”, we use a fixed holdout period equal to 3, 5, and 7 years to derive the composite (ensemble) model. As comparing the different approaches for combining forecasting models are mostly based on estimates of overall predictive performance, we present three approaches in Table 4.5. In the first approach entitled “only holdout”, we only use a set of different forecasting models to make the ensemble model by different holdouts. As we can see, some models are exhibit better performance when compared with the ensemble model. Even in average error, the TBATS shows a lower error than the ensemble.

The second approach is named “Holdout and selection”. This approach uses a multiple of the range of SMAPE values across all methods to evaluate the distance of each model to the remaining others as shown above in the Pseudocode (Table 4.1). The model’s with SMAPE values higher than the range indicator are considered poor forecasters and eliminated from the ensemble forecast. The results in Table 4.5 highlight the improvement in the accuracy of the Bayesian model ensemble (BME) when pursuing the Holdout and selection approach, ranking first among all tested methods. The final proposed approach is a combination of the two previous ones. It combines the best forecasting models fitted using each model’s optimal holdout selection. The accuracy of the ensemble is dramatically improved, leaving the individual learning algorithms at a reasonable distance.

The results show that some models exhibit better performance when compared with the ensemble model (BME). For instance, the average error of the TBATS model across the three holdout periods is smaller than that of the BME. Table 4.5 presents the results aggregated across all countries, with individual countries’ results available as supplementary material in a Mendeley dataset (Ashofteh et al. 2021c). The results in Table 4.5 show that the accuracy of the BME approach improves when pursuing the selection approach for each holdout, with the composite model now ranking first among all tested methods. The first row of models shows the Bayesian Model Ensemble (BME) with the SMAPE equal to 0.112 with fix holdout 3 for all models, which is the classical approach. In the same row, the second strategy with model selection shows improvement in SMAPE (0.103) for the same holdout. The final proposed approach, named “model selection plus dynamic holdouts”, which is a combination of the two previous ones, also shows better SMAPE, equal to

0.102. It combines the best forecasting models fitted using each model's optimal holdout selection, and the accuracy of the ensemble is improved, leaving the individual learning algorithms at a reasonable distance.

TABLE 4.5. RANKING THE MODELS AND ENSEMBLES ACCORDING TO THE ACCURACY MEASURE.

| Models | The model's error (SMAPE) in average | | | | | | | | (3)Model selection; dynamic holdouts | Total error of Models | Rank |
|--------|--------------------------------------|-------|-------|---------|---------------------------|-------|-------|---------|--|-----------------------|------|
| | (1) Only holdout | | | | (2) Holdout and selection | | | | | | |
| | ho=3 | ho=5 | ho=7 | Average | ho=3 | ho=5 | ho=7 | Average | | | |
| | | | | | | | | | | | |
| BME | 0.112 | 0.181 | 0.191 | 0.161 | 0.103 | 0.125 | 0.136 | 0.121 | 0.102 | 0.128 | 1 |
| TBATS | 0.120 | 0.150 | 0.172 | 0.147 | 0.114 | 0.143 | 0.177 | 0.145 | 0.119 | 0.137 | 2 |
| ETS | 0.125 | 0.200 | 0.185 | 0.170 | 0.110 | 0.138 | 0.158 | 0.135 | 0.117 | 0.141 | 3 |
| ARIMA | 0.133 | 0.178 | 0.214 | 0.175 | 0.107 | 0.145 | 0.166 | 0.139 | 0.114 | 0.143 | 4 |
| SNAIVE | 0.124 | 0.181 | 0.212 | 0.172 | 0.114 | 0.142 | 0.164 | 0.140 | 0.121 | 0.144 | 5 |
| STL | 0.117 | 0.180 | 0.201 | 0.166 | 0.118 | 0.155 | 0.169 | 0.147 | 0.121 | 0.145 | 6 |
| NNETAR | 0.141 | 0.194 | 0.210 | 0.182 | 0.106 | 0.150 | 0.181 | 0.146 | 0.106 | 0.145 | 7 |
| HWA | 0.134 | 0.193 | 0.222 | 0.183 | 0.117 | 0.154 | 0.179 | 0.150 | 0.128 | 0.154 | 8 |
| MLP | 0.130 | 0.220 | 0.240 | 0.197 | 0.123 | 0.140 | 0.169 | 0.144 | 0.123 | 0.155 | 9 |
| HWM | 0.148 | 0.195 | 0.256 | 0.200 | 0.124 | 0.157 | 0.156 | 0.146 | 0.128 | 0.158 | 10 |
| ELM | 0.139 | 0.227 | 0.242 | 0.203 | 0.114 | 0.150 | 0.203 | 0.156 | 0.122 | 0.16 | 11 |
| SSA | 0.160 | 0.190 | 0.231 | 0.194 | 0.136 | 0.168 | 0.188 | 0.164 | 0.139 | 0.166 | 12 |
| RWF | 0.153 | 0.289 | 0.362 | 0.268 | 0.111 | 0.141 | 0.184 | 0.145 | 0.123 | 0.179 | 13 |

Source: Author's preparation.

Figure 4-3 summarizes the empirical results showing that the performance of the proposed ensemble model with a new layered learning approach exhibits the highest predictive accuracy when compared to both the single forecasting methods used and the ensemble strategies considering fixed holdouts and fixed holdout with model selection. It shows that the proposed approach improves the predictive performance at each step of the learning process illustrated in Figure 4-2.

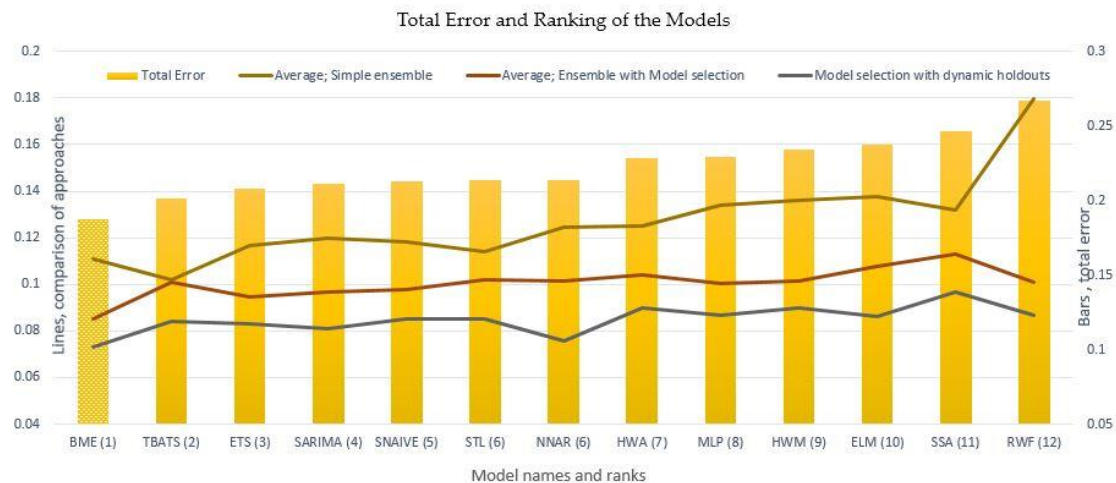


FIGURE 4-3 - COMPARING THE ACCURACY OF THE MODELS.

4.3.2.2 Model excluded in model selection

In Table 4.6, reports the distribution of the models excluded in the model selection procedure and ranks them according to their contribution to the composite model. The vertical comparison of the results gives us insight into the contribution of the different models to the ensemble, while the horizontal comparison is useful to assess the rate of contribution across different holdout periods.

The results show, first, that all models are excluded several times from the BME model space as a result of the model confidence set selection procedure, highlighting that the set of best-performing forecasters differs between countries, i.e., their predictive accuracy is population and period-specific. This is not surprising and can be explained by the differential patterns observed in respiratory disease data. The variability in the model's out-of-sample forecasting accuracy also reveals its ability to capture diverse features of mortality data. Second, the results suggest that combining models is a way to leverage their strengths and minimize their downsides. The results on the contribution of single forecasters to the composite model show that the best contributor – the ETS model – has an exclusion rate substantially smaller than that of the worst forecaster, the RWF model. Moreover, the results suggest that increasing the holdout has a slightly positive effect on some models (e.g., the ETS, SNAIVE, NNAR, MLP, and RWF models), and a negative effect on other (e.g., the SARIMA, HWA, ELM, and SSA methods), and in others a neutral effect (e.g., TBATS, STL and HWM). This variation in the contribution rates

from the best model to the worst one and from the lowest holdout period to the highest one suggests a potentially positive effect on the final forecasting accuracy of the ensemble model by selecting the best holdout for each model along with selecting the best forecasters to the model confidence set finally used to forecast

TABLE 4.6. CONTRIBUTION RATE OF THE MODELS IN THE ENSEMBLE.

| Models | The model's exclusion frequency | | | | | | | | Rank |
|--------|--|--------|-------|--------|-------|--------|-------|--------|------|
| | (2)In model selection layer for each holdout | | | | | | | | |
| | ho=3 | | ho=5 | | ho=7 | | Ave. | | |
| | Freq. | Prop. | Freq. | Prop. | Freq. | Prop. | Freq. | Prop. | |
| ETS | 10 | 4.61% | 8 | 3.56% | 8 | 4.30% | 9 | 4.31% | 1 |
| TBATS | 12 | 5.53% | 13 | 5.78% | 9 | 4.84% | 11 | 5.26% | 2 |
| STL | 13 | 5.99% | 11 | 4.89% | 11 | 5.91% | 12 | 5.74% | 3 |
| ARIMA | 13 | 5.99% | 13 | 5.78% | 14 | 7.53% | 13 | 6.22% | 4 |
| SNAIVE | 18 | 8.29% | 13 | 5.78% | 14 | 7.53% | 15 | 7.18% | 5 |
| HWA | 13 | 5.99% | 19 | 8.44% | 17 | 9.14% | 16 | 7.66% | 6 |
| HWM | 19 | 8.76% | 17 | 7.56% | 17 | 9.14% | 18 | 8.61% | 7 |
| NNETAR | 23 | 10.60% | 21 | 9.33% | 13 | 6.99% | 19 | 9.09% | 8 |
| MLP | 22 | 10.14% | 24 | 10.67% | 14 | 7.53% | 20 | 9.57% | 9 |
| ELM | 17 | 7.83% | 28 | 12.44% | 18 | 9.68% | 21 | 10.05% | 10 |
| SSA | 27 | 12.44% | 21 | 9.33% | 18 | 9.68% | 22 | 10.53% | 11 |
| RWF | 30 | 13.82% | 37 | 16.44% | 33 | 17.74% | 33 | 15.79% | 12 |

Source: Author's preparation.

Table 4.7 presents the contribution ranks, the exclusion frequency, and the proportion of the selected models with the best holdout for the BME approach with dynamic holdouts. The results show that the contribution of single learners to the ensemble changes when compared with that obtained with model selection only (Table 4.6), highlighting again the importance of combining model selecting with holdout period calibration.

TABLE 4.7. THE MODEL'S EXCLUSION FREQUENCY FOR THE ENSEMBLE WITH DYNAMIC HOLDOUTS.

| | TBATS | STL | ETS | HWA | ARIMA | SNAIVE | HWM | ELM | MLP | SSA | NNETAR | RWF |
|-------------------|-------|-----|-----|-----|-------|--------|-----|-----|-----|-----|--------|-----|
| Frequency | 13 | 15 | 17 | 17 | 18 | 19 | 19 | 21 | 23 | 25 | 29 | 37 |
| Proportion | 5% | 6% | 7% | 7% | 7% | 8% | 8% | 8% | 9% | 10% | 11% | 14% |
| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

Source: Author's preparation.

Chapter IV – New Ensemble Learning Strategy for Panel Time-Series Forecasting

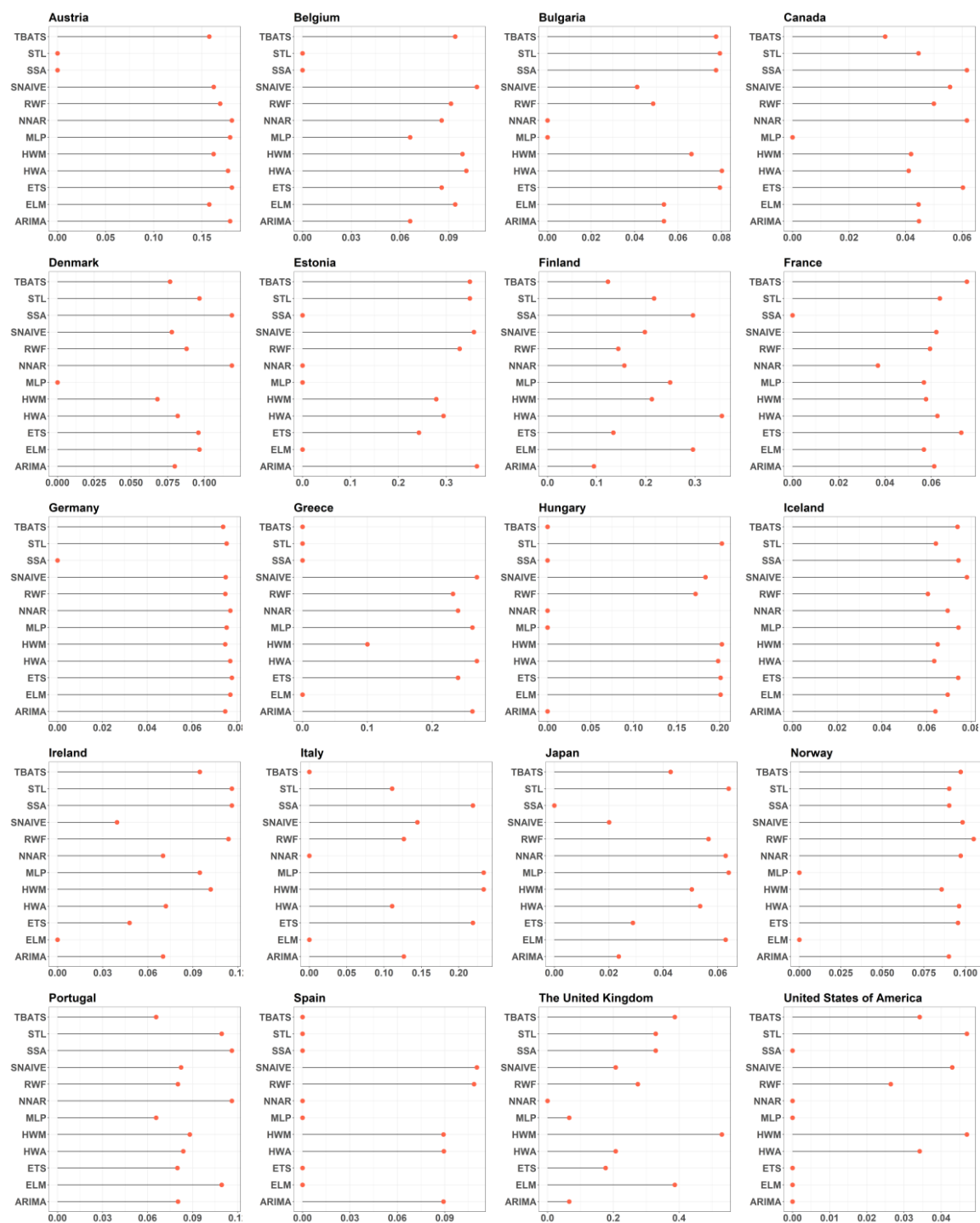


FIGURE 4-4 - BME MODEL CONFIDENCE SET AND ESTIMATED WEIGHTS PER COUNTRY.

Figure 4-4 reports the BME model confidence set (vertical axis) and corresponding posterior probability (horizontal axis) for selected countries. As we used SMAPE criteria to select the set of models and respective weights, the given zero weight indicates excluding that individual model from the BME

forecast combination. We can observe that the model's contribution in the ensemble varies among countries and the ensemble model consistently performs well in all countries.

4.3.2.3 *Algorithmic efficiency analysis*

We analyze the algorithmic efficiency of each method, i.e., the amount of computational resources used by the algorithm, by measuring the time spent in fitting the ensemble model with each approach and using it to predict the maximum likely run-time of a new given time series (Table 4.8).

TABLE 4.8. THE METHODOLOGY EFFECT ON THE RUN-TIME AND COMPUTATIONAL EFFICIENCY.

| Models | Run-time analysis of obtaining ensemble model (in minute) | | | | | | | | | | | |
|--------|---|------|------|------|-----------------------|------|------|------|------------------------------------|------|------|------|
| | Only holdout | | | | Holdout and selection | | | | Model selection & dynamic holdouts | | | |
| | ho=3 | ho=5 | ho=7 | Ave. | ho=3 | ho=5 | ho=7 | Ave. | ho=3 | ho=5 | ho=7 | Ave. |
| ART | 2.97 | 2.86 | 2.39 | 2.74 | 3.03 | 2.65 | 2.41 | 2.70 | 3.29 | 2.96 | 2.64 | 2.96 |
| STD | 0.72 | 0.72 | 0.52 | 0.65 | 0.70 | 0.60 | 0.54 | 0.61 | 0.84 | 0.71 | 0.70 | 0.75 |
| LCL | 2.79 | 2.68 | 2.26 | 2.58 | 2.85 | 2.50 | 2.27 | 2.54 | 3.08 | 2.78 | 2.46 | 2.77 |
| UCL | 3.15 | 3.04 | 2.52 | 2.90 | 3.21 | 2.80 | 2.55 | 2.85 | 3.50 | 3.14 | 2.82 | 3.15 |

Source: Author's preparation. Notes: ART: Average run-time, STD: Standard deviation, LCL: Lower confidence limit, UCL: Upper confidence limit.

The proposed method fits the models considering the three holdout periods to select the best holdout for each model. As a result, we expect that it drives the run-time at least three times more than the two other approaches. This is expected since the underlying model is a multi-step forecasting method. However, if we look at the average of run-time and their mean confidence intervals for the three approaches, we could see that they are not significantly different. It shows that our proposed method is efficient in terms of computation time.

4.3.2.4 *Excess mortality analysis*

The proposed ensemble learning for panel time-series with selecting strategy and dynamic holdouts (discussed in Section 4.2 and in the preceding paragraphs of section 4.3) was used to forecast the number of deaths caused by different kinds of respiratory diseases for a subset of 61 countries in 2020. Additionally, the COVID-19 deaths were extracted for the same year from the

COVID-19 Weekly Epidemiological Update of the World Health Organization (WHO) with data as received from national authorities, as of 3 January 2021, which has a proper coverage on the whole period of 2020 (World Health Organization, 2021).

TABLE 4.9 - COMPARISON BETWEEN FORECASTING DEATHS FOR RESPIRATORY DISEASES AND ACTUAL COVID-19 DEATHS.

| Row | Country | ⁽¹⁾ Alpha-3 | Country No | Population | ⁽²⁾ RD TD | ⁽³⁾ COVID TD | STD_RD TD | STD_COVI D TD |
|-----|-------------|------------------------|------------|------------|----------------------|-------------------------|-----------|---------------|
| 1 | Armenia | ARM | 51 | 2957.728 | 83 | 2850 | -0.417 | -0.297 |
| 2 | Australia | AUS | 36 | 25203.2 | 16554 | 909 | 2.288 | -0.337 |
| 3 | Austria | AUT | 40 | 8955.108 | 234 | 6214 | -0.392 | -0.227 |
| 4 | Azerbaijan | AZE | 31 | 10047.719 | 294 | 2703 | -0.382 | -0.3 |
| 5 | Bahamas | BHS | 44 | 389.486 | 21 | 170 | -0.427 | -0.352 |
| 6 | Belarus | BLR | 112 | 9452.409 | 205 | 153 | -0.397 | -0.353 |
| 7 | Belgium | BEL | 56 | 11539.326 | 1571 | 19693 | -0.172 | 0.052 |
| 8 | Bulgaria | BGR | 100 | 7000.117 | 412 | 7644 | -0.363 | -0.198 |
| 9 | Canada | CAN | 124 | 37411.038 | 1766 | 15679 | -0.14 | -0.031 |
| 10 | Chile | CHL | 152 | 18952.035 | 992 | 16724 | -0.268 | -0.01 |
| 11 | Costa Rica | CRI | 188 | 5047.561 | 170 | 2185 | -0.403 | -0.311 |
| 12 | Croatia | HRV | 191 | 4130.299 | 68 | 4072 | -0.419 | -0.272 |
| 13 | Cuba | CUB | 192 | 11333.484 | 1689 | 146 | -0.153 | -0.353 |
| 14 | Cyprus | CYP | 196 | 1198.574 | 19 | 129 | -0.427 | -0.353 |
| 15 | Czechia | CZE | 203 | 10689.213 | 738 | 11960 | -0.309 | -0.108 |
| 16 | Denmark | DNK | 208 | 5771.877 | 595 | 1345 | -0.333 | -0.328 |
| 17 | Egypt | EGY | 818 | 100388.076 | 4626 | 7741 | 0.329 | -0.196 |
| 18 | El Salvador | SLV | 222 | 6453.55 | 452 | 1351 | -0.356 | -0.328 |
| 19 | Estonia | EST | 233 | 1325.649 | 68 | 244 | -0.419 | -0.351 |
| 20 | Finland | FIN | 246 | 5532.159 | 53 | 561 | -0.422 | -0.344 |
| 21 | France | FRA | 250 | 65129.731 | 4733 | 64543 | 0.347 | 0.98 |
| 22 | Germany | DEU | 276 | 83517.046 | 5815 | 34272 | 0.524 | 0.354 |
| 23 | Greece | GRC | 300 | 10473.452 | 2000 | 4921 | -0.102 | -0.254 |
| 24 | Guatemala | GTM | 320 | 17581.476 | 1726 | 4827 | -0.147 | -0.256 |
| 25 | Hungary | HUN | 348 | 9684.68 | 344 | 9884 | -0.374 | -0.151 |
| 26 | Iceland | ISL | 352 | 339.037 | 17 | 29 | -0.428 | -0.355 |
| 27 | Ireland | IRL | 372 | 4882.498 | 316 | 2252 | -0.379 | -0.309 |
| 28 | Italy | ITA | 380 | 60550.092 | 4792 | 74985 | 0.356 | 1.196 |
| 29 | Japan | JPN | 392 | 126860.299 | 39818 | 3548 | 6.107 | -0.282 |
| 30 | Kuwait | KWT | 414 | 4207.077 | 291 | 937 | -0.383 | -0.337 |
| 31 | Kyrgyzstan | KGZ | 417 | 6415.851 | 253 | 1359 | -0.389 | -0.328 |
| 32 | Latvia | LVA | 428 | 1906.74 | 103 | 668 | -0.414 | -0.342 |

| | | | | | | | | |
|----|-------------------|-----|-----|------------|-------|--------|--------|--------|
| 33 | Lithuania | LTU | 440 | 2759.631 | 185 | 1644 | -0.4 | -0.322 |
| 34 | Maldives | MDV | 462 | 530.957 | 5 | 48 | -0.43 | -0.355 |
| 35 | Malta | MLT | 470 | 440.377 | 39 | 220 | -0.424 | -0.351 |
| 36 | Mauritius | MUS | 480 | 1269.67 | 82 | 10 | -0.417 | -0.356 |
| 37 | Mexico | MEX | 484 | 127575.529 | 5956 | 126507 | 0.547 | 2.263 |
| 38 | Montenegro | MNE | 499 | 627.988 | 12 | 690 | -0.428 | -0.342 |
| 39 | Netherlands | NLD | 528 | 17097.123 | 1206 | 11565 | -0.232 | -0.117 |
| 40 | New Zealand | NZL | 554 | 4783.062 | 232 | 25 | -0.392 | -0.355 |
| 41 | North Macedonia | MKD | 807 | 2083.458 | 36 | 2522 | -0.425 | -0.304 |
| 42 | Norway | NOR | 578 | 5378.859 | 528 | 436 | -0.344 | -0.347 |
| 43 | Philippines | PHL | 608 | 108116.622 | 15580 | 9253 | 2.128 | -0.164 |
| 44 | Poland | POL | 616 | 37887.771 | 5347 | 29119 | 0.448 | 0.247 |
| 45 | Portugal | PRT | 620 | 10226.178 | 2097 | 7045 | -0.086 | -0.21 |
| 46 | Qatar | QAT | 634 | 2832.071 | 12 | 245 | -0.428 | -0.351 |
| 47 | Republic of Korea | KOR | 410 | 51225.321 | 3712 | 962 | 0.179 | -0.336 |
| 48 | Rep. of Moldova | MDA | 498 | 4043.258 | 221 | 3020 | -0.394 | -0.293 |
| 49 | Romania | ROU | 642 | 19364.558 | 1484 | 15919 | -0.187 | -0.026 |
| 50 | Serbia | SRB | 688 | 8772.228 | 419 | 3288 | -0.362 | -0.288 |
| 51 | Singapore | SGP | 702 | 5804.343 | 906 | 29 | -0.282 | -0.355 |
| 52 | Slovakia | SVK | 703 | 5457.012 | 476 | 2317 | -0.352 | -0.308 |
| 53 | Slovenia | SVN | 705 | 2078.654 | 145 | 2889 | -0.407 | -0.296 |
| 54 | Spain | ESP | 724 | 46736.782 | 3042 | 50442 | 0.069 | 0.688 |
| 55 | Suriname | SUR | 740 | 581.363 | 39 | 123 | -0.424 | -0.353 |
| 56 | Sweden | SWE | 752 | 10036.391 | 665 | 8727 | -0.321 | -0.175 |
| 57 | Switzerland | CHE | 756 | 8591.361 | 428 | 7049 | -0.36 | -0.21 |
| 58 | The UK | GBR | 826 | 67530.161 | 6943 | 74570 | 0.71 | 1.188 |
| 59 | Turkey | TUR | 792 | 83429.607 | 1658 | 21295 | -0.158 | 0.085 |
| 60 | Ukraine | UKR | 804 | 43993.643 | 1089 | 18854 | -0.252 | 0.034 |
| 61 | US of America | USA | 840 | 329064.917 | 16554 | 345253 | 2.288 | 6.791 |

Source: Author's preparation. Notes: (1) Abbreviation code of the country (Three letters); (2) Respiratory diseases deaths; (3) WHO COVID-19 deaths.

Table 4.9 represents the forecasting of the total deaths for respiratory diseases (RD_TD), which is concluded as an aggregation of monthly forecasting of death for each country. The last two columns show the standardized values of total respiratory disease deaths and COVID-19 deaths to be used for calculating the

correlation. The Pearson correlation for all 61 countries is 0.34, which is statistically significant (P-value = 0.007). As it is shown in Table 4.9, we considered the European countries, Canada, the United States of America, and the United Kingdom from the list to calculate the correlation. The selection criteria was related to the official statistics maturity (SDDS+, SDDS, GDDS), the models of corruption in official statistics (Georgiou, 2021), and quality level of deaths data according to the WHO ranking discussed in section 4.1.

The correlation is increased dramatically to 0.94 (P-value =0.000). It could be because of a higher quality of the Official Statistics in these countries as Ashofteh and Bravo (2020) showed that there is significant variation in the quality of COVID-19 datasets reported worldwide. A recent study suggests that data science and new technologies are expected to play a significant role in improving data quality at National Statistical Offices in the future (Ashofteh and Bravo, 2021b).

The comparison of respiratory diseases and Covid-19 deaths are shown in Figures 4.5 and 4.6.

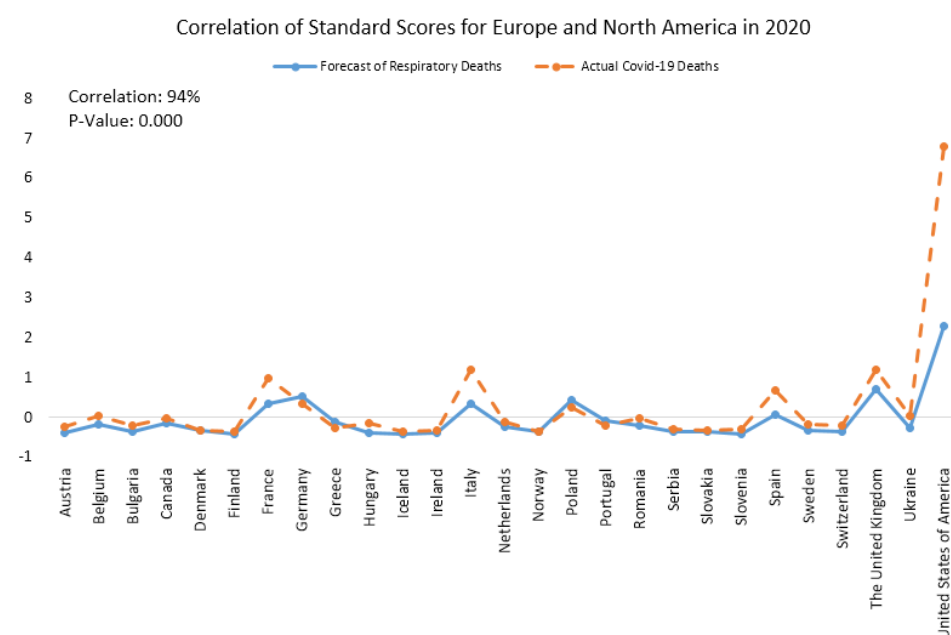


FIGURE 4-5 - RESPIRATORY DISEASES DEATHS AND COVID-19 DEATHS FOR EUROPE AND NORTH AMERICA IN 2020.

For most countries, the reported COVID-19 deaths have somehow “replaced” the otherwise respiratory deaths that would have anyway occurred based on extrapolating past respiratory disease trends. Concerning the factors affecting

COVID-19 mortality, our research results show a high correlation between respiratory deaths and COVID-19 deaths, which are consistent with clinical manifestations and epidemiological studies. For example, compared with other countries, countries with a high expectancy of respiratory diseases had a higher excess mortality, which is at the macro country level. At the individual level, the higher number of deaths for respiratory diseases could be considered as more susceptibility of population to COVID-19 symptoms, and the greater risk of death. This study shows the comparison of the different countries and their policies' effectiveness could cause an evaluation bias without considering their backgrounds to respiratory diseases.

Figure 4.5 shows that European countries and North America were sensitive to respiratory diseases and it boosted their excess mortality caused by the COVID-19 pandemic, however, Figure 4-6 shows that some countries have dealt with COVID-19 in 2020 better than others in respect to their vulnerability to respiratory diseases. It shows that in countries in which the forecast of respiratory diseases significantly exceeds the confirmed COVID-19 deaths (e.g., Japan and the Philippines, Figure 4-6) the management of the pandemic crisis succeeded in reducing excess mortality. The results from these two Figures align with a recent study that indicated a much lower overall excess-mortality burden due to COVID-19 in Japan than in Europe and the USA (Kawashima et al. 2020). To describe the possible reason, Yorifuji, et al. (2021) suggest that in Japan, the public health regulations aimed at preventing COVID-19 may incidentally reduce mortality related to respiratory diseases such as influenza, and it decreased the net excess mortality.

Additionally, in response to the vulnerability to respiratory diseases in various countries, Japan and the Philippines have provided a good example for the rest of the world in terms of controlling the positive effect of respiratory death numbers on the COVID-19 deaths. This similar situation of these two countries might testify to the importance of Philippine and Japan agreements on COVID-19 response support. As it is reported on the website of the department of foreign affairs of Philippines, the Japanese Government has been unstinting in its commitment to the Philippines' recovery efforts, previously pledging

over JPY100 billion assistance in emergency and standby loans and the recent donation of 1 million Japan-manufactured AstraZeneca vaccines¹².

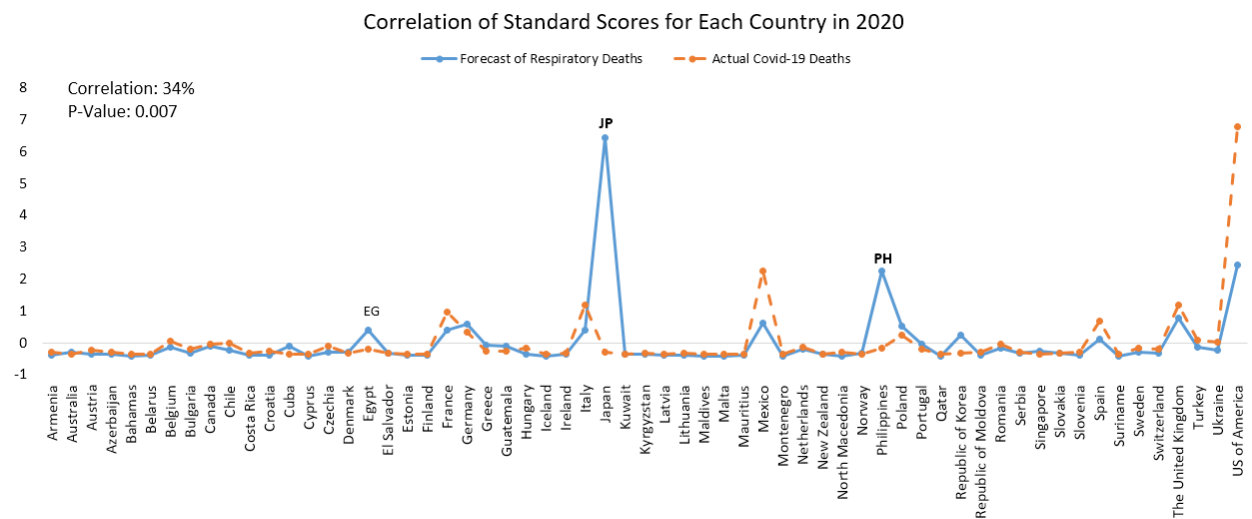


FIGURE 4-6 - RESPIRATORY DISEASES DEATHS AND COVID-19 DEATHS FOR EACH COUNTRY IN 2020.

Figure 4-6 shows the same situation for the Republic of Korea, which is geographically in the vicinity of these two countries. The result from comparison of the forecast respiratory deaths and actual COVID-19 deaths for the Republic of Korea is align with the results of very recent study about estimation of excess mortality in Korea by Shin et. al., (2021), which shows that the mortality in 2020 was similar to the historical trend. This similarity among neighbours might clarifies the importance of international cooperation and sharing the resources for successfully controlling the effects of pandemics. Besides, as these countries are geographically close to each other, meteorological factors could be also influential in this similarity, and it requires more in-depth research.

As a final result, in addition to respiratory deaths effect on the pandemic's deaths, international cooperation, optimal scheduling and utilization of medical resources, large-scale virus testing, protect and manage the health care of the elderly, lockdowns, vaccination, and controlling the borders are

¹² <https://dfa.gov.ph/dfa-news/dfa-releasesupdate/29206-philippines-and-japan-sign-agreements-on-covid-19-response-support-and-on-scholarship-grants-for-civil-servants>

examples of other factors, which may lead to different conclusions for different countries. However, accurate and timely estimation of respiratory deaths also seems an important factor to be considered in the comparison of multiple countries.

4.4 DISCUSSION AND CONCLUSION

Our goal was to obtain a benchmark to evaluate the excess of mortality derived from COVID-19 as a common framework for all countries. In this sense, we investigated a new technique of ensemble learning for panel time series forecasting of respiratory diseases and we summarized the empirical results from individual models, from a simple ensemble model, from an ensemble with model selection, and from an ensemble with model selection and dynamic holdouts.

According to the performance of the models, presented in table 4.5, the ensemble models provide better performance than the remaining methods on average. Table 4.5 provided clear evidence on the competitiveness of our method in terms of predictive performance when compared to the state of the arts and even the ensemble model without the holdout and model selection layer.

In terms of the candidate models to contribute to the ensemble, Tables 4.6 and 4.7 show the positive effect on prediction accuracy by selecting the best holdout for each model and removing the outlier models from the ensemble. Additionally, one can see that some of the state-of-the-art approaches overperformed the neural networks' time series models. The possible reason for the underperforming of the complex neural network approaches could be the non-stationary elements such as trend component. These models are known not to work well with the trend data, which is exactly the case for most of our time series. However, neural networks' time series models are proven to perform well when the time series data are nonlinear, stationary, and have sudden changes through the layering hierarchy. Therefore, we expect them to add value to the Ensemble for mostly de-trended time series. Additionally, recurrent neural networks like LSTM and GRU have the potential to

outperform time series models and they could be further explored for the ensemble in future studies.

The variation in the performance for each model shows the necessity to improve each of them separately by choosing the best holdout, and further try to find the best models to contribute to the ensemble without overfitting. The proposed indicator in Formula (4-3) removes only the models that are very far from the others, to avoid the significant bias in the set of candidate forecasters. The final ensemble model shows a significant improvement in the accuracy in comparison with the other ensembles and each individual state-of-arts.

We use the new ensemble strategy to forecast the number of deaths from respiratory diseases in 2020, for a sample of 61 countries. The correlation between the standardized values of the respiratory diseases' deaths and the COVID-19 deaths were positive and statistically significant. It recommends we consider the forecasted values of the respiratory diseases as a covariate to evaluate the effective strategies of different countries, such as lockdown rules or relaxing of border control regulations. Japan and the Philippines are candidates with our study for more investigation in this regard, and they are more eligible than other countries with only a low death toll. It could be possible that the experience of these countries with high mortality caused by respiratory diseases has played a relevant role in managing the pandemic.

It could be relevant in this pandemic to focus more on the death toll than the cumulative number of patients. According to the nature of pandemics, it is challenging to control its spread; however, the main concern could be controlling the severe cases and the patients with a high likelihood of death. These countries with a high number of respiratory diseases that could manage the pandemic reasonably could be more recommendable for further studies on their policies and health strategies in comparison with the countries with only a low rate of mortality.

To sum up, this chapter describes an initial attempt at proposing a new approach for ensemble forecasting tasks. The main motivation of this study was the observation that the performance of the ensemble model has the potential of improvement based on choosing the best holdout for each candidate model and choosing the best outcomes based on the dynamics of the observed values of the main series. In experiments using the 61 time series from

respiratory diseases death from 2000 to 2016, aggregating some selected forecasting models using our approach provides a consistent advantage in terms of accuracy and leads to better predictive performance. Furthermore, this study offers a correction on the total number of positive cases of COVID-19, according to the expected number of deaths caused by respiratory diseases as the result of our ensemble model.

Finally, this study highlighted the situation of Japan and the Philippines as two candidates for further studies. These two countries are in a category with high vulnerability to this pandemic; however, they could manage it well. As a result, they are recommended as the best practices by this study regardless of their higher death toll in comparison with some other countries. Additionally, it is interesting to see that for most countries the reported COVID deaths have actually sort of “replaced” the otherwise respiratory deaths that would have anyway occurred, based on extrapolating the past trends of respiratory deaths.

■

CHAPTER FIVE - LIFE TABLE FORECASTING IN COVID-19 TIMES: AN ENSEMBLE LEARNING APPROACH

In this Chapter, we will use the new ensemble time series model that we developed in the previous Chapter. Stochastic mortality modeling play a critical role in public pension design, population, and public health projections and the design, pricing, and risk management of life insurance contracts and longevity-linked securities. There is no general method to forecast mortality rates applicable to all situations especially for unusual years such as the COVID-19 pandemic. In this Chapter, we investigate the feasibility of using an ensemble of traditional and machine learning time series methods to empower forecasts of age-specific mortality rates for groups of countries that share common longevity trends. We use Generalized Age-Period-Cohort stochastic mortality models to capture age and period effects, apply K-means clustering to time series to group countries following common longevity trends, and use ensemble learning to forecast future longevity and annuity price markers. To calibrate models, we use data for 14 European countries from 1960 to 2018. The results show that the ensemble method presents the best robust results overall with minimum RMSE in the presence of structural changes in the shape of time series at the time of COVID-19¹³.

¹³ Please cite this chapter as: Ashofteh, A., & Bravo, J. M. (2021). Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach. 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), 2021, pp. 1-6,

DOI: <https://doi.org/10.23919/CISTI52073.2021.9476583>

5.1 INTRODUCTION

Public and private pension schemes and life insurance companies offer individuals an ex-ante efficient risk pooling mechanism that addresses the (individual) uncertainty of death through the delivery of a lifetime annuity, redistributing income in a welfare-enhancing manner (Alho et al., 2012; Ashofteh & Bravo, 2021; Ayuso et al., 2021a,b; Bravo & Herce, 2020; Bravo, 2016, 2019, 2021). Annuity providers and life settlement investors face long-run solvency challenges to provide guaranteed lifetime income due to uncertain financial returns and systematic (non-diversifiable) longevity risk (Simões et al., 2021). Longevity risk management solutions include product re-design, risk-sharing arrangements between pensioners/policyholders and providers, natural hedging, liability selling via an insurance or reinsurance contract (pension buy-outs/ins, bulk annuity transfers), and, more recently, capital-market-based solutions (e.g., CAT mortality bonds, survivor/longevity bonds, Index-based longevity swaps q-forwards, S-forwards, longevity options) (Bravo, 2020; Bravo & Nunes, 2021; Bravo, 2021; Bravo & El Mekkaoui de Freitas, 2018; Bravo & Pereira da Silva, 2006). Stochastic mortality models play a critical role in the design, pricing, and risk management of life insurance contracts and longevity-linked securities, in public pension design (e.g., through automatic indexation of pension age to life expectancy) and in population and public health projections (J. Bravo & Coelho, 2019). The traditional approach to age specific mortality rate forecasting is to use a single believed to be best model selected from a set of candidates using some method or criteria (e.g., BIC, AIC), often neglecting model risk for statistical inference purposes. To this end, in the actuarial, financial and demographic literature, several single and multi-population discrete-time and continuous-time stochastic mortality models have been proposed (see, e.g., Brouhns, Denuit, & Vermunt, 2002; Denuit & Goderniaux, 2005; Hunt & Blake, 2021; Hyndman & Ullah, 2007; Lee & Carter, 1992; and references therein). A recent strand of literature involves the use of an adaptive Bayesian Model Ensemble (BME) of heterogeneous methods. The procedure involves both the selection of the subset of superior models and the determination of optimal weights (Bravo et al., 2021a,b; Ayuso et al., 2021b; Bravo & Ayuso, 2020, 2021; Breiman, 1996; Wiśniowski et al., 2015). The ongoing COVID-19 pandemic outbreak

highlighted the importance of analyzing the impact of adverse mortality and morbidity shocks on insured annuity and pension scheme portfolios. If in the early stages of the pandemic the disease has predominantly impacted mortality at high ages, it remains to be seen whether the catastrophic mortality event has the potential to permanently affect human longevity prospects at all ages, in a uniform or heterogeneous way across socioeconomic groups, either directly or indirectly, or if it is just a temporary shock that accelerated the termination of the life of certain groups with pre-existing significant co-morbidities (Ashofteh & Bravo, 2020). This Chapter expands previous research by exploring the use of ensemble learning for multi-population age-specific mortality forecasting, life table construction, and annuity pricing. We first use a Generalized Age-Period-Cohort (GAPC) stochastic mortality model accounting for age and period effects to fit the data to individual countries. Second, we apply K-means clustering for time series to group countries according to similar longevity trends. Third, we forecast age- and country-specific mortality rates by using a BME of traditional and machine learning heterogeneous models comprising ARIMA models, Multilayer Perceptron (MLP), and Singular spectrum analysis (SSA). Fourth, we use mortality forecasts to compute life expectancy measures and life annuity prices. We evaluate the model's performance pre-and post-COVID-19 and discuss the pandemic implications for the insurance industry. The datasets used in this study comprise mortality data (deaths and exposure to risk) for 14 European countries (Austria, Belgium, Denmark, Finland, France, Germany, Italy, Luxembourg, Netherlands, Norway, Portugal, and Sweden), USA, Canada, Australia, and Japan from 1960 to 2018. Our results suggest that the proposed ensemble time series model performs better than merely extrapolating the mortality patterns observed in each age interval. The remaining of the Chapter is organized as follows. Section 2 outlines the key concepts and methods used in the Chapter. Section 3 reports summary results for the forecasted pension age together with the reference period (and cohort) life expectancy measures, and critically discusses the results. Finally, conclusions are presented in section 4.

5.2 MATERIALS AND METHODS

5.2.1 GAPC stochastic mortality models

Generalised Age-Period-Cohort mortality models are a class of parametric models that link a response variable with a linear or bilinear predictor structure consisting of a series of factors dependent on age of the individual, x ; calendar effects, t ; and year of birth effects, $c = t - x$. GAPC models fit into the general class of generalized nonlinear models (GNM), with a structure that includes a random component, a systematic component, a link function, a set of parameter constraints to ensure identifiability and time series methods for forecasting and simulating the period and cohort indexes. The random component specifies whether the number of deaths recorded at age x in year t , $D_{x,t}$, follows a Poisson distribution $D_{x,t} \sim P(\mu_{x,t}E_{x,t}^c)$, with $E(D_{x,t}/E_{x,t}^c) = \mu_{x,t}$, or a Binomial distribution $D_{x,t} \sim B(q_{x,t}E_{x,t}^0)$, with $E(D_{x,t}/E_{x,t}^0) = q_{x,t}$, where $E_{x,t}^0$ and $E_{x,t}^c$ denote, respectively, the population initially or centrally exposed-to-risk, and $q_{x,t}$ is the one-year death probability for an individual aged x last birthday in year t . The systematic component links a response variable to an appropriate linear predictor $\eta_{x,t}$

$$\eta_{x,t} = \alpha_x + \sum_{i=1}^N \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x}, \quad (5-1)$$

where $\exp(\alpha_x)$ denotes the general shape of the mortality schedule across age, $\beta_x^{(i)} \kappa_t^{(i)}$ is a set of N age-period terms describing the mortality trends, with each time index $\kappa_t^{(i)}$ contributing in specifying the general mortality trend and $\beta_x^{(i)}$ modulating its effect across ages, and the term $\gamma_{t-x} \equiv \gamma_c$ accounts for the cohort effect c with $\beta_x^{(0)}$ modulating its effect across ages. The period $\kappa_t^{(i)}$ and cohort γ_{t-x} indices are stochastic processes. The specification is complemented with a set of parameter constraints to ensure unique parameter estimates (Hunt & Blake, 2020). Parameter estimates are obtained using maximum-likelihood (ML) methods. For illustration, in this Chapter we use one member of the GAPC family of models, the standard age-period Lee-Carter model under a Poisson setting for the number of deaths (Brouhns et al., 2002), defined for each population as:

$$\eta_{x,t} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)}, \quad (5-2)$$

where $\exp(\alpha_x)$ denotes the general shape of the mortality schedule across age, $\beta_x^{(1)} \kappa_t^{(1)}$ is an age-period term describing the mortality trends, with the time index $\kappa_t^{(1)}$ capturing the general mortality trend and $\beta_x^{(1)}$ tempering its effect across ages. The framework assumes $D_{x,t}$ follows a Poisson distribution $D_{x,t} \sim \mathcal{P}(\mu_{x,t} E_{x,t}^c)$ with $\mathbb{E}(D_{x,t}/E_{x,t}^c) = \mu_{x,t}$ and log canonical link. To forecast age-specific mortality rates, we first calibrate model (2) to each country population data from 1960 to 2018 and for ages in the range 50 – 90. Second, to forecast mortality rates, we assume the age vectors α_x and $\beta_x^{(1)}$ in equation (2) remain constant over time and model the period index $\kappa_t^{(1)}$ using a BME of traditional univariate ARIMA (p, d, q) , Multilayer Perceptron and Singular spectrum analysis models. Before that, we apply K-means clustering for time series to each country $\kappa_t^{(1)}$ series to group countries according to similar longevity trends. Third, to close life tables at high ages $x > 90$, we use the log-quadratic model proposed in (Denuit & Goderniaux, 2005). The datasets used in this study comprise mortality data and full pension age data. Mortality data are obtained from the Human Mortality Database (“Human Mortality Database,” 2021) and consist of observed death counts, $D_{x,t}$, and exposure-to-risk, $E_{x,t}$, classified by age at death ($x = 0, \dots, 110+$), year of death ($t = 1960, \dots, 2018$) and sex.

5.2.2 K-Means Clustering of time trend indices

We use the dynamic time warping (DTW) distance and its corresponding lower bounds (LBs) for clustering the individual time series $\kappa_{t,j}^{(1)}$ representing the longevity trends observed at time t in country j , into similar groups. DTW is a technique to cluster the data points in an ordered sequence and to measure similarity between two temporal sequences that do not align exactly in time or length. Given the time series $\kappa_{t,j}^{(1)} = (\kappa_{1,j}^{(1)}, \kappa_{2,j}^{(1)}, \dots, \kappa_{n,j}^{(1)})$ for country j , and $\kappa_{t,i}^{(1)} = (\kappa_{1,i}^{(1)}, \kappa_{2,i}^{(1)}, \dots, \kappa_{n,i}^{(1)})$ for country i , the optimized DTW distance would be obtained by the following squared root of the sum of squared distances between each element in $\kappa_{t,i}^{(1)}$ and its nearest point in $\kappa_{t,j}^{(1)}$:

$$DTW(\kappa_{t,j}^{(1)}, \kappa_{t,i}^{(1)}) = \min_{\pi} \sqrt{\sum_{(i,j) \in \pi} d(\kappa_{1,j}^{(1)}, \kappa_{1,i}^{(1)})^2}, \quad (5-3)$$

where $\pi = [\pi_0, \dots, \pi_L]$ is a path satisfying the properties:

- a) It is a list of index pairs $\pi_l = (i_l, j_l)$ with $0 \leq i_l < n$ and $0 \leq j_l < m$.
- b) $\pi_0 = (0,0)$ and $\pi_L = (n-1, m-1)$.
- c) For all $k > 0$, $\pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:
 - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
 - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

5.2.3 Ensemble learning of time series

Bayesian model-ensemble or averaging is an application of Bayesian theory to model selection and inference under model uncertainty. The approach overcomes the problem of deriving conclusions based on a single assumed to be "best" model by conditioning the statistical inference on the entire ensemble of statistical models (or a subset of them) initially considered. Following (J. M. Bravo et al., 2021), let each candidate model be denoted by M_l , $l = 1, \dots, K$ representing a set of probability distributions comprising the likelihood function $L(y|\theta_l, M_l)$ of the observed data y in terms of model specific parameters θ_l and a set of prior probability densities $p(\theta_l|M_l)$. Let Δ denote a quantity of interest present in all models (e.g., the future values of y). The marginal posterior distribution across all models is

$$p(\Delta|y) = \sum_{k=1}^K p(\Delta|y, M_k)p(M_k|y), \quad (5-4)$$

where $p(\Delta|y, M_k)$ denotes the forecast PDF based on model M_k alone, and $p(M_k|y)$ is the posterior probability of model M_k given the observed data with $\sum_{k=1}^K p(M_k|y) = 1$. To compute model weights $p(M_k|y)$, for each population we first rank models according to their out-of-sample predictive accuracy and then use the normalized exponential function,

$$p(M_k|y) = \frac{\exp(-|\xi_k|)}{\sum_{l=1}^K \exp(-|\xi_l|)}, k = 1, \dots, K, \quad (5-5)$$

with $\xi_k = S_k / \max\{S_l\}_{l=1, \dots, K}$ and S_k is forecasting error for model k and population g . The normalized exponential function assigns larger weights to models with smaller forecasting error, with weights decaying exponentially. The ensemble model set used in this Chapter comprises the classical univariate $ARIMA(p, d, q)$ time series models, the Multilayer Perceptron (MLP) model and the Singular spectrum analysis (SSA) model. The MLP is a kind of NNAR-Neural Network Autoregression Model. It is more complicated and advanced than "nnetar" with three components in the form of $NNAR(p, P, k)$, in which p denotes the number of lagged values that are used as inputs and usually is chosen based on an information criterion, like AIC, P denotes the number of seasonal lags, and k denotes the number of hidden nodes.

TABLE 5.1 - SUMMARY OF THE LEARNING ALGORITHMS.

| ID | Algorithm | Parameters | Value |
|-------|---|--|--|
| ARIMA | The Auto-Regressive Integrated Moving Average | Auto | |
| MLP | Multilayer Perceptron for time series | Comb hd.auto.type hd.max | Mode Valid 5 |
| SSA | Singular spectrum analysis | Kind svd.method L neig force.decompose mask | 1d-ssa Auto 12 NULL TRUE NULL |

Source: Author's preparation.

Singular spectrum analysis (SSA) is used as one of the high-quality modeling approaches for forecasting mortality rate (Mahmoudvand et al., 2017). The calibration of the SSA is an important but not easy task in a standalone modeling approach. It depends upon two basic parameters of the window length and the number of eigentriples used for reconstruction. The choice of improper values for these parameters yields incomplete reconstruction, and the forecasting results might be misleading. In this study, we set a length equal to 12 and eigentriples equal to NULL. Whenever possible, we use the Box-Cox

transformation of the learning algorithms. The other hyper-parameters are summarized in Table 5.1.

5.2.4 Life expectancy and life annuity computation

Let ${}_t p_x(t)$ denote the τ -year survival rate of a reference population cohort aged x at time t , defined as ${}_t p_x(t) := \exp\left(-\int_0^\tau \mu_{x+s}(s)ds\right)$, where $\mu_x(t)$ is a stochastic force of mortality process. For the discretized stochastic process, we assume that $\mu_{x+\xi}(t+\varepsilon) = \mu_x(t)$ for any $0 \leq \xi, \varepsilon < 1$, from which $\mu_x(t)$ is approximated by the central death rate $m_x(t)$ and $p_x(t) = \exp(-m_x(t))$. The complete period life expectancy for an x -year old individual in year t is computed as follows

$$e_{x,g}^P(t) := \frac{1}{2} + \sum_{k=1}^{\omega-x} \exp\left(-\sum_{j=0}^{k-1} m_{x+j,g}(t)\right), \quad (5-6)$$

with ω denoting the highest attainable age. The corresponding life annuity price is computed as follows:

$$a_{x,g}^P(t) := \sum_{k=1}^{\omega-x} {}_k p_x(t) (1+y)^{-k}, \quad (5-7)$$

with y denoting the guaranteed interest rate, set at 1% in this study. Without loss of generality, we discard annuity providers default risk when pricing life annuity contracts.¹⁴

5.3 RESULTS

In Figure 5-1, we represent, as a representative case, the crude estimates of mortality rates (in log scale) for Portugal (total population) for selected ages from 1960 to 2018.

¹⁴ For a discussion of credit risk see, e.g., (Ashofteh, 2018; Ashofteh & Bravo, 2019, 2021; Chamboko & Bravo, 2019b) and references therein.

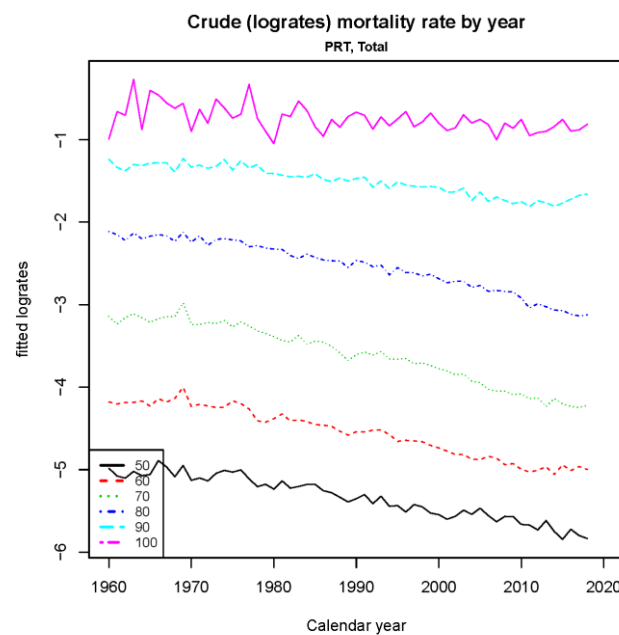


FIGURE 5-1 - CHANGES IN THE TOTAL LOG MORTALITY RATES WITH RESPECT TO BOTH AGE AND YEAR OVER THE PERIOD 1960-2018 IN PORTUGAL.

The results show a declining trend in mortality at all ages, more pronounced at younger ages, explained by improved economic, social and health conditions, changing lifestyles, developments in medical treatments and medicines. Figure 5-2 reports the Lee-Carter parameter estimates for Portugal.

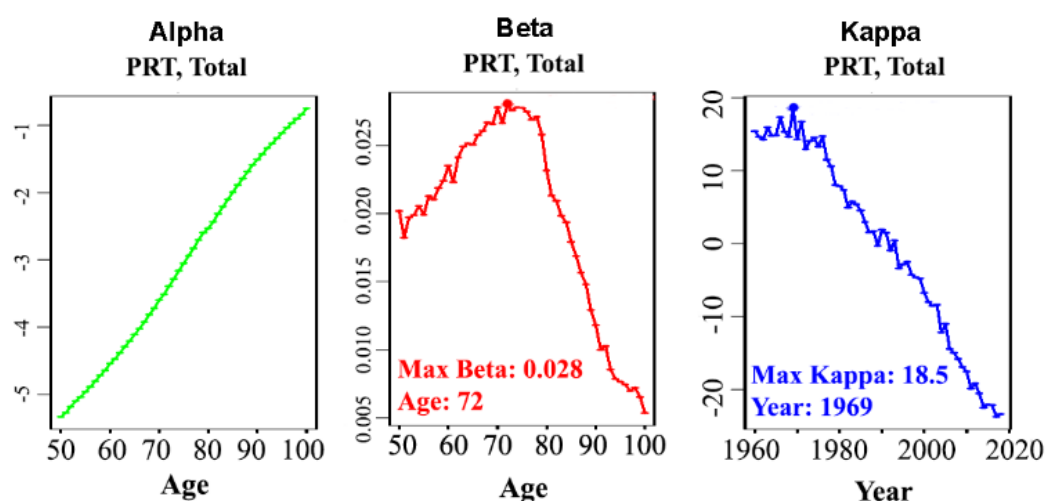


FIGURE 5-2 - THE PARAMETERS OF THE POISSON LEE-CARTER MODEL OVER THE PERIOD 1960 -2018. PORTUGAL WITH MAXIMUM BETA AT AGE 72 AND MAXIMUM KAPPA AT THE YEAR 1969.

The α_x pattern shows a normal increase in mortality rates with age, as observed in developed and developing countries. The $\beta_x^{(1)}$ estimates show that the

mortality decline observed over this period was not homogeneous across ages, with the age group 65-75 exhibiting the highest longevity gains. The time trend parameter estimates $\kappa_t^{(1)}$ for Portugal highlight an almost linear decline in mortality observed during the last six decades.

We observe a regular increasing trend in period (and cohort) life expectancy at all ages, as illustrated in Figure 5-3 at age 70. However, with COVID-19 the estimation of the life expectancy is decreasing dramatically for different ages according to Figure 5-3.

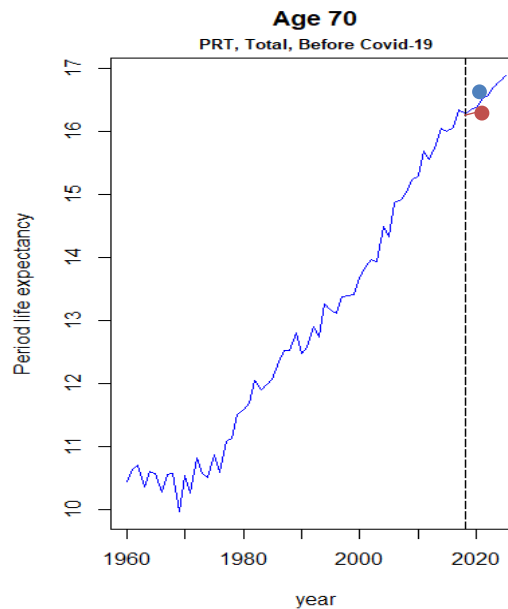


FIGURE 5-3 - LIFE EXPECTANCY/ANNUITY PRICES; BEFORE (BLUE) AND AFTER (RED) THE COVID19.

The necessity of this approach will be more prominent by referring to the recent studies on long-term consequences of COVID19 and its long-term health effects. Figure 5-4 shows the dendrogram of hierarchical cluster analysis for the time series of $\kappa_t^{(1)}$ for the countries analyzed in this study. The results suggest, for both genders, similar longevity trends in

- the United States of America and Canada;
- France, Belgium, Italy, and Austria;
- Sweden and Finland.

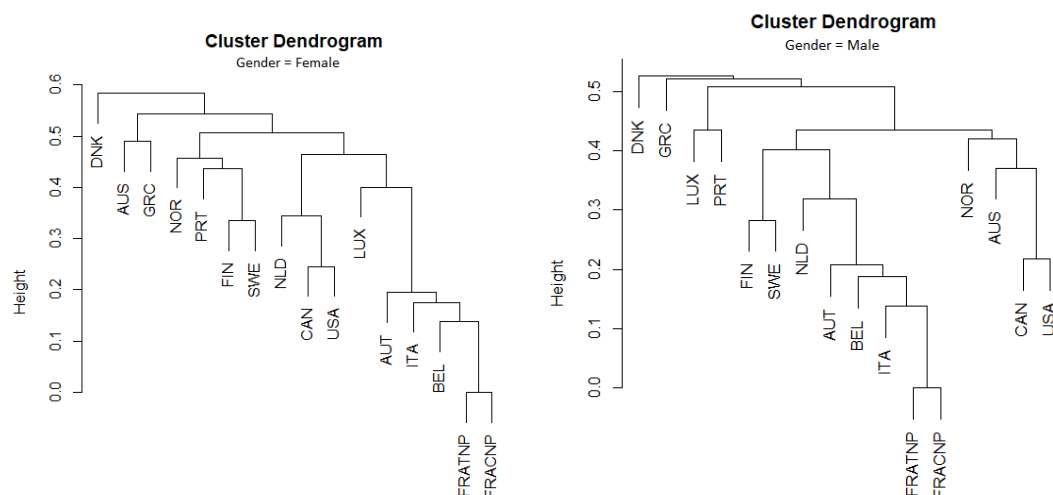


FIGURE 5-4 - CLUSTER DENDROGRAMS OF THE BASELINE TIME SERIES PARAMETERS BY COUNTRY AND GENDER.

Australia is in the same cluster as the USA in males, however for females, it falls into the same cluster as Germany. Therefore, if we choose the USA and Germany, then it is possible to ignore Canada and Australia. With the same logic, we choose Portugal, France, and Denmark. The time series for Germany was reported till 2017, and it was excluded because of incompleteness in the time series. As a result, the final list of the selected countries to cover all of the possible baseline time series of $\kappa_t^{(1)}$ is Denmark, France, Portugal, and the USA.

The cluster analysis results allowed the identification of common longevity trends among groups of countries and were then used to apply the BME approach when forecasting the time trend parameter.

The resulting prospective life tables were then used to compute: (i) estimates of the price of a life annuity paying 1 monetary unit for life; (ii) period life expectancy and (iii) the probability of death, for all ages and years. Tables 5.2 and 5.3 report representative results for all ages in the age interval 75-90 considering the pre- and post- COVID-19 mortality conditions using the BME ensemble learning approach and the traditional univariate ARIMA models.

Table 5.2 shows that the estimations based on ARIMA do not respond effectively to the changes based on COVID19. The maximum difference is for the age 75 and the differences are decreasing as the age is increased. Similar results can be found for the life expectancy and probability of death. Although the case of Portugal is considered as representative of the developed countries

for the comparison of different approaches, however, we did these comparisons on the other countries and observed similar results in almost all ages.

TABLE 5.2 - LIFE ANNUITY, LIFE EXPECTANCY AND PROBABILITY OF DEATH BASED ON ARIMA FORECASTING OF KAPPA FOR THE YEAR 2020.

| 2020 | Life annuity | | Life Expectancy | | Prob. of Death | |
|------|------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|
| AGE | ARIMA <i>Before COVID19</i> | ARIMA <i>After COVID19</i> | ARIMA <i>Before COVID19</i> | ARIMA <i>After COVID19</i> | ARIMA <i>Before COVID19</i> | ARIMA <i>After COVID19</i> |
| 75 | 11.1503 | 11.1267 | 12.5923 | 12.5652 | 0.0234 | 0.0235 |
| 76 | 10.5311 | 10.5084 | 11.8814 | 11.8555 | 0.0263 | 0.02643 |
| 77 | 9.9236 | 9.9016 | 11.1887 | 11.1639 | 0.0297 | 0.02987 |
| 78 | 9.3299 | 9.3085 | 10.5162 | 10.4922 | 0.0334 | 0.03356 |
| 79 | 8.749 | 8.7281 | 9.8625 | 9.8392 | 0.0384 | 0.03853 |
| 80 | 8.1892 | 8.1686 | 9.2362 | 9.2134 | 0.0436 | 0.04397 |
| 81 | 7.6483 | 7.6297 | 8.6346 | 8.6141 | 0.0501 | 0.05044 |
| 82 | 7.1321 | 7.1153 | 8.0636 | 8.0451 | 0.0569 | 0.05719 |
| 83 | 6.6377 | 6.6224 | 7.5195 | 7.5028 | 0.0648 | 0.06515 |
| 84 | 6.1687 | 6.1548 | 7.006 | 6.9908 | 0.0727 | 0.07299 |
| 85 | 5.7185 | 5.7057 | 6.5158 | 6.5019 | 0.0833 | 0.08367 |
| 86 | 5.3008 | 5.289 | 6.0627 | 6.0499 | 0.0947 | 0.09501 |
| 87 | 4.9137 | 4.9027 | 5.6444 | 5.6325 | 0.1066 | 0.10691 |
| 88 | 4.5549 | 4.5444 | 5.2582 | 5.2469 | 0.1198 | 0.12013 |
| 89 | 4.2267 | 4.2165 | 4.9058 | 4.895 | 0.1373 | 0.13758 |
| 90 | 3.9482 | 3.9381 | 4.6069 | 4.5961 | 0.1411 | 0.14159 |
| RMSE | 0.0171 | | 0.019 | | 0.0003 | |

Source: Author's preparation.

The same RMSE of forecasts of the mentioned parameters are computed and reported in Table 5.3. The proposed ensemble model followed the increase of the deaths caused by the pandemic effectively. It shows small changes, however, it is robust enough to not be changed dramatically and bias the risk estimation by biased forecasts of price annuities. Finally, the comparison of RMSE's of ensemble models and ARIMA models show that the ensemble approach is more accurate and robust than ARIMA approach in all considered ages for the representative countries.

TABLE 5.3 - LIFE ANNUITY, LIFE EXPECTANCY AND PROBABILITY OF DEATH BASED ON ENSEMBLE FORECASTING OF KAPPA FOR THE YEAR 2020.

| 2020 | Life annuity | | Life Expectancy | | Prob. of Death | |
|------|-----------------------------------|----------------------------------|-----------------------------------|----------------------------------|-----------------------------------|----------------------------------|
| AGE | ENS. <i>Before COVID19</i> | ENS. <i>After COVID19</i> | ENS. <i>Before COVID19</i> | ENS. <i>After COVID19</i> | ENS. <i>Before COVID19</i> | ENS. <i>After COVID19</i> |
| 75 | 11.1652 | 11.1637 | 12.6093 | 12.6076 | 0.0233 | 0.0233 |
| 76 | 10.5453 | 10.544 | 11.8976 | 11.896 | 0.0262 | 0.0262 |
| 77 | 9.9372 | 9.9359 | 11.2041 | 11.2026 | 0.0296 | 0.0296 |
| 78 | 9.3428 | 9.3415 | 10.5307 | 10.5293 | 0.0333 | 0.0333 |
| 79 | 8.7612 | 8.76 | 9.8761 | 9.8748 | 0.0382 | 0.0383 |
| 80 | 8.2006 | 8.1995 | 9.2489 | 9.2477 | 0.0435 | 0.0435 |
| 81 | 7.6591 | 7.658 | 8.6466 | 8.6454 | 0.0499 | 0.05 |
| 82 | 7.1423 | 7.1413 | 8.0748 | 8.0737 | 0.0567 | 0.0567 |
| 83 | 6.6472 | 6.6463 | 7.53 | 7.529 | 0.0646 | 0.0646 |
| 84 | 6.1776 | 6.1767 | 7.0157 | 7.0148 | 0.0725 | 0.0725 |
| 85 | 5.7268 | 5.726 | 6.5247 | 6.5238 | 0.0831 | 0.0831 |
| 86 | 5.3084 | 5.3077 | 6.0709 | 6.0701 | 0.0945 | 0.0945 |
| 87 | 4.9208 | 4.9201 | 5.6521 | 5.6513 | 0.1064 | 0.1064 |
| 88 | 4.5615 | 4.5609 | 5.2652 | 5.2645 | 0.1196 | 0.1196 |
| 89 | 4.2328 | 4.2322 | 4.9124 | 4.9117 | 0.137 | 0.1371 |
| 90 | 3.954 | 3.9534 | 4.613 | 4.6124 | 0.1408 | 0.1409 |
| RMSE | 0.001 | | 0.0011 | | 0 | |

Source: Author's preparation.

5.4 CONCLUSIONS

Despite the popularity and long tradition of ensemble learning methods in the statistical and forecasting literature, model combination has received little attention in the actuarial, demographic or pension literature with some noticeable exceptions (see, e.g., L. Breiman (1996), M. Ayuso et. al. (2021)). This Chapter expands previous research by exploring the use of ensemble learning for multi-population age-specific mortality forecasting, life table construction and annuity pricing, taking data for 14 European countries, USA, Canada, Australia, and Japan from 1960 to 2018. To illustrate the empirical results, Portugal was picked as one of developed countries, and the age interval was selected based on the most impressable ages with both COVID19 and senility. Our results suggest that the proposed ensemble time series model combining traditional ARIMA models, MLP as a neural network autoregression approach, and SSA as a non-parametric technique in the field of time series analysis

performs better than merely extrapolating the mortality patterns observed in a given age interval. The ensemble method exhibits the best robust results overall with minimum RMSE in the presence of structural changes in the shape of time series at the time of COVID19. Further research should investigate the use of alternative distance functions to measure the similarity between time series with invariance, taking into account amplitude invariance and offset invariance, and experiment with alternative time series clustering techniques which do not depend on an averaging function (e.g., hierarchical clustering). Further research should explore the implications of using BME in combination with time series distance functions in pricing multi-population longevity-linked securities.■

CHAPTER SIX - CONCLUSIONS

6.1 SUMMARY OF FINDINGS AND CONTRIBUTIONS

This thesis presented data science solutions for risk management in financial services of different economic sectors, which would lead to better financial stability especially at the time of crisis.

The study covered almost all stages of applying data science for risk analysis including Big Data gathering (web scraping and text mining), feature engineering, machine learning modelling, time series forecasting, and deploying the results associated with the topic. The studies reported in chapters 2, 3, 4, and 5 contribute with both theoretical and practical insights.

The data preparation stage revealed that data collection during a stressed situation in economy or society is not qualified as the normal situations and some measurement errors are involved, which should be treated in a different way than standard data quality frameworks.

The feature engineering stage revealed that our suggested index could improve the quality of explainable models even more than black-box neural networks with automatic feature engineering inside.

The machine learning modelling stage showed that with our new approach, we could approve the responsible artificial intelligence conditions such as fairness and transparency by using explainable models such as Logistic Regression with high accuracy.

Findings of the time series forecasting with panel data, ensemble learning, layered learning, and clustering approach showed that our new approach has the potential to diagnose the rapid changes in trend at the time of crisis, better than existed models.

Applying these approaches to financial Big Data revealed new opportunities to better risk management of new banks and online banks. The reproducible PySpark code of this section is published online and it could be considered for applying to the platforms of loan providers. Additionally, research on online

credit risk demonstrates the significant role of our new approach on improving confidentiality and privacy, financial inclusion, and compliance risk impacts.

Overall, we hope that by providing these new methods for machine learning and time series modelling, this dissertation helps to design and implement better risk management portals, increasing the accuracy of financial models, technology adoption and usage of financial institutions, providing better and more efficient financial services to people, and contributing to the better financial stability by using new technologies.

6.2 LIMITATIONS

We acknowledge that this dissertation has several limitations.

Availability of data related to financial activities could be considered as the first limitation for these kinds of studies. Financial institutions have restrictions to disclose their financial data according to the regulations.

Additionally, the quality of data is a concern. Especially when the scope of the study is including the time of crisis; means that even if the dataset is available for these kinds of analyses, the pre-processing stage of modelling could be still a challenge for the researchers.

For dealing with Big Data and also the ensemble time series modelling for panel data, the computational power is another important factor to get the results in a proper time. For instance, each iteration of our proposed models took about 24 hours on a high-performance desktop computer. It would be more challenging when the researcher is developing a new approach and needs to repeat it after each improvement in the theories or practical sections.

6.3 FUTURE RESEARCH

Currently, several financial institutions are starting to develop mobile applications for money transferring, granting loans, and providing insurance services. Taking into account the current high usage of mobile devices and financial apps, adopting the proposed approaches to the small amount of memory and processor limitation of cell phones is an interesting avenue of research. Using the call detail records of mobile service providers for the risk management of financial institutions is another interesting topic. For time

series models in chapter 4, could suggest further studies for optimization of θ in formula (4-3), i.e., investigating some dynamic selection of optimum θ for better performance. Additionally, as we could see that usual neural networks could not model time series adequately, especially for incomplete/limited data in early epidemic, authors could recommend the recurrent neural networks like LSTM and GRU for future studies, and their effect on accuracy and computation time and other resources, which might have the potential to outperform ensemble time series models with reasonable increase in computation power requirement. Considering non-linear meta-learning approach instead of a linear one, and prediction intervals instead of point forecast, could be recommended as the next step. Besides, analysing heterogeneous and homogeneous countries by classification techniques could be considered as another layer after the application of the forecasting methods. Therefore, a clustering analysis might be useful to be implemented based on the notion of excess mortality.■

REFERENCES

- Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73, 1–10.
<https://doi.org/10.1016/j.eswa.2016.12.020>
- Acosta-González, E., & Fernández-Rodríguez, F. (2014). Forecasting Financial Failure of Firms via Genetic Algorithms. *Computational Economics*, 43(2), 133–157.
<https://doi.org/10.1007/s10614-013-9392-9>
- Aiolfi, M. and Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, 135(1): 31–53.
- Akyuz, A. O., Uysal, M., Bulbul, B. A., & Uysal, M. O. (2017). Ensemble approach for time series analysis in demand forecasting: Ensemble learning. In *Proceedings - 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2017* (pp. 7–12). Institute of Electrical and Electronics Engineers Inc.
<https://doi.org/10.1109/INISTA.2017.8001123>
- Alho, J., Bravo, J., & Palmer, E. (2012). Annuities and Life Expectancy in NDC. In *Nonfinancial Defined Contribution Pension Schemes in a Changing Pension World* (pp. 395–436, doi: 10.1596/9780821394786_ch22). The World Bank.
https://doi.org/10.1596/9780821394786_ch22
- Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4), 589. <https://doi.org/10.2307/2978933>
- Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 1(1), 29–54.
[https://doi.org/10.1016/0378-4266\(77\)90017-6](https://doi.org/10.1016/0378-4266(77)90017-6)
- Andrawis R.R., Atiya, A.F., & El-Shishiny, H. (2011). Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. *International Journal of Forecasting* 27(3): 672–688.
- Arminger, G., Enache, D., & Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. *Computational Statistics*, 12(2), 293–310. Retrieved from
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4801
- Ashofteh, A. (2018). Mining Big Data in statistical systems of the monetary financial institutions (MFIs). In *International Conference on Advanced Research Methods and Analytics (CARMA)* (p. doi: 10.4995/carma2018.2018.8570). Valencia: Universitat Politecnica de Valencia. <https://doi.org/10.4995/carma2018.2018.8570>
- Ashofteh, A., & Bravo, J. M. (2019). A non-parametric-based computationally efficient approach for credit scoring. In *Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao. Associacao Portuguesa de Sistemas de Informacao*.
- Ashofteh A. and Bravo J. M. (2020a). Corona-virus disease (COVID-19) Data-set with Improved Measurement Errors of Referenced Official Data Sources. *Mendeley Data [Distributor]*, V2 [Version], with reference to dashboard. doi: 10.17632/nw5m4hs3jr.3 , available from: <http://dx.doi.org/10.17632/nw5m4hs3jr.3>
- Ashofteh, A., & Bravo, J. M. (2020b). A study on the quality of novel coronavirus (COVID-19) official datasets. *Statistical Journal of the IAOS*, 36(2), 291–301.
<https://doi.org/10.3233/SJI-200674>
- Ashofteh, A., & Bravo, J. M. (2021a). A Conservative Approach for Online Credit Scoring. *Expert Systems with Applications*, 114835. <https://doi.org/10.1016/j.eswa.2021.114835>
- Ashofteh, A., & Bravo, J. M. (2021b). Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical

References

- systems. *Statistical Journal of the IAOS*, Preprint(Preprint), 1–19.
<https://doi.org/10.3233/SJI-210841>
- Ashofteh, A., & Bravo, J. M. (2021c). Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach. 16th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, p. 1-6. <https://doi.org/10.23919/CISTI52073.2021.9476583>
- Ashofteh, A., & Bravo, J. M. (2021d) Spark Code: A Novel Conservative Approach for Online Credit Scoring [Source Code]. <https://doi.org/10.24433/CO.1963899.v1>
- Ashofteh, A., Bravo, J. M., & Ayuso, M. (2021a). An Ensemble Learning Strategy for Panel Time Series Forecasting of Excess Mortality during the COVID-19 Pandemic. (In press).
- Ashofteh, A., Bravo, J. M., & Ayuso, M. (2021b). A Novel Layered Learning Approach for Forecasting Respiratory Disease Excess Mortality during the COVID-19 pandemic. Proceeding of CAPSI 2021. 21^a Conferência da Associação Portuguesa de Sistemas de Informação, "Sociedade 5.0: Os desafios e as Oportunidades para os Sistemas de Informação". [21th Portuguese Association of Information Systems Conference]. Associação Portuguesa de Sistemas de Informação
- Ashofteh, A., Bravo, J. M., & Ayuso, M. (2021c). Time Series Data for Monthly Respiratory Diseases Deaths in 61 Countries from 2000 to 2016 - Mendeley Data. Mendeley Data, <https://doi.org/10.17632/gjj68bmv8d.2>
- Avery, R. B., Bostic, R. W., Calem, P. S., & Canner, G. B. (2000). Credit scoring: Statistical issues and evidence from credit-bureau files. *Real Estate Economics*, 28(3), 523–547. <https://doi.org/10.1111/1540-6229.00811>
- Ayuso, M., Bravo, J. M., Holzmann, R., & Palmer, E. (2021b). Automatic indexation of pension age to life expectancy: When policy design matters. *Risks*, 9(5), 96. <https://doi.org/10.3390/risks9050096>
- Ayuso, Mercedes, Bravo, J. M., & Holzmann, R. (2021a). Getting life expectancy estimates right for pension policy: Period versus cohort approach. *Journal of Pension Economics and Finance*, 20(2), 212–231, doi: 10.1017/S1474747220000050. <https://doi.org/10.1017/S1474747220000050>
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635. <https://doi.org/10.1057/palgrave.jors.2601545>
- Baesens, Bart, Roesch, D., & Scheule, H. (2016). Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS - Bart Baesens, Daniel Roesch, Harald Scheule - Google Books. Retrieved from <https://books.google.com/books?hl=en&lr=&id=ornsDAAAQBAJ&oi=fnd&pg=PR11&dq=Credit+risk+analytics:+Measurement+techniques,+applications,+and+examples+in+SAS&ots=NHbvLqOdKi&sig=RshAgweYNZ8JI0Monkk137s8Y60>
- Banerjee, A., L. Pasea, S. Harris, A. Gonzalez-Izquierdo, A. Torralbo, L. Shallcross, M. Noursadeghi, D. Pillay, N. Sebire, C. Holmes, C. Pagel, W.K. Wong, C. Langenberg, B. Williams, S. Denaxas and H. Hemingway. (2020). Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. *The Lancet*. 395: 1715–25.
- Bates, J. M. & Granger, C. W. J. (1969). The Combination of Forecasts, *Journal of the Operational Research Society*, 20:4, 451-468, DOI: 10.1057/jors.1969.103
- Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, 29(4), 563–574. <https://doi.org/10.1016/j.ijforecast.2013.04.003>
- Bjorkegren, D., & Grissen, D. (2018). Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2611775>
- Bravo, J. M. (2020). Longevity-Linked Life Annuities: A Bayesian Model Ensemble Pricing Approach. CAPSI 2020 Proceedings. Retrieved from

References

- <https://aisel.aisnet.org/capsi2020/29>
- Bravo, J. M., & Coelho, E. (2019). Forecasting Subnational Demographic Data using Seasonal Time Series Methods. *Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao 2019*, 24 [CAPSI 2019]. Available at: <https://aisel.aisnet.org/capsi2019/24>.
- Bravo, J. M., & Ayuso, M. (2021). Forecasting the retirement age: A Bayesian Model Ensemble Approach. *Advances in Intelligent Systems and Computing*, Volume 1365 AIST, 123 – 135 [2021 World Conference on Information Systems and Technologies, WorldCIST 2021] Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_12. In: Rocha Á., Adeli H., Dzemyda G., Moreira F., Ramalho Correia A.M. (eds) Trends and Applications in Information Systems and Technologies.
- Bravo, J. M., Ayuso, M. (2020). Previsões de mortalidade e de esperança de vida mediante combinação Bayesiana de modelos: Uma aplicação à população portuguesa. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, E40, 128–144
- Bravo, J. M., Ayuso, M., Holzmann, R., & Palmer, E. (2021). Addressing the Life Expectancy Gap in Pension Policy. *Insurance: Mathematics and Economics*, 99, 200–221. <https://doi.org/10.1016/j.insmatheco.2021.03.025>
- Bravo, J.M., Ayuso, M., Holzmann, R., and Palmer, E. (2021). Intergenerational Actuarial Fairness when Longevity Increases: Amending the Retirement Age. Preprint.
- Bravo, J. M., & Ayuso, M. (2020). Mortality and life expectancy forecasts using bayesian model combinations: An application to the Portuguese population. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, 2020(E40), 128–144, doi: 10.17013/risti.40.128-145. <https://doi.org/10.17013/risti.40.128-145>
- Bravo, J. M., & Herce, J. A. (2020). Career breaks, broken pensions? Long-run effects of early and late-career unemployment spells on pension entitlements. *Journal of Pension Economics and Finance*, 1–27, DOI: 10.1017/S1474747220000189. <https://doi.org/10.1017/S1474747220000189>
- Bravo, J. M., & Nunes, J. P. V. (2021). Pricing longevity derivatives via Fourier transforms. *Insurance: Mathematics and Economics*, 96, 81–97, doi: 10.1016/j.insmatheco.2020.10.008. <https://doi.org/10.1016/j.insmatheco.2020.10.008>
- Bravo, J. M., & El Mekkaoui de Freitas, N. (2018). Valuation of longevity-linked life annuities. *Insurance: Mathematics and Economics*, 78, 212–229, doi: 10.1016/j.insmatheco.2017.09.009. <https://doi.org/10.1016/j.insmatheco.2017.09.009>
- Bravo, J. M. (2016). Taxation of pensions in Portugal: A semi-dual income tax system. *CESifo DICE Report*, 14(1), 14–23.
- Bravo, J. M. (2019). Funding for longer lives. Retirement wallet and risk-sharing annuities. *EKONOMIAZ*, 96(2), 268–291. Retrieved from <https://ideas.repec.org/a/ekz/ekonoz/2019212.html>
- Bravo, J. M. (2021). Pricing Participating Longevity-Linked Life Annuities: A Bayesian Model Ensemble Approach. *European Actuarial Journal*, <https://doi.org/10.1007/s13385-021-00279-w>.
- Brazdil, P., Carrier, C., Soares, C., & Vilalta, R. (2009, November 22). *Metalearning - Applications to Data Mining. Cognitive Technologies*. Springer Berlin Heidelberg. Retrieved from [https://books.google.com/books?hl=en&lr=&id=-Gsi_cxZGpcC&oi=fnd&pg=PA1&dq=Brazdil,+P.B.+\(Ed.\),+2009.+Metalearning:+applications+to+data+mining,+Cognitive+technologies.+Springer,+&ots=wkZEoWxtMg&sig=o7ZkVPhvJJROwAdWP1hnGIHm7To](https://books.google.com/books?hl=en&lr=&id=-Gsi_cxZGpcC&oi=fnd&pg=PA1&dq=Brazdil,+P.B.+(Ed.),+2009.+Metalearning:+applications+to+data+mining,+Cognitive+technologies.+Springer,+&ots=wkZEoWxtMg&sig=o7ZkVPhvJJROwAdWP1hnGIHm7To)
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1023/A:1018054314350>
- Brouhns, N., Denuit, M., & Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3), 373–393, doi: 10.1016/S0167-6687(02)00185-3. [https://doi.org/10.1016/S0167-6687\(02\)00185-3](https://doi.org/10.1016/S0167-6687(02)00185-3)

References

- Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking and Finance*, 72, 218–239. <https://doi.org/10.1016/j.jbankfin.2016.07.015>
- Capistrán, C., Timmermann, A., & Aiolfi, M. 2010. Forecast combinations, Working Papers.
- Castro, V. (2013). Macroeconomic determinants of the credit risk in the banking system: The case of the GIPSI. *Economic Modelling*, 31(1), 672–683. <https://doi.org/10.1016/j.econmod.2013.01.027>
- Cerqueira, V., Torgo, L., & Soares, C. (2020). Early Anomaly Detection in Time Series: A Hierarchical Approach for Predicting Critical Health Episodes. Retrieved from <https://arxiv.org/pdf/2010.11595.pdf>
- Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.
- Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. *Risk Management*, 18(4), 264–287. <https://doi.org/10.1057/s41283-016-0006-4>
- Chamboko, R., & Bravo, J. M. (2019a). Frailty correlated default on retail consumer loans in Zimbabwe. *International Journal of Applied Decision Sciences*, 12(3), 257–270. <https://doi.org/10.1504/IJADS.2019.100436>
- Chamboko, R., & Bravo, J. M. (2020). A Multi-State Approach to Modelling Intermediate Events and Multiple Mortgage Loan Outcomes. *Risks*, 8(2), 64. <https://doi.org/10.3390/risks8020064>
- Chamboko, R., & Bravo, J. M. V. (2019b). Modelling and forecasting recurrent recovery events on consumer loans. *International Journal of Applied Decision Sciences*, 12(3), 271–287. <https://doi.org/10.1504/IJADS.2019.100440>
- Checchi, F. and L. Roberts. (2005). Interpreting and using mortality data in humanitarian emergencies. *Humanitarian Practice Network*, 52. Interpreting and using mortality data in humanitarian emergencies.
- Chen H., Hailey D., Wang N., and Yu P. A. (2014). Review of data quality assessment methods for public health information systems, *Int J Environ Res Public Health*. 11(5): 5170–5207. doi: 10.3390/ijerph110505170.
- Cheng P, Gilchrist A, Robinson KM, Paul L. (2009). The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. *Health Inf Manag*. 38(1):35-46.
- Chinese center for disease control and prevention (Chinese CDC), National health commission updates [Internet]. Available from: <http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm#NHCMar18>
- Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. In *Personal and Ubiquitous Computing* (Vol. 17, pp. 433–450). Springer. <https://doi.org/10.1007/s00779-011-0490-1>
- Cleofas-Sánchez, L., García, V., Marqués, A. I., & Sánchez, J. S. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing Journal*, 44, 144–152. <https://doi.org/10.1016/j.asoc.2016.04.005>
- Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess (with Discussion). *Journal of Official Statistics*, 6, 3–73. Retrieved from [http://cs.wellesley.edu/~cs315/Papers/stl statistical model.pdf](http://cs.wellesley.edu/~cs315/Papers/stl%20statistical%20model.pdf)
- Cui, S., Wang, Y., Wang, D., Sai, Q., Huang, Z., and Cheng, T. C. E. (2021). A two-layer nested heterogeneous ensemble learning predictive method for COVID-19 mortality. *Applied Soft Computing*, vol. 113, p. 107946.
- De Angelis, D., Presanis, A. M., Birrell, P., Tomba, G. S., and Housed, T. (2015). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*.

References

- 10: 83–87. DOI: 10.1016/j.epidem.2014.09.004.
- De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496), 1513–1527. <https://doi.org/10.1198/jasa.2011.tm09771>
- De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, 120 (1): 190–204.
- Deng, Y., Quigley, J. M., & Van Order, R. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2), 275–307. <https://doi.org/10.1111/1468-0262.00110>
- Denuit, M., & Goderniaux, A.-C. (2005). Closing and projecting life tables using log-linear models. *Bulletin of the Swiss Association of Actuaries*, (1), 29–48.
- Do, T.-M.-T., & Gatica-Perez, D. (2010). By Their Apps You Shall Understand Them: Mining Large-scale Patterns of Mobile Phone Usage. *Telecommunications Policy*, 24(ii), 27:1–27:10. <https://doi.org/10.1145/1899475.1899502>
- Don, E., Hongru, D. & Lauren, G. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5), 533–534.
- Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using Conditional Generative Adversarial Networks Improving Artificial Data Generation View project Factors impacting on mobile payment adoption on user perspective View project Georgios Douzas Effective d. *Expert Systems With Applications*, 91, 464–471. <https://doi.org/10.1016/j.eswa.2017.09.030>
- Erdogan, B. E. (2013). Prediction of bankruptcy using support vector machines: An application to bank bankruptcy. *Journal of Statistical Computation and Simulation*, 83(8), 1543–1555. <https://doi.org/10.1080/00949655.2012.666550>
- European Centre for Disease Prevention and Control, Data on the geographic distribution of COVID-19 cases worldwide [Internet]. Available from: <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>
- Feng, X., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing Journal*, 65, 139–151. <https://doi.org/10.1016/j.asoc.2018.01.021>
- Fensterstock, A. (2005). Credit scoring and the next step. *Business Credit*, 107, 46.
- Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 1–37. <https://doi.org/10.1145/2523813>
- García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10), 2044–2064. <https://doi.org/10.1016/j.ins.2009.12.010>
- García, V., Marqués, A. I., & Sánchez, J. S. (2012). On the use of data filtering techniques for credit risk prediction with instance-based models. *Expert Systems with Applications*, 39(18), 13267–13276. <https://doi.org/10.1016/j.eswa.2012.05.075>
- García, Vicente, Marqués, A. I., & Sánchez, J. S. (2019a). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion*, 47, 88–101. <https://doi.org/10.1016/j.inffus.2018.07.004>
- García, Vicente, Marqués, A. I., Sánchez, J. S., & Ochoa-Domínguez, H. J. (2019b). Dissimilarity-Based Linear Models for Corporate Bankruptcy Prediction. *Computational Economics*, 53(3), 1019–1031. <https://doi.org/10.1007/s10614-017-9783-4>
- Georgiou, A. V. (2021). The manipulation of official statistics as corruption and ways of understanding it. *Statistical Journal of the IAOS*, Preprint(Preprint), 1–21.

References

- <https://doi.org/10.3233/sji-200667>
- Gerardi, K., Shapiro, A., & Willen, P. (2007). Subprime outcomes: Risky mortgages, homeownership and foreclosure.
- Gicić, A., & Subasi, A. (2019). Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. *Expert Systems*, 36(2), e12363. <https://doi.org/10.1111/exsy.12363>
- Ha, S. H. (2010). Behavioral assessment of recoverable credit of retailer's customers. *Information Sciences*, 180(19), 3703–3717. <https://doi.org/10.1016/j.ins.2010.06.012>
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 160(3), 523–541. <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
- Ho Ha, S., & Krishnan, R. (2012). Predicting repayment of the credit card debt. *Computers and Operations Research*, 39(4), 765–773. <https://doi.org/10.1016/j.cor.2010.10.032>
- Huang, Guang-Bin & Zhu, Qin-Yu & Siew, Chee. (2004). Extreme learning machine: A new learning scheme of feedforward neural networks. *IEEE International Conference on Neural Networks — Conference Proceedings*. 2. 985–990 vol.2. 10.1109/IJCNN.2004.1380068.
- Human Mortality Database. (2021). Retrieved from www.mortality.org
- Hunt, A., & Blake, D. (2021). On the structure and classification of mortality models. *North American Actuarial Journal*, 25:sup1, S215-S234. DOI: 10.1080/10920277.2019.1649156.
- Hyndman, R. J., & Shahid Ullah, M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51(10), 4942–4956. <https://doi.org/10.1016/j.csda.2006.07.028>
- Hyndman, R. J., Booth, H. & Yasmeeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography* 50(1), 261–283.
- Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., ... Yasmeeen, F. (2020). *Forecasting Functions for Time Series and Linear Models*. Retrieved from <https://pkg.robjhyndman.com/forecast/>
- Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. [OTexts.com/fpp3](https://www.otexts.com/fpp3).
- International Standard ISO 3166-1, Codes for the representation of names of countries and their subdivisions--Part 1: Country codes, ISO 3166-1: 2006 (E/F), International Organization on Standardization (Geneva, 2006) [Internet]. The latest version is available from: http://www.iso.org/iso/home/standards/country_codes.htm.
- Islam, N., Shkolnikov, V. M., Acosta, R. J., Klimkin, I., Kawachi, I., Irizarry, R. A., Alicandro, G., Khunti, K., Yates, T., Jdanov, D. A., White, M., Lewington, S., & Lacey, B. (2021). Excess deaths associated with covid-19 pandemic in 2020: age and sex disaggregated time series analysis in 29 high income countries. *BMJ (Clinical research ed.)*, 373, n1137.
- Jacobson, T., & Roszbach, K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of Banking and Finance*, 27(4), 615–633. [https://doi.org/10.1016/S0378-4266\(01\)00254-0](https://doi.org/10.1016/S0378-4266(01)00254-0)
- Jones, S., & Hensher, D. A. (2004, October 1). Predicting firm financial distress: A mixed logit model. *Accounting Review*. American Accounting Association. <https://doi.org/10.2308/accr.2004.79.4.1011>
- Jordan, M. I. (2013). On statistics, computation and scalability. *Bernoulli*, 19(4), 1378–1390. <https://doi.org/10.3150/12-BEJSP17>
- Jose, V.R., & Winkler, R. L. (2008). Simple robust averages of forecasts: some empirical results. *International Journal of Forecasting* 24(1): 163--169.
- Kawashima T. et al. (2020). Excess all-cause deaths during coronavirus disease pandemic,

References

- Japan, January-May 2020. Emerging Infectious Diseases. Centers for Disease Control and Prevention 27(3): 789–795.
- Kellison, B., & Wortham, G. (2003). Bureau of Business Research • McCombs School of Business • The University of Texas at Austin SPECIAL ISSUE. Retrieved from <https://repositories.lib.utexas.edu/handle/2152/15187>
- Khairalla, M. A., Ning, X., AL-Jallad, N. T., & El-Faroug, M. O. (2018). Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model. *Energies*, 11(6), 1605. <https://doi.org/10.3390/en11061605>
- Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(4), 795–816. <https://doi.org/10.1111/rssb.12050>
- Kontis, V., Bennett, J.E., Rashid, T. et al. (2020). Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries. *Nature Medicine* 26, 1919–1928.
- Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13), 5125–5131. <https://doi.org/10.1016/j.eswa.2013.03.019>
- Kwak, W., Shi, Y., & Kou, G. (2012). Bankruptcy prediction for Korean firms after the 1997 financial crisis: Using a multiple criteria linear programming data mining approach. *Review of Quantitative Finance and Accounting*, 38(4), 441–453. <https://doi.org/10.1007/s11156-011-0238-z>
- Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419), 659–671, doi: 10.1080/01621459.1992.10475265. <https://doi.org/10.1080/01621459.1992.10475265>
- Lensberg, T., Eilifsen, A., & McKee, T. E. (2006). Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research*, 169(2), 677–697. <https://doi.org/10.1016/j.ejor.2004.06.013>
- Leon, D.A., V.M. Shkolnikov, L. Smeeth, P. Magnus, M. Pechholdová & C.I. Jarvis. (2020). COVID-19: a need for real-time monitoring of weekly excess deaths. *Lancet*. 395: e81.
- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using Semi-supervised Support Vector Machines. *Expert Systems with Applications*, 74, 105–114. <https://doi.org/10.1016/j.eswa.2017.01.011>
- Liang, F., Cheng, Y., Song, Q., Park, J., & Yang, P. (2013). A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association*, 108(501), 325–339. <https://doi.org/10.1080/01621459.2012.746061>
- Liu, Z. F., & Pan, S. (2018). Fuzzy-Rough Instance Selection Combined with Effective Classifiers in Credit Scoring. *Neural Processing Letters*, 47(1), 193–202. <https://doi.org/10.1007/s11063-017-9641-3>
- Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65, 465–470. <https://doi.org/10.1016/j.engappai.2016.12.002>
- Ma, P., Mahoney, M. W., Yu, B., & Yu Ma, B. (2015). A Statistical Perspective on Algorithmic Leveraging. *Journal of Machine Learning Research* (Vol. 16). <https://doi.org/10.5555/2789272.2831141>
- Maclaurin, D., & Adams, R. P. (2015). Firefly Monte Carlo: Exact MCMC with subsets of data. In *IJCAI International Joint Conference on Artificial Intelligence* (Vol. 2015-Janua, pp. 4289–4295). International Joint Conferences on Artificial Intelligence.

References

- Mahmoudvand, R., Konstantinides, D., & Rodrigues, P. C. (2017). Forecasting mortality rate by multivariate singular spectrum analysis. *Applied Stochastic Models in Business and Industry*, 33(6), 717–732, doi: 10.1002/asmb.2274. <https://doi.org/10.1002/asmb.2274>
- Makridakis, R. Winkler Averages of forecasts: Some empirical results *Management Science* (1983), pp. 987–996
- Maldonado, S., Peters, G., & Weber, R. (2020). Credit scoring using three-way decisions with probabilistic rough sets. *Information Sciences*, 507, 700–714. <https://doi.org/10.1016/j.ins.2018.08.001>
- Noh, H. J., Roh, T. H., & Han, I. (2005). Prognostic personal credit risk model considering censored information. *Expert Systems with Applications*, 28(4), 753–762. <https://doi.org/10.1016/j.eswa.2004.12.032>
- Nyitrai, T., & Virág, M. (2019). The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences*, 67, 34–42. <https://doi.org/10.1016/j.seps.2018.08.004>
- Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing Journal*, 74, 26–39. <https://doi.org/10.1016/j.asoc.2018.10.004>
- Pedro, J. S., Proserpio, D., & Oliver, N. (2015). Mobiscore: Towards universal credit scoring from mobile phone data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9146, pp. 195–207). Springer Verlag. https://doi.org/10.1007/978-3-319-20267-9_16
- Plawiak, P., Abdar, M., & Rajendra Acharya, U. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing Journal*, 84, 105740. <https://doi.org/10.1016/j.asoc.2019.105740>
- Premachandra, I. M., Bhabra, G. S., & Sueyoshi, T. (2009). DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research*, 193(2), 412–424. <https://doi.org/10.1016/j.ejor.2007.11.036>
- Quittner, J. (2003). Credit cards: sub-prime's tech dilemma: with delinquencies and charge-offs on the rise, the industry examines the role of automated decisioning. *Bank Technology*, 16(1), 19–23.
- Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian Model Averaging to calibrate forecast ensembles. *Journal of American Meteorological society*, 133: 1155--1174.
- Rimmer, J. (2005). Contemporary Changes in Credit Scoring. *Credit Control*, 26(4), 56–60. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=17602788&site=ehost-live>
- Rutherford G., McFarland W., Spindler H., White K., Patel S., Grasse J.A., Sabin K., Smith N., Tache S., Garcia J.C., Stoneburner R. Public health triangulation: approach and application to synthesizing data to understand national and local HIV epidemics. *BMC Public Health* 10:447.
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*. Retrieved from <https://academic.oup.com/bioinformatics/article-abstract/23/19/2507/185254>
- Samuels, J.D., & Sekkel, R.M. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, 33(1): 48--60.
- Sarlija, N., Bensic, M., & Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications*, 36(5), 8778–8788. <https://doi.org/10.1016/j.eswa.2008.11.042>
- Schifano, E. D., Wu, J., Wang, C., Yan, J., & Chen, M. H. (2016). Online Updating of Statistical Inference in the Big Data Setting. *Technometrics*, 58(3), 393–403.

References

- <https://doi.org/10.1080/00401706.2016.1142900>
- Scortichini, M. et al. (2020). Excess mortality during the COVID-19 outbreak in Italy: a two-stage interrupted time-series analysis, *International Journal of Epidemiology*, 49(6): 1909–1917.
- Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1), 127–135. <https://doi.org/10.1016/j.eswa.2004.08.009>
- Shin, M. S., Sim, B., Jang, W. M., & Lee, J. Y. (2021). Estimation of Excess All-cause Mortality during COVID-19 Pandemic in Korea. *Journal of Korean Medical Science*, 36(39), 1–10. <https://doi.org/10.3346/JKMS.2021.36.E280>
- Simões, C., Oliveira, L., & Bravo, J. M. (2021). Immunization Strategies for Funding Multiple Inflation-Linked Retirement Income Benefits. *Risks*, 9(4), 60, doi: 10.3390/risks9040060. <https://doi.org/10.3390/risks9040060>
- Skyler, S., Eric, M., Isaac, M., & Felix, K. (2017). Mobile phone-based Credit Scoring. *NetMob2017*.
- Slowinski, R., & Zopounidis, C. (1995). Application of the Rough Set Approach to Evaluation of Bankruptcy Risk. *Intelligent Systems in Accounting, Finance and Management*, 4(1), 27–41. <https://doi.org/10.1002/j.1099-1174.1995.tb00078.x>
- Song, Q., & Liang, F. (2015). A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 77(5), 947–972. <https://doi.org/10.1111/rssb.12095>
- Standard country or area codes for statistical use (ST/ESA/STAT/SER.M/49/Rev.3) [Internet], available from: <http://unstats.un.org/unsd/methods/m49/m49.htm>.
- Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, 50(2), 277–289. <https://doi.org/10.1287/opre.50.2.277.426>
- Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2), 149–172. [https://doi.org/10.1016/S0169-2070\(00\)00034-0](https://doi.org/10.1016/S0169-2070(00)00034-0)
- Tian, Y., Yong, Z., & Luo, J. (2018). A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines. *Applied Soft Computing Journal*, 73, 96–105. <https://doi.org/10.1016/j.asoc.2018.08.021>
- Tong, E. N. C., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, 218(1), 132–139. <https://doi.org/10.1016/j.ejor.2011.10.007>
- Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. In *IEEE International Conference on Neural Networks - Conference Proceedings (Vol. 1, pp. 90–95)*. IEEE. <https://doi.org/10.1109/icnn.1996.548872>
- United Nations, Department of economic and social affairs, Population division. *World population prospects 2019*, Online edition. Rev. 1.
- Ventura Bravo, J. M., & Pereira da Silva, C. M. (2006). Immunization using a stochastic-process independent multi-factor model: The Portuguese experience. *Journal of Banking and Finance*, 30(1), 133–156, doi: 10.1016/j.jbankfin.2005.01.006. <https://doi.org/10.1016/j.jbankfin.2005.01.006>
- Verkasalo, H., López-Nicolás, C., Molina-Castillo, F. J., & Bouwman, H. (2010). Analysis of users and non-users of smartphone applications. *Telematics and Informatics*, 27(3), 242–255. <https://doi.org/10.1016/j.tele.2009.11.001>
- Volkov, A., Benoit, D. F., & Van den Poel, D. (2017). Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination. *Decision Support Systems*, 98, 59–68. <https://doi.org/10.1016/j.dss.2017.04.008>
- Wiśniowski, A., Smith, P. W. F., Bijak, J., Raymer, J., & Forster, J. J. (2015). Bayesian

References

- Population Forecasting: Extending the Lee-Carter Method. *Demography*, 52(3), 1035–1059. <https://doi.org/10.1007/s13524-015-0389-y>
- World Health Organization, D. of I. E. and R. (2018). Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2016. Retrieved August 19, 2019, from https://www.who.int/healthinfo/global_burden_disease/GHE2016_Death-Rates-country.xls?ua=1
- World Health Organization. (2021). Coronavirus Disease (COVID-19) Situation Reports. Coronavirus disease (COVID-19)/Situation reports. Retrieved from <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>
- World Health Organization. Geneva, Switzerland: Coronavirus disease, (2019). (COVID-19) situation reports [Internet]. Available from: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>.
- World Health Organization; Geneva, Switzerland. (2012). Framework and standards for country health information systems. Second edition.
- Xia, Y. (2019). A Novel Reject Inference Model Using Outlier Detection and Gradient Boosting Technique in Peer-to-Peer Lending. *IEEE Access*, 7, 92893–92907. <https://doi.org/10.1109/ACCESS.2019.2927602>
- Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, 93, 182–199. <https://doi.org/10.1016/j.eswa.2017.10.022>
- Yorifuji, T. , Matsumoto, N., & Takao, S. (2021). Excess all-cause mortality during the COVID-19 outbreak in Japan. *Journal of Epidemiology*, 31(1): 90–92.
- Zhang, H., & Liu, Q. (2019). Online Learning Method for Drift and Imbalance Problem in Client Credit Assessment. *Symmetry*, 11(7), 890. <https://doi.org/10.3390/sym11070890>
- Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M., & Chen, H. (2017). An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach. *Computational Economics*, 49(2), 325–341. <https://doi.org/10.1007/s10614-016-9562-7>
- Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, 22, 59. <https://doi.org/10.2307/2490859>

APPENDIX

APPENDIX-1 PYSPARK CODE FOR PROPOSED MACHINE LEARNING APPROACH - CHAPTER III.

```
from pyspark.sql.types import IntegerType
from pyspark.sql.types import DoubleType
# loading train data set
file_location = "/FileStore/tables/paper_train1.csv"
file_type = "csv"
# CSV options
infer_schema = "false"
first_row_is_header = "true"
delimiter = ","
# The applied options are for CSV files. For other file types, these will be
ignored.
train = spark.read.format(file_type) \
    .option("inferSchema", infer_schema) \
    .option("header", first_row_is_header) \
    .option("sep", delimiter) \
    .load(file_location).cache()
# loading test data set
file_location = "/FileStore/tables/paper_valid1.csv"
file_type = "csv"
# CSV options
infer_schema = "false"
first_row_is_header = "true"
delimiter = ","
valid = spark.read.format(file_type) \
    .option("inferSchema", infer_schema) \
```

Appendixes

```
.option("header", first_row_is_header) \
.option("sep", delimiter) \
.load(file_location).cache()
from pyspark.sql.functions import *
train = train.withColumn("loan_amnt", train.loan_amnt.cast("float"))\
    .withColumn("emp_length", train.emp_length.cast("float"))\
    .withColumn("annual_inc", train.annual_inc.cast("float"))\
    .withColumn("dti", train.dti.cast("float"))\
    .withColumn("delinq_2yrs", train.delinq_2yrs.cast("float"))\
    .withColumn("revol_util", regexp_replace("revol_util", "%",
""").cast("float"))\
    .withColumn("total_acc", train.total_acc.cast("float"))\
    .withColumn("credit_length_in_years",
train.credit_length_in_years.cast("float"))\
    .withColumn("int_rate", regexp_replace("int_rate", "%",
""").cast("float"))\
    .withColumn("remain", train.remain.cast("float"))\
    .withColumn("issue_year", train.issue_year.cast("float"))\
    .withColumn("phi_loan_amnt", train.phi_loan_amnt.cast("float"))\
    .withColumn("phi_emp_length",
train.phi_emp_length.cast("float"))\
    .withColumn("phi_annual_inc", train.phi_annual_inc.cast("float"))\
    .withColumn("phi_dti", train.phi_dti.cast("float"))\
    .withColumn("phi_delinq_2yrs",
train.phi_delinq_2yrs.cast("float"))\
    .withColumn("phi_revol_util", regexp_replace("phi_revol_util", "%",
""").cast("float"))\
    .withColumn("phi_total_acc", train.phi_total_acc.cast("float"))\
    .withColumn("phi_credit_length_in_years",
train.phi_credit_length_in_years.cast("float"))\
    .withColumn("phi_int_rate", regexp_replace("phi_int_rate", "%",
""").cast("float"))\
    .withColumn("CRI", train.CRI.cast("float"))\
```


Appendixes

```
.withColumn("train_flag", train.train_flag.cast("float"))
valid = valid.withColumn("loan_amnt", valid.loan_amnt.cast("float"))\
    .withColumn("emp_length", valid.emp_length.cast("float"))\
    .withColumn("annual_inc", valid.annual_inc.cast("float"))\
    .withColumn("dti", valid.dti.cast("float"))\
    .withColumn("delinq_2yrs", valid.delinq_2yrs.cast("float"))\
    .withColumn("revol_util", regexp_replace("revol_util", "%",
""").cast("float"))\
    .withColumn("total_acc", valid.total_acc.cast("float"))\
    .withColumn("credit_length_in_years",
valid.credit_length_in_years.cast("float"))\
    .withColumn("int_rate", regexp_replace("int_rate", "%",
""").cast("float"))\
    .withColumn("remain", valid.remain.cast("float"))\
    .withColumn("issue_year", valid.issue_year.cast("float"))\
    .withColumn("phi_loan_amnt", valid.phi_loan_amnt.cast("float"))\
    .withColumn("phi_emp_length",
valid.phi_emp_length.cast("float"))\
    .withColumn("phi_annual_inc", valid.phi_annual_inc.cast("float"))\
    .withColumn("phi_dti", valid.phi_dti.cast("float"))\
    .withColumn("phi_delinq_2yrs",
valid.phi_delinq_2yrs.cast("float"))\
    .withColumn("phi_revol_util", regexp_replace("phi_revol_util", "%",
""").cast("float"))\
    .withColumn("phi_total_acc", valid.phi_total_acc.cast("float"))\
    .withColumn("phi_credit_length_in_years",
valid.phi_credit_length_in_years.cast("float"))\
    .withColumn("phi_int_rate", regexp_replace("phi_int_rate", "%",
""").cast("float"))\
    .withColumn("CRI", valid.CRI.cast("float"))\
    .withColumn("train_flag", valid.train_flag.cast("float"))
train.registerTempTable("train")
train.write.parquet('AA_DFW_ALL.parquet', mode='overwrite')
```

Appendixes

```
valid.registerTempTable("valid")
valid.write.parquet('AA_DFW_ALL.parquet', mode='overwrite')
print(" >>>>>>> " + str(train.count())+ " loans opened by TRAIN data_set for
model training!")

print(" >>>>>>> " + str(valid.count())+ " loans opened by VALID data_set for
model validation!")

print(" == imbalance of the loan train and valid datasets ==")

print(" >>>>>>> Train dataset: " +
str(train.groupby('default_loan').count().collect()))

print(" >>>>>>> Test dataset: " +
str(valid.groupby('default_loan').count().collect()))

# Set the response and predictor variables and set up regression models
with train and test datasets.

Y = "default_loan"

categoricals = ["phi_term_month", "home_ownership", "purpose",
"addr_state", "verification_status", "application_type"]

numerics = ["CRI", "phi_loan_amnt", "phi_emp_length", "phi_annual_inc",
"phi_dti", "phi_delinq_2yrs", "phi_revol_util", "phi_total_acc",
"phi_credit_length_in_years", "phi_int_rate"]

X = categoricals + numerics

%sh

/databricks/python/bin/pip install plotnine matplotlib==2.2.2

%sh

/databricks/python/bin/pip install PyPI mlflow[extras]

# (1) define the model function

# to build Grid of GLM models and Standardization + CrossValidation

import sklearn.metrics as metrics

import pandas as pd

from plotnine import *

from plotnine.data import meat

from mizani.breaks import date_breaks

from mizani.formatters import date_format

from pyspark.ml import Pipeline
```

Appendixes

```
from pyspark.ml.feature import StandardScaler, StringIndexer,
OneHotEncoder, Imputer, VectorAssembler

from pyspark.ml.classification import LogisticRegression

from pyspark.ml.evaluation import BinaryClassificationEvaluator

from pyspark.ml.tuning import CrossValidator, ParamGridBuilder

import mlflow

import mlflow.spark

from pyspark.mllib.evaluation import BinaryClassificationMetrics

from pyspark.ml.linalg import Vectors

# setting the parameters

maxIter = 10

## we start with mlflow.start_run() which essentially start tracking what
we are doing in this notebook in databricks

with mlflow.start_run():

    labelCol = "default_loan"

    indexers = list(map(lambda c: StringIndexer(inputCol=c,
outputCol=c+"_idx", handleInvalid = "keep"), categoricals))

    ohes = list(map(lambda c: OneHotEncoder(inputCol=c + "_idx",
outputCol=c+"_class"), categoricals))

    imputers = Imputer(inputCols = numerics, outputCols = numerics)

    featureCols = list(map(lambda c: c+"_class", categoricals)) + numerics

    model_matrix_stages = indexers + ohes + \

        [imputers] + \

        [VectorAssembler(inputCols=featureCols,
outputCol="features"), \

        StringIndexer(inputCol= labelCol, outputCol="label")]

    scaler = StandardScaler(inputCol="features",

        outputCol="scaledFeatures",

        withStd=True,

        withMean=True)
```

Appendixes

```
## here, we build the logistic regression model with parameters equal to
variables for elasticNet regression

lr = LogisticRegression(maxIter=maxIter, featuresCol = "scaledFeatures")

# Create parameter grid

params = ParamGridBuilder() \
    .addGrid(lr.regParam, [0.01, 0.1, 0.3, 1.0, 10.0]) \
    .addGrid(lr.elasticNetParam, [0.0, 0.5, 0.99]) \
    .build()

##now, we define a pipeline which includes everything from standardizing
the data, imputing missing values and encoding for categorical columns

pipeline = Pipeline(stages=model_matrix_stages+[scaler]+[lr])

cv=CrossValidator(estimator=pipeline, estimatorParamMaps=params,
evaluator=BinaryClassificationEvaluator(), numFolds=10)

glm_model = cv.fit(train)

## Log Params and Model

## The important part for mlflow of model tracking and reproduceability of
the input parameters that we may want to review and take an action.

mlflow.log_param("algorithm", "SparkML_GLM_regression") # we put a
name for the algorithm that we used

mlflow.log_param("regParam", regParam)

mlflow.log_param("maxIter", maxIter)

mlflow.log_param("elasticNetParam", elasticNetParam)

mlflow.spark.log_model(glm_model, "glm_model")      # here we log
the model itself

##Evaluate and Log ROC Curve

lr_summary = glm_model.stages[len(glm_model.stages)-1].summary

roc_pd = lr_summary.roc.toPandas()

fpr = roc_pd["FPR"]

tpr = roc_pd["TPR"]

roc_auc = metrics.auc(roc_pd["FPR"], roc_pd["TPR"])

## Set Max F1 Threshold (for predicting the loan default with a balance
between true-positives and false-positives)

fMeasure = lr_summary.fMeasureByThreshold
```

Appendixes

```
maxFMeasure = fMeasure.groupBy().max("F-Measure").select("max(F-
Measure)").head()

madFMeasure = maxFMeasure["max(F-Measure)"]

fMeasure = fMeasure.toPandas()

bestThreshold = float ( fMeasure[ fMeasure["F-Measure"] ==
maxFMeasure] ["threshold"])

lr.setThreshold(bestThreshold)

## Evaluate and Log Metrics (here we score the customers)

def extract(row):

    return (row.remain,) + tuple(row.probability.toArray().tolist()) +
(row.label,) + (row.prediction,)

def score(model,data):

    pred = model.transform(data).select("remain", "probability", "label",
"prediction")

    pred = pred.rdd.map(extract).toDF(["remain", "p0", "p1", "label",
"prediction"])

    return pred

def auc(pred):

    metric = BinaryClassificationMetrics(pred.select("p1", "label").rdd)

    return metric.areaUnderROC

glm_train = score(glm_model, train)
glm_valid = score(glm_model, valid)
glm_train.registerTempTable("glm_train")
glm_valid.registerTempTable("glm_valid")
print( "GLM Training AUC :" + str( auc(glm_train)))
print( "GLM Validation AUC :" + str(auc(glm_valid)))

## here we log the auc values and the area under the curve for the models
metrics as we defined before for training as well as validation dataset

mlflow.log_metric("train_auc", auc(glm_train))

mlflow.log_metric("valid_auc", auc(glm_valid))

pandas_df = glm_valid.toPandas()
```

Appendixes

```
pd.crosstab(pandas_df.label, pandas_df.prediction,
values=pandas_df.remain, aggfunc="count").round(2)

##Evaluate and Log ROC Curve

glm_model2 = pipeline.fit(valid)

lr_summary = glm_model2.stages[len(glm_model2.stages)-1].summary

roc_pd = lr_summary.roc.toPandas()

fpr2 = roc_pd["FPR"]

tpr2 = roc_pd["TPR"]

roc_auc = metrics.auc(roc_pd["FPR"], roc_pd["TPR"])

# Now, we use matplotlib to draw the AUC of the model

import matplotlib.pyplot as plt

plt.title("Receiver Operating Characteristic (ROC)")

plt.plot(fpr, tpr, "b", label = "AUC = %0.2f" % roc_auc)

plt.legend(loc = "lower right")

plt.plot([0, 1], [0, 1], "r--")

plt.xlim([0, 1])

plt.ylim([0, 1])

plt.ylabel("True Positive Rate")

plt.xlabel("False Positive Rate")

display(plt.show())

display(glm_valid.groupby("label",
"prediction").agg((sum(col("remain")))).alias("sum_net")))

# SUPPORT VECTOR MACHINE

# required changes for Linear support vector machine

lsvc = LinearSVC(maxIter=maxIter)

params = ParamGridBuilder() \

    .addGrid(lsvc.regParam, [0.1, 0.99, 10, 100]) \

    .build()

# Chain indexer and lsvc in a Pipeline

#now, we define a pipeline which includes everything from standardizing
the data, imputing missing values and encoding for categorical columns
```

Appendixes

```
pipeline_lsvc = Pipeline(stages=model_matrix_stages+[lsvc])
# Train model. This also runs the indexer.

cv = CrossValidator(estimator=pipeline_lsvc,
estimatorParamMaps=params, evaluator=BinaryClassificationEvaluator(),
numFolds=10)

lsvc_model = cv.fit(train)
def extract(row):
    return (row.remain,) + (row.label,) + (row.prediction,)
def score(model,data):
    pred = model.transform(data).select("remain", "label", "prediction")
    pred = pred.rdd.map(extract).toDF(["remain", "label", "prediction"])
    return pred
def auc(pred):
    metric = BinaryClassificationMetrics(pred.select("prediction", "label").rdd)
    return metric.areaUnderROC

## Evaluate and Log Metrics (here we score the customers)
lsvcm_train = score(lsvc_model, train)
lsvcm_valid = score(lsvc_model, valid)
print( "lsvcM Training AUC :" + str( auc(lsvcm_train)))
print( "lsvcM Validation AUC :" + str(auc(lsvcm_valid)))

lsvcm_valid = lsvc_model.transform(valid).select("remain", "label",
"prediction")
# lsvcm_valid= pred.rdd.map(extract).toDF(["remain", "label",
"prediction"])
pandas_df = lsvcm_valid.toPandas()
pd.crosstab(pandas_df.label, pandas_df.prediction,
values=pandas_df.remain, aggfunc="count").round(2)
display(lsvcm_valid.groupBy("label",
"prediction").agg((sum(col("remain"))).alias("sum_net"))))

# RANDOM FORESTS
# required changes for Random forests
```

Appendixes

```
# here, we define a RF model.

rf = RandomForestClassifier(labelCol="label", featuresCol="features")

params = ParamGridBuilder() \
    .addGrid(rf.numTrees, [3, 5, 10]) \
    .build()

# Chain indexer and RF in a Pipeline

#now, we define a pipeline which includes everything from standardizing
the data, imputing missing values and encoding for categorical columns

pipeline_rf = Pipeline(stages=model_matrix_stages+[rf])

# Train model. This also runs the indexer.

cv = CrossValidator(estimator=pipeline_rf,
estimatorParamMaps=params, evaluator=BinaryClassificationEvaluator(),
numFolds=10)

rf_model = cv.fit(train)

## Evaluate and Log Metrics (here we score the customers)

def extract(row):

    return (row.remain,) + tuple(row.probability.toArray().toList()) +
(row.label,) + (row.prediction,)

def score(model,data):

    pred = model.transform(data).select("remain", "probability", "label",
"prediction")

    pred = pred.rdd.map(extract).toDF(["remain", "p0", "p1", "label",
"prediction"])

    return pred

def auc(pred):

    metric = BinaryClassificationMetrics(pred.select("p1", "label").rdd)

    return metric.areaUnderROC

## Evaluate and Log Metrics (here we score the customers)

rfm_train = score(rf_model, train)

rfm_valid = score(rf_model, valid)

rfm_train.registerTempTable("rfm_train")

rfm_valid.registerTempTable("rfm_valid")
```


Appendixes

```
print( "RFM Training AUC :" + str( auc(rfm_train)))
print( "RFM Validation AUC :" + str(auc(rfm_valid)))
## here we log the auc values and the area under the curve for the models
metrics as we defined before for training as well as validation dataset

mlflow.log_metric("train_auc", auc(rfm_train))
mlflow.log_metric("valid_auc", auc(rfm_valid))
pandas_df = rfm_valid.toPandas()
pd.crosstab(pandas_df.label, pandas_df.prediction,
values=pandas_df.remain, aggfunc="count").round(2)
display(rfm_valid.groupby("label",
"prediction").agg((sum(col("remain"))).alias("sum_net"))))
```

APPENDIX-2 CORRECTIONS IN THE DATASET.

Corrections in the dataset

- Name of countries:
 - Curaçao is changed to Curaçao;
 - Falkland Islands (Malvinas) is changed to Falkland Islands [Islas Malvinas];
 - North Macedonia is changed to North Macedonia [FYROM];
 - Republic of Korea is changed to Republic of Korea (South);
 - Reunion is changed to Réunion;
 - Saint Helena ex. dep. is changed to Saint Helena;
 - United Kingdom of Great Britain and Northern Ireland is changed to The United Kingdom;
 - Venezuela (Bolivarian Republic of) is changed to Venezuela;
 - Wallis and Futuna Islands is changed to Wallis and Futuna;
 - Aland Islands was not found in death table. We do not have data for proportion of respiratory disease in this region.
- The following countries report for less than eight years and we did not consider their time series.

Albania; Bahrain; Barbados; Bosnia and Herzegovina; Brazil; Brunei; Georgia; Panama; Saint Lucia; Seychelles; Tajikistan; Trinidad and Tobago; Uruguay; Uzbekistan; Mongolia; Saint Vincent and the Grenadines; Venezuela.

- Turkey has data for less than eight years; however, it has data for the most last years. As a result, it could be included in the research.
- Kazakhstan and Russian Federation were removed because they did not report data for recent years up to 2016.
- Death for China was not reported in the UNdata. As a result, china is not in our final data set.

APPENDIX-3 SOFTWARE R CODE FOR PROPOSED ENSEMBLE LEARNING STRATEGY FOR PANEL TIME- SERIES FORECASTING – CHAPTER IV.

```
# -----  
# Forecasting Deaths of respiratory diseases using Seasonal Time Series Methods  
# -----  
# 1.    Preparing the dataset - choosing countries with high completeness and good  
quality of data. Calculating the proportion and number of deaths for respiratory  
infections.  
# 2.    Estimating the missing values  
# 3.    Adopting the program to change the start date for each country and not using  
the missing values imputation method to go backward. (It decreases the accuracy and  
increase the probability of overfitting)  
# 4.    Applying our proposed min-max accuracy measure to remove dynamically  
inappropriate models for each country. Therefore, the prediction accuracy of the  
ensemble method will increase in some cases 10 times.  
# 5.    Using our proposed methodology to extract the frequency of models  
contribution in the ensemble. The best situation will happen, when the contribution of  
models are equal. It means that if a model is not recognized as a good predictor for  
some countries but it shows a good performance for other countries.  
# 6.    Making the program dynamic to test a set of holdouts and choose the best  
holdout for each method and each country for calculating the ensemble model.  
# 7.    Finally, calibrating the parameters of the models for the best performance. It  
changed the rank of models based on our min-max accuracy measure, and some worst  
models improved so much. Therefore, the balance in models contribution to our final  
ensemble model is increased.  
# -----  
# R software preparation  
# -----  
rm(list = ls(all = TRUE))  
graphics.off()  
close.screen(all = TRUE)  
erase.screen()  
windows.options(record = TRUE)  
options(digits = 10)  
# Check if the Rtools40 (gcc 8.3.0) is installed. We need to compile R packages from  
source that  
# contain C/C++/Fortran. By default, R for Windows installs the precompiled binary  
packages from CRAN, for which you do not need rtools. Run the following command  
and the result should be "C:\\rtools40\\usr\\bin\\make.exe". If not, install last version of  
Rtools and follow the instruction to make a text file .Renviron in your Documents folder
```

Appendixes

which contains the line: `PATH = "${RTOOLS40_HOME}\usr\bin;${PATH}"` . Save it in the document folder.

```
Sys.which("make")
# Please change to your directory
setwd('~')
library(pacman)
if (!require("pacman"))
  install.packages("pacman", type = "source")
if (!require("pacman"))
  install.packages("imputeTS", type = "source")
# -----
# Parameter Specification
# -----
LastYear = 2016      # Last year available in the database
clevel = 0.95        # Confidence Level
YearMin = 2000       # First year considered in the estimation sample
TargetYr = 2020      # Target year
holdout_set=c(3,4,5,6,7,8,9,10) # holdout period
Ens.crit = 'SMAPE'    # Criteria for computing model weights
set.seed(6121974)
mod.names                                     <-
c('SNAIVE','RWF','HWA','HWM','ETS','ARIMA','TBATS','STL','NNAR','MLP','EL
M','SSA','ENS')
# -----
# Functions
# -----
# Function-1
# Out-of-sample Goodness-of-fit forecasting measures (Validation Period)
fac.fun <- function (act, pred, predUB=0, predLB=0){
  AE <- ae(act, pred)
  APE <- ape(act, pred)
  bias <- bias(act, pred)
  CE <- ce(act, pred)
  MAE <- mae(act, pred)
  MAPE <- mape(act, pred)
  MASE <- mase(act, pred, step_size = 12)
  MDAE <- mdae(act, pred)
  MSE <- mse(act, pred)
  Pbias <- percent_bias(act, pred)
  RAE <- rae(act, pred)
  RMSE <- rmse(act, pred)
  RRSE <- rrse(act, pred)
  RSE <- rse(act, pred)
```

Appendixes

```
SE <- se(act, pred)
SMAPE <- smape(act, pred)
SSE <- sse(act, pred)
LAD <- max(abs(act-pred))
SAD <- min(abs(act-pred))
CFE <- sum(abs(act-pred))
CPFE <- (sum(abs(act-pred))/sum(act))*100
CVRMSE <- RMSE/mean(act)
CICount <- length(which((act-predUB)>0))+length(which((act-predLB)<0))
fac <- rbind(bias=bias, CE=CE, MAPE=MAPE, MASE=MASE, MDAE=MDAE,
            MSE=MSE, MSLE=MSLE, Pbias=Pbias, RAE=RAE, RMSE=RMSE,
            RMSLE=RMSLE,
            RRSE=RRSE, RSE=RSE, SMAPE=SMAPE, SSE=SSE, LAD=LAD,
            SAD=SAD, CFE=CFE, CPFE=CPFE, CVRMSE=CVRMSE,
            CICount=CICount)
return(list(AE=AE, APE=APE, SE=SE, SLE=SLE, fac=fac))}
# -----
# Function-2
model_weights <- function(error) {
  pr <- error/max(error)
  exp(-abs(pr))/sum(exp(-abs(pr)))}
# -----
# Function-3
calculate_summaries <- function(m, weight, cl=0.05) {
  qs <- quantile(m, probs = c(cl/2, 0.5, 1-cl/2), na_rm=TRUE)
  data.frame(
    mean = mean(m),
    median = qs[2],
    sd = sd(m),
    lb = qs[1],
    ub = qs[3],
    wgt.mean = m %*% weight)}
# -----
# Country cycle START - Missing values imputation - Model selection
# -----
mod_exc_list <- NULL
mod_exc <- NULL
r=1 # reset for first loop
ho = min(holdout_set) # reset for second loop
pdf(file = "D:/0-Corona virus/PAPER3-FORECASTING/R/output/plots.pdf",
    width = 4,
    height = 4)
```

Appendixes

```
##### START: Loop of countries (first
loop)
for (r in 3:nc){
  dxt <- NULL      # It is necessary to reset dxt for new country, if not, new models
  # will be added to the models of last country
  ENS <- NULL
  par(mfrow=c(1, 1))
  cnt <- country[r]
  UNData$Year = as.numeric(as.character(UNData$Year))
  YearMin = 2000
  # -----
  # Start: Missing values imputation for country r in the loop
  # -----
  UNData$death_respiratory = as.numeric(as.character(UNData$death_respiratory))
  cnt.data <- subset(UNData, Country==cnt & Year>=YearMin
    & Month!='Total'
    & Month!='January - March'
    & Month!='April - June'
    & Month!='July - September'
    & Month!='October - December'
    & Month!='Unknown')
  cnt.data <- cnt.data[with(cnt.data, order(Year,Month_no)), ]
  ymin <- min(cnt.data$Year)
  ymax <- max(cnt.data$Year)
  YearMin = ymin
  print("//////////\\\\\\\\\\\\\\\\")
  print(paste(cnt, ' | ', r, '/', nc, ' | ', ymin, '-', ymax, sep=""))
  print("-----")
  if (dim(subset(cnt.data, Year==ymin))[1]<12) {ymin=ymin+1}
  if (dim(subset(cnt.data, Year==ymax))[1]<12) {ymax=ymax-1}
  cnt.ts <- ts(cnt.data$death_respiratory, start=c(ymin,1),
    end=c(ymax,12),frequency=12)
  imp.seasplit <- na_seasplit(cnt.ts)
  plotNA.imputations(cnt.ts, imp.seasplit, legend = TRUE, main = paste(cnt, ' | ', ymin, '-',
    'ymax, ' Imputation', sep="),
    ylab = "Number of death")
  statsNA(cnt.ts)
  stl_imp = stl(imp.seasplit, "periodic")
  plot(stl_imp, main = paste(cnt, ' | ', ymin, '-', ymax, ' | ' Time series decomposition',
    sep="))
  par(mfrow=c(2,2))
```

Appendixes

```
trend_stl_imp <- stl_imp$time.series[,2]
plot(as.ts(trend_stl_imp), main = "Main time series", ylab="", xlab="")
trend_imp = ma(imp.seasplit, order = 12, centre = T)
detrend_imp = imp.seasplit / trend_imp
plot(as.ts(detrend_imp), main = "Detrend", ylab="", xlab="")
m_imp = t(matrix(data = detrend_imp, nrow = 12))
seasonal_imp = colMeans(m_imp, na.rm = T)
plot(as.ts(rep(seasonal_imp,12)), main = "Seasonality", ylab="", xlab="")
random_imp = imp.seasplit / (trend_imp * seasonal_imp)
plot(as.ts(random_imp), main = "Random noise", ylab="", xlab="")
#####
# Training & test sets
counter = 1
err1 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err2 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err3 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err4 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err5 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err6 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err7 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err8 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err9 <- array(0, dim=c(ncrit,nm,nc), dimnames = list(crit.names,mod.names,country))
err10 <- array(0, dim=c(ncrit,nm,nc), dimnames =
list(crit.names,mod.names,country))
err11 <- array(0, dim=c(ncrit,nm,nc), dimnames =
list(crit.names,mod.names,country))
err12 <- array(0, dim=c(ncrit,nm,nc), dimnames =
list(crit.names,mod.names,country))
SNAIVEls <- as.list( NULL )
RWFls <- as.list( NULL )
ETSls <- as.list( NULL )
HWAls <- as.list( NULL )
HWMls <- as.list( NULL )
ARIMAls <- as.list( NULL )
STLls <- as.list( NULL )
NNARls <- as.list( NULL )
TBATSls <- as.list( NULL )
MLPls <- as.list( NULL )
ELMls <- as.list( NULL )
SSAls <- as.list( NULL )
SGls <- as.list( NULL )
##### START: Loop of Holdout set (second
loop)
```

Appendixes

```
# Holdout loop : fitting the models for all values of holdout_set
for (ho in holdout_set) {
  if ( (ymax-ho+1) < ymin+3 ) { break } # This condition will terminate the second loop
  if holdout is
      # not appropriate for the min year of time series. It will save
      # enough degrees of freedom and length for models and
  guarantee

      # the continuance of program.
  tic() # start of run-time calculation
  train <- window(imp.seasplit, start=c(ymin,1), end=c(ymax-ho,12))
  if (calend.adj == 'Y') { train <- train/monthdays(train)}
  test <- window(imp.seasplit, start=c(ymax-ho+1,1))
  ht <- length(test)      # length of the test set
  h <- (TargetYr-ymax+ho)*12 # Forecasting horizon in months
  mdays <- 1
  if (calend.adj == 'Y') {
    mdays <- monthdays(test)}
  print(paste("***** For holdout:",ho,"*****", sep = " "))
  #####
  # Model fitting
  STLs[[ho]] <- stlf(train, h, lambda="auto", biasadj=TRUE, robust = FALSE, t.window
= 6, s.window = 6)
  print(paste('(r,/,nc, ) ', cnt,": Model stlf for holdout ",ho," is passed!", sep = ""))
  checkresiduals(STLs[[ho]], plot = FALSE)

  SNAIVEs[[ho]] <- snaive(train, drift=F, lambda=0, level=clevel, biasadj=TRUE, h=h)
  print(paste('(r,/,nc, ) ', cnt,": Model snaive for holdout ",ho," is passed!", sep = ""))
  checkresiduals(SNAIVEs[[ho]], plot = FALSE)
  ARIMAs[[ho]] <- auto.arima(train, lambda=0, biasadj=TRUE)
  print(paste('(r,/,nc, ) ', cnt,": Model arima for holdout ",ho," is passed!", sep = ""))
  checkresiduals(ARIMAs[[ho]], plot = FALSE)
  ETSs[[ho]] <- ets(train, model = "ZAA",lambda="auto", ic = "bic", restrict = TRUE,
    allow.multiplicative.trend = TRUE)
  print(paste('(r,/,nc, ) ', cnt,": Model ETS for holdout ",ho," is passed!", sep = ""))
  checkresiduals(ETSs[[ho]], plot = FALSE)
  TBATs[[ho]] <- tbats(train, biasadj=TRUE)
  print(paste('(r,/,nc, ) ', cnt,": Model TBATS for holdout ",ho," is passed!", sep = ""))
  HWMs[[ho]] <- hw(train, h, seasonal = c('multiplicative'), level=clevel)
  print(paste('(r,/,nc, ) ', cnt,": Model HWM for holdout ",ho," is passed!", sep = ""))
  checkresiduals(HWMs[[ho]], plot = FALSE)
  HWAs[[ho]] <- hw(train, h, seasonal = c('additive'), level=clevel)
  print(paste('(r,/,nc, ) ', cnt,": Model HWA for holdout ",ho," is passed!", sep = ""))
  checkresiduals(HWAs[[ho]], plot = FALSE)
```


Appendixes

```

RWFls[[ho]] <- rwf(train, drift=F, h, lambda="auto", level=clevel, biasadj=TRUE)
print(paste('(r,/,nc, ) ', cnt,": Model Random Walk Forecasts (RWF) for holdout
",ho," is passed!", sep = ""))
checkresiduals(RWFls[[ho]], plot = FALSE)
ELM <- forecast(elm(train, type=c("lasso"), hd=500 ,comb=c("mean"), reps = 200,
difforder=NULL, h=h, level=clevel, set.lag = TRUE, allow.det.season,
det.type = "auto"))
ELMls[[ho]] <- elm(train, type=c("lasso"), hd=500 ,comb=c("mean"), reps = 200,
difforder=NULL)
print(paste('(r,/,nc, ) ', cnt,": Model Extreme Learning Machines (ELM) for holdout
",ho," is passed!", sep = ""))
MLPls[[ho]] <- mlp(train, comb='mode', hd.auto.type='valid', hd.max = 5)
print(paste('(r,/,nc, ) ', cnt,": Model Multilayer Perceptron (MLP) for holdout ",ho,"
is passed!", sep = ""))
NNARls[[ho]] <- nnetar(train, P = 2, size = 1, decay=0.001, lambda="auto", repeats =
100, MaxNWts=2000)
print(paste('(r,/,nc, ) ', cnt,": Model NNETAR for holdout ",ho," is passed!", sep =
""))
SSAls[[ho]] <- ssa(train, kind="1d-ssa", svd.method="auto", L=12, neig = NULL,
force.decompose = TRUE, mask = NULL)
print(paste('(r,/,nc, ) ', cnt,": Model Sigular spectrum analysis (SSA) for holdout
",ho," is passed!", sep = ""))
SSA2 <- forecast(SSAls[[ho]], h=h, groups = list(1:6), bootstrap = TRUE, len = 48,
R = 10, method = c("vector"), interval = "prediction",
level=clevel, only.intervals = TRUE,drop = TRUE, drop.attributes = FALSE,
cache = TRUE)
ts <- forecast(SSAls[[ho]], h=h)
print(ggplot2::autoplot(imp.seasplit) +
autolayer(SSA2,series="SSA", PI=FALSE) +
xlab("Year") + ylab("counts") +
ggtitle(paste(cnt, "Monthly Deaths", sep=': ')))
Rssa::ssa.capabilities(SSAls[[ho]])
SGls[[ho]] <- savgolay(train, width = 4, degree = 2)
print(paste('(r,/,nc, ) ', cnt,": Model Savitzgy-Golay Smoothing (savgolay) for
holdout ",ho," is passed!", sep = ""))
checkresiduals(SGls[[ho]], plot = FALSE)
SG[[ho]] <- ts(SG[[ho]], start=c(ymin,1), end=c(ymax,12),frequency=12)
#####
#Model selecting (best forecasting according to different holdout values)
ETSselect <- forecast(ETSls[[ho]], bootstrap = TRUE, simulate = TRUE,h=h,
level=clevel, biasadj=TRUE)
accuracy(train, ETSselect)
ARIMAselect <- forecast(ARIMAls[[ho]], h=h)

```

Appendixes

```
accuracy(train, ARIMAselect)
NNARselect <- forecast(NNARls[[ho]], PI=T, level=clevel, h=h, npaths=nsim)
accuracy(train, NNARselect)
TBATSselect <- forecast(TBATSls[[ho]], h=h)
accuracy(train, TBATSselect)
MLPselect <- forecast(MLPls[[ho]], h=h, level=clevel)
accuracy(train, MLPselect)
ELMselect <- forecast(ELMls[[ho]], h=h, level=clevel)
accuracy(train, ELMselect)
SSAselect <- forecast(SSAls[[ho]], h=h, groups = list(1:6), bootstrap = TRUE, len = h,
R = 10, method = c("vector"), interval = "prediction",
                level=clevel, only.intervals = TRUE, drop = TRUE, drop.attributes =
FALSE, cache = TRUE)
accuracy(train, SSAselect)
if ( counter == 1 ) {
SNAIVE <- SNAIVEls[[ho]]
RWF  <- RWFls[[ho]]
ETS  <- ETSselect
HWA  <- HWAls[[ho]]
HWM  <- HWMls[[ho]]
ARIMA <- ARIMAselect
STL  <- STLls[[ho]]
NNAR  <- NNARselect
TBATS <- TBATSselect
MLP   <- MLPselect
ELM   <- ELMselect
SSA   <- SSAselect
ho_SNAIVE<-ho_RWF<-ho_HWA<-ho_HWM<-ho_ETS<-ho_ARIMA<-
ho_NNAR<-ho_MLP<-ho_ELM<-ho_SSA<-ho_TBATS<-ho_STL<-ho}
print(paste("***** ", '(',r,',',nc, ') ', cnt," - holdout:",ho," *****", sep = ""))
counter_1 <- (counter-1)
# selecting the best SNAIVE
fac.UN[,1,r] <- fac.fun(SNAIVEls[[ho]]$mean[1:ht]*mdays, test)$fac
err1[[counter]] <- fac.UN[Ens.crit,1,r]
if (counter_1 > 0) {
  if (err1[[counter]] < err1[[counter_1]]) {
    SNAIVE <- SNAIVEls[[ho]]
    ho_SNAIVE <- ho} }
# selecting the best RWF
fac.UN[,2,r] <- fac.fun(RWFls[[ho]]$mean[1:ht]*mdays, test)$fac
err2[[counter]] <- fac.UN[Ens.crit,2,r]
if (counter_1 > 0) {
  if (err2[[counter]] < err2[[counter_1]]) {RWF <- RWFls[[ho]]}
```

Appendixes

```
ho_RWF<-ho}
# selecting the best ETS
fac.UN[,3,r] <- fac.fun(ETSselect$mean[1:ht]*mdays, test)$fac #ETSselect is an
atomic vector and ETSselect[[ho]]
err3[[counter]] <- fac.UN[Ens.crit,3,r] # is a number. these
method_nameselect methods
if (counter_1 > 0) { # do not need [[ho]] like others.
  if (err3[[counter]] < err3[[counter_1]]) {ETS <- ETSselect}
  ho_ETS<-ho}
# selecting the best HWA
fac.UN[,4,r] <- fac.fun(HWAIs[[ho]]$mean[1:ht]*mdays, test)$fac
err4[[counter]] <- fac.UN[Ens.crit,4,r]
if (counter_1 > 0) {
  if (err4[[counter]] < err4[[counter_1]]) {HWA <- HWAIs[[ho]]}
  ho_HWA <- ho}
# selecting the best HWM
fac.UN[,5,r] <- fac.fun(HWMIls[[ho]]$mean[1:ht]*mdays, test)$fac
err5[[counter]] <- fac.UN[Ens.crit,5,r]
if (counter_1 > 0) {
  if (err5[[counter]] < err5[[counter_1]]) {HWM <- HWMIls[[ho]]}
  ho_HWM <- ho}
# selecting the best ARIMA
fac.UN[,6,r] <- fac.fun(ARIMAselect$mean[1:ht]*mdays, test)$fac
err6[[counter]] <- fac.UN[Ens.crit,6,r]
if (counter_1 > 0) {
  if (err6[[counter]] < err6[[counter_1]]) {ARIMA <- ARIMAselect}
  ho_ARIMA <- ho}
# selecting the best STL
fac.UN[,7,r] <- fac.fun(STLls[[ho]]$mean[1:ht]*mdays, test)$fac
err7[[counter]] <- fac.UN[Ens.crit,7,r]
if (counter_1 > 0) {
  if (err7[[counter]] < err7[[counter_1]]) {STL <- STLls[[ho]]}
  ho_STL <- ho}
# selecting the best NNAR
fac.UN[,8,r] <- fac.fun(NNARselect$mean[1:ht]*mdays, test)$fac
err8[[counter]] <- fac.UN[Ens.crit,8,r]
if (counter_1 > 0) {
  if (err8[[counter]] < err8[[counter_1]]) {NNAR <- NNARselect}
  ho_NNAR <- ho}
# selecting the best TBATS
fac.UN[,9,r] <- fac.fun(TBATSselect$mean[1:ht]*mdays, test)$fac
err9[[counter]] <- fac.UN[Ens.crit,9,r]
if (counter_1 > 0) {
```

Appendixes

```
    if (err9[[counter]] < err9[[counter_1]]) {TBATS <- TBATSselect}
    ho_TBATS <- ho}
# selecting the best MLP
fac.UN[,10,r] <- fac.fun(MLPselect$mean[1:ht]*mdays, test)$fac
err10[[counter]] <- fac.UN[Ens.crit,10,r]
if (counter_1 > 0) {
  if (err10[[counter]] < err10[[counter_1]]) {MLP <- MLPselect}
  ho_MLP <- ho}
# selecting the best ELM
fac.UN[,11,r] <- fac.fun(ELMselect$mean[1:ht]*mdays, test)$fac
err11[[counter]] <- fac.UN[Ens.crit,11,r]
if (counter_1 > 0) {
  if (err11[[counter]] < err11[[counter_1]]) {ELM <- ELMselect}
  ho_ELM <- ho}
# selecting the best SSA
fac.UN[,12,r] <- fac.fun(SSAselect$mean[1:ht]*mdays, test)$fac
err12[[counter]] <- fac.UN[Ens.crit,12,r]
if (counter_1 > 0) {
  if (err12[[counter]] < err12[[counter_1]]) {SSA <- SSAselect}
  ho_SSA <- ho}
SG <- savgolay(train, width = 4, degree = 2)
SG <- ts(SGls[[ho]], start=c(ymin,1), end=c(ymax,12),frequency=12)
counter = counter + 1 # counter will be reset to 1 before running the loop again
exectime <- toc(quiet = FALSE) # end of run-time calculation
run_time <- c(exectime$tic,exectime$toc, exectime$toc-exectime$tic )
exectime_name=paste(r,"_", "Run-
time", "_", cnt.data$Country[1], "ho", holdout, ".csv", sep = "")
write.csv(run_time, exectime_name) # save run-time
}
##### END: Loop of holdout loop
# -----
# START: The model ensemble
# -----
ho <- min(holdout_set) # different models are selected by different holdouts. This min
holdout makes the
test <- window(imp.seasplit, start=c(ymax-ho+1,1))
ht <- length(test) # set again the length of the test set for the model ensemble
h <- (TargetYr-ymax+ho)*12 # set again the Forecasting horizon in months for the
model ensemble
mdays <- 1
if (calend.adj == 'Y') {
  mdays <- monthdays(test)}
# Forecasting accuracy for final models
```

Appendixes

```
fac.UN[,1,r] <- fac.fun(SNAIVE$mean[1:ht]*mdays, test)$fac
fac.UN[,2,r] <- fac.fun(RWF$mean[1:ht]*mdays, test)$fac
fac.UN[,3,r] <- fac.fun(HWA$mean[1:ht]*mdays, test)$fac
fac.UN[,4,r] <- fac.fun(HWM$mean[1:ht]*mdays, test)$fac
fac.UN[,5,r] <- fac.fun(ETS$mean[1:ht]*mdays, test)$fac
fac.UN[,6,r] <- fac.fun(ARIMA$mean[1:ht]*mdays, test)$fac
fac.UN[,7,r] <- fac.fun(TBATS$mean[1:ht]*mdays, test)$fac
fac.UN[,8,r] <- fac.fun(STL$mean[1:ht]*mdays, test)$fac
fac.UN[,9,r] <- fac.fun(NNAR$mean[1:ht]*mdays, test)$fac
fac.UN[,10,r] <- fac.fun(MLP$mean[1:ht]*mdays, test)$fac
fac.UN[,11,r] <- fac.fun(ELM$mean[1:ht]*mdays, test)$fac
fac.UN[,12,r] <- fac.fun(SSA$mean[1:ht]*mdays, test)$fac
dxt1 <- cbind(SNAIVE$mean, RWF$mean, HWA$mean, HWM$mean, ETS$mean,
ARIMA$mean,
              TBATS$mean, STL$mean, NNAR$mean, MLP$mean, ELM$mean,
              SSA$mean)
# -----
# Start: Removing OUTLIER models - the high error models remove from ensemble
# model calculation
# -----
# START: calculation of the max and min for total error by using all error measures
max_error <-
as.numeric(max(fac.UN[Ens.crit,1,r],fac.UN[Ens.crit,2,r],fac.UN[Ens.crit,3,r],fac.UN
[Ens.crit,4,r],
              fac.UN[Ens.crit,5,r],fac.UN[Ens.crit,6,r],fac.UN[Ens.crit,7,r],fac.UN[Ens.crit,8,r],

fac.UN[Ens.crit,9,r],fac.UN[Ens.crit,10,r],fac.UN[Ens.crit,11,r],fac.UN[Ens.crit,12,r])
)
min_error <-
as.numeric(min(fac.UN[Ens.crit,1,r],fac.UN[Ens.crit,2,r],fac.UN[Ens.crit,3,r],fac.UN[
Ens.crit,4,r],

fac.UN[Ens.crit,5,r],fac.UN[Ens.crit,6,r],fac.UN[Ens.crit,7,r],fac.UN[Ens.crit,8,r],

fac.UN[Ens.crit,9,r],fac.UN[Ens.crit,10,r],fac.UN[Ens.crit,11,r],fac.UN[Ens.crit,12,r])
)
if (is.na(min_error)) { min_error <- 0}
id_error <- (min_error+max_error)/2
fac.UN[Ens.crit,1,r] fac.UN[Ens.crit,2,r] fac.UN[Ens.crit,3,r]
              fac.UN[Ens.crit,4,r]
fac.UN[Ens.crit,5,r] fac.UN[Ens.crit,6,r] fac.UN[Ens.crit,7,r]
              fac.UN[Ens.crit,8,r]
```

Appendixes

```

fac.UN[Ens.crit,9,r] fac.UN[Ens.crit,10,r] fac.UN[Ens.crit,11,r]
      fac.UN[Ens.crit,12,r]
mod.names <-
c('SNAIVE','RWF','HWA','HWM','ETS','ARIMA','TBATS','STL','NNAR','MLP','EL
M','SSA','ENS')
j=1 c=1
method_list = list()
for (j in 1:(nm-1)) {
  if (fac.UN[Ens.crit,j,r] > id_error) {
    print(paste("Attention: For country ", '(',r,',',nc, ')', cnt, ", ", mod.names[j], " model
is excluded!", sep = "")) #output: message included removed models
    mod_exc <- rbind(mod_exc, data.frame(cnt, mod.names[j])) #mod_exc: list of
countries-excluded models
    fac.UN[,j,r] <- 0 #error to zero to recognize the possible mistake in the loop.
  }
  else {method_list[j] <- mod.names[j]
    fac.UN[,c,r]=fac.UN[,j,r] #make fac.UN in order for further ENS calculations.
IMPORTANT: no more in order based on mod.names!!
    c=c+1}
  }
method_list <- method_list[-which(sapply(method_list, is.null))]
# to make a list of excluded models per country
ifelse (r == 1 ,mod_exc_list <- mod_exc, mod_exc_list <- cbind(mod_exc_list,
mod_exc))
#####
method_length <- length(method_list)
if (method_length < 1) { next }
dxt <- eval(parse(text = as.name(method_list[[1]])))$mean
j=2
for (j in 2:method_length-1) {
  dxt <- cbind(dxt,eval(parse(text = as.name(method_list[[j]])))$mean)
  }
#defining the names for columns
j=1
for (j in 1:method_length) {
  colnames(dxt)[j] <- paste("Column",j,":",method_list[[j]], sep = "" )
  }
# -----
# END: Removing the high error models from ensemble model calculation
# -----
# Ensemble model weights
method_list2 <- data.frame(method_list)

```

Appendixes

```
mw <- array(0, dim=c(nc,method_length,ncrit), dimnames = list(country, method_list,
crit.names)) for (k2 in 1:ncrit) {
  mw[r,1:method_length,k2] <- model_weights(fac.UN[k2,1:method_length,r])
}
SNAIVE_ENS <- window(SNAIVE$mean, start=c(ymax-ho+1,1))
RWF_ENS <- window(RWF$mean, start=c(ymax-ho+1,1))
HWA_ENS <- window(HWA$mean, start=c(ymax-ho+1,1))
HWM_ENS <- window(HWM$mean, start=c(ymax-ho+1,1))
ETS_ENS <- window(ETS$mean, start=c(ymax-ho+1,1))
ARIMA_ENS <- window(ARIMA$mean, start=c(ymax-ho+1,1))
TBATS_ENS <- window(TBATS$mean, start=c(ymax-ho+1,1))
STL_ENS <- window(STL$mean, start=c(ymax-ho+1,1))
NNAR_ENS <- window(NNAR$mean, start=c(ymax-ho+1,1))
MLP_ENS <- window(MLP$mean, start=c(ymax-ho+1,1))
ELM_ENS <- window(ELM$mean, start=c(ymax-ho+1,1))
SSA_ENS <- window(SSA$mean, start=c(ymax-ho+1,1))
# to check the dimensions, should be equal:
dim.data.frame(mw[r,k])
dim.data.frame(fac.UN[k,1:method_length,r])
# Forecasted monthly death counts & CI by country
SNAIVE.cnt[[cnt]] <- SNAIVE_ENS      RWF.cnt[[cnt]]      <- RWF_ENS
  HWA.cnt[[cnt]] <- HWA_ENS
HWM.cnt[[cnt]] <- HWM_ENS      ETS.cnt[[cnt]] <- ETS_ENS
ARIMA.cnt[[cnt]] <- ARIMA_ENS
TBATS.cnt[[cnt]] <- TBATS_ENS    STL.cnt[[cnt]] <- STL_ENS
NNAR.cnt[[cnt]] <- NNAR_ENS
MLP.cnt[[cnt]] <- MLP_ENS      ELM.cnt[[cnt]]      <- ELM_ENS
SSA.cnt[[cnt]] <- SSA_ENS
dxt_ENS <- na.omit(dxt)
min_len_model <- min(length(SNAIVE), length(RWF), length(HWA), length(HWM),
length(ETS), length(ARIMA),
length(TBATS), length(STL), length(NNAR), length(MLP),
length(ELM), length(SSA))
ifelse (length(SNAIVE) == min_len_model, model_with_min_len <- SNAIVE ,
ifelse (length(RWF) == min_len_model, model_with_min_len <- RWF ,
ifelse (length(HWA) == min_len_model, model_with_min_len <- HWA ,
ifelse (length(HWM) == min_len_model, model_with_min_len <- HWM ,
ifelse (length(ETS) == min_len_model, model_with_min_len <- ETS ,
ifelse (length(ARIMA) == min_len_model, model_with_min_len <- ARIMA ,
ifelse (length(TBATS) == min_len_model, model_with_min_len <- TBATS ,
ifelse (length(STL) == min_len_model, model_with_min_len <- STL ,
ifelse (length(NNAR) == min_len_model, model_with_min_len <- NNAR ,
model_with_min_len <- SSA)))))))))
```

Appendixes

```

h <- length(rownames(data.frame(dxt_ENS, stringsAsFactors = FALSE)))
month.names1 <- rownames(data.frame(model_with_min_len, stringsAsFactors =
FALSE))
month.names <- month.names1[(length(month.names1)-h+1):
(length(month.names1))]
ENS <- array(0, dim=c(h,6), dimnames = list(month.names,
c('mean','median','sd','lb','ub','Wmean')))
for (i in 1:h){
  ENS[i,] <- as.numeric(calculate_summaries(dxt_ENS[i,],mw[r,,Ens.crit]))}

ENS.cnt[[cnt]] <- ENS #output of ensemble model
fac.UN[,13,r] <- fac.fun(ENS[,6][1:ht]*mdays, test)$fac #error
# -----
# END: The model ensemble
# -----
# Final result: all models forecasts in dxt
dxt <- cbind(dxt_ENS, ENS=ENS[,6])
col_no = method_length+1
colnames(dxt)[col_no] <- paste("dxt.Column",col_no,":","ENS", sep = "" )
colnames(dxt) <- mod.names
dxt.cnt[[cnt]] <- dxt # FINAL OUTPUT for country r, ready to be saved in SAVE
RESULT section below.
# Plots
ENS_ts <- ts(ENS[,6], start=c((TargetYr-min_len_model),1),
end=c(TargetYr,12),frequency=12)
print(ggplot2::autoplot(imp.seasplit) +
  autolayer(SNAIVE,series="SNAIVE", PI=FALSE) +
  autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
  xlab("Year") + ylab("counts") +
  ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
  autolayer(RWF, series="RWF", PI=FALSE) +
  autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
  xlab("Year") + ylab("counts") +
  ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
  autolayer(HWA, series="HWA", PI=FALSE) +
  autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
  xlab("Year") + ylab("counts") +
  ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)

```


Appendixes

```
print(ggplot2::autoplot(imp.seasplit) +
      autolayer(HWM, series="HWM", PI=FALSE) +
      autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
      xlab("Year") + ylab("counts") +
      ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
      autolayer(ETS, series="ETS", PI=FALSE) +
      autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
      xlab("Year") + ylab("counts") +
      ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
      autolayer(ARIMA, series="ARIMA", PI=FALSE) +
      autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
      xlab("Year") + ylab("counts") +
      ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
      autolayer(TBATS, series="TBATS", PI=FALSE) +
      autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
      xlab("Year") + ylab("counts") +
      ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
      autolayer(STL, series="STL", PI=FALSE) +
      autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
      xlab("Year") + ylab("counts") +
      ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
      autolayer(NNAR, series="NNAR", PI=FALSE) +
      autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
      xlab("Year") + ylab("counts") +
      ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
      autolayer(MLP, series="MLP", PI=FALSE) +
      autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
      xlab("Year") + ylab("counts") +
      ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
```

Appendixes

```

    autolayer(ELM, series="ELM", PI=FALSE) +
    autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
    xlab("Year") + ylab("counts") +
    ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
  )
print(ggplot2::autoplot(imp.seasplit) +
  autolayer(SSA, series="SSA", PI=FALSE) +
  autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
  xlab("Year") + ylab("counts") +
  ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
print(ggplot2::autoplot(imp.seasplit) +
  autolayer(ENS_ts, series="ENS_ts", PI=FALSE) +
  xlab("Year") + ylab("counts") +
  geom_point()+
  ggtitle(paste(cnt, "Monthly Deaths", sep=': '))
)
# -----
-----
# Save full data & results
# -----
-----
file_name_country1=paste("MDeaths_UN","_",cnt.data$Country[1],".RData",sep =
"")
save.image(file=file_name_country1)
file_name_country2=paste(r,"_", "MODELS_ERRORS","_",cnt.data$Country[1],".cs
v",sep = "")
write.csv(fac.UN[,r], file_name_country2)
file_name_country3=paste(r,"_", "MODELS_EXCLUDED","_",cnt.data$Country[1],".
csv",sep = "")
write.csv(mod_exc, file_name_country3)
file_name_country4=paste(r,"_", "MODELS_PREDS","_",cnt.data$Country[1],".csv",
sep = "")
write.csv(dxt, file_name_country4)
file_name_country5=paste(r,"_", "MODELS_ENS","_",cnt.data$Country[1],".csv",se
p = "")
write.csv(ENS_ts, file_name_country5)
# -----
# Country cycle END
# -----
}
##### END: Loop of countries
dev.off()    # finish the output pdf file with all plots

```

Appendixes

```
write.csv(fac.UN[,r], "MODELS_ERRORS_TOTAL.csv")
write.csv(mod_exc, "MODELS_EXCLUDED_TOTAL.csv")
write.csv(dxt, "MODELS_PRED_TOTAL.csv")
write.csv(ENS_ts, "MODELS_ENS_TOTAL.csv")■
```



Data Science for Finance: Targeted Learning from (Big) Data for Economic Stability and Financial Risk Management
Afshin Ashofteh

PhD