UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ARTHUR RICARDO DE SOUSA VITÓRIA

Supporting Public Health Policy Decisions Through Live Birth Predictions for Health Regions of Goiás with Machine Learning

Goiânia
2023

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[x] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Arthur Ricardo de Sousa Vitoria

3. Título do trabalho

Supporting Public Health Policy Decisions Through Live Birth Predictions for Health Regions of Goiás with Machine Learning

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Arlindo Rodrigues Galvão Filho**, **Professor do Magistério Superior**, em 08/05/2023, às 10:50, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Arthur Ricardo De Sousa Vitoria**, **Discente**, em 08/05/2023, às 13:06, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3729611** e o código CRC **6B63CBCE**.

Referência: Processo nº 23070.016640/2023-16 SEI nº 3729611

ARTHUR RICARDO DE SOUSA VITÓRIA

Supporting Public Health Policy Decisions Through Live Birth Predictions for Health Regions of Goiás with Machine Learning

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.
Área de concentração: Ciência da Computação.
Linha de Pesquisa: Sistemas Inteligentes e Aplicações.

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho.

Goiânia
2023

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Vitoria, Arthur Ricardo Sousa
Supporting Public Health Policy Decisions Through Live Birth Predictions for Health Regions of Goiás with Machine Learning [manuscrito] / Arthur Ricardo Sousa Vitoria. - 2023.
XLV, 45 f.

Orientador: Prof. Arlindo Rodrigues Galvão Filho.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Cidade de Goiás, 2023.
Bibliografia.
Inclui lista de figuras.

1. Machine Learning. 2. Live Births Prediction. 3. Univariate Time Series. 4. Multivariate Time Series. I. Galvão Filho, Arlindo Rodrigues, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **04/2023** da sessão de Defesa de Dissertação de **Arthur Ricardo de Sousa Vitoria**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos onze dias do mês de abril de dois mil e vinte e três, a partir das dezoito e trinta horas, via sistema de webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Supporting Public Health Policy Decisions Through Live Birth Predictions for Health Regions of Goiás with Machine Learning"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Arlindo Rodrigues Galvão Filho (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Clarimar José Coelho (PUC/|GO), membro titular externo; Professor Doutor Anderson da Silva Soares (INF/UFG), membro titular interno. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Arlindo Rodrigues Galvão Filho, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos onze dias do mês de abril de dois mil e vinte e três.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Arlindo Rodrigues Galvão Filho**, **Professor do Magistério Superior**, em 11/04/2023, às 19:53, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **CLARIMAR JOSÉ COELHO**, **Usuário Externo**, em 11/04/2023, às 19:53, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Anderson Da Silva Soares**, **Professor do Magistério Superior**, em 11/04/2023, às 19:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Arthur Ricardo De Sousa Vitoria**, **Discente**, em 11/04/2023, às 20:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3659468** e o código CRC **0AE623A1**.

Referência: Processo nº 23070.016640/2023-16 SEI nº 3659468

Resumo

Vitória, Arthur. **Supporting Public Health Policy Decisions Through Live Birth Predictions for Health Regions of Goiás with Machine Learning**. Goiânia, 2023. 45p. Dissertação de Mestrado. Programa de Pós Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

Modelos de previsão de séries temporais estão se tornando cada vez mais comuns em aplicações de saúde e administração, pois podem ser ferramentas confiáveis de apoio à decisão. A taxa de nascidos vivos é um índice de saúde diretamente ligado à saúde materna e neonatal, e sua previsão pode ajudar gestores de saúde a antecipar recursos para serviços obstétricos e pediátricos. Assim, o objetivo deste trabalho é prever o número de nascidos vivos no estado de Goiás (Brasil) para um horizonte de 24 meses, fornecendo informações úteis para apoiar o planejamento e a implementação de políticas públicas. Este estudo investiga duas abordagens distintas: univariada e multivariada, permitindo uma melhor compreensão e gestão da hierarquia territorial brasileira. Ambas as abordagens são avaliadas com dados fornecidos pelo Sistema de Informação sobre Nascidos Vivos do Departamento de Informação do Sistema Único de Saúde (SINASC-DATASUS). O conjunto de dados é composto por 252 registros mensais do número de nascidos vivos para as 18 regiões de saúde de Goiás. Os resultados foram mensurados pela capacidade de previsão pelo Erro Percentual Médio Absoluto (*Mean Average Percentual Error*, MAPE) e Erro Médio Absoluto (*Mean Absolute Error*, MAE). Para a abordagem univariada utilizando a LMU, a média de MAPE e MAE alcançada foi de 6,4614 e 19,9136, respectivamente. A abordagem multivariada foi combinada com o método *K-means* para agrupar séries temporais similares usando o empenamento dinâmico do tempo (*dynamic time warping*) como medida de similaridade, gerando um resultado médio de 5,5985 e 18,1360 para MAPE e MAE, respectivamente.

Palavras–chave

<Aprendizado de Máquina, Predição de Nascidos Vivos, Séries Temporais Univariadas, Séries Temporais Multivariadas>

Abstract

Vitória, Arthur. **<Supporting Public Health Policy Decisions Through Live Birth Predictions for Health Regions of Goiás with Machine Learning>**. Goiânia, 2023. 45p. MSc. Dissertation. Programa de Pós Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

The use of forecasting models is becoming even more common in healthcare and administration applications because they can be reliable decision support tools. The live birth rate is a health index that is directly linked with maternal and newborn health, and its prediction can assist health managers to anticipate resources destined for obstetric and pediatric services. Thus, the objective of this work is to forecast the number of live births in the state of Goiás (Brazil) for a 24-month horizon, providing useful information to support the planning and implementation of public policies. This study investigates two distinct approaches: univariate and multivariate, allowing a better understanding and management of the Brazilian territorial hierarchy. Both approaches are evaluated with data provided by the information system on live births of the information department of the single health system (SINASC-DATASUS). The dataset is composed of 252 monthly records of the number of live births for the 18 health regions of Goiás. The results were measured in prediction ability by Mean Absolute Percentual Error (MAPE) and Mean Absolute Error (MAE). For the univariate approach using a LMU, the average MAPE and MAE achieved were 6.4614 and 19.9136, respectively. The multivariate approach was combined with the K-means method for clustering similar time series using a dynamic time warping measure, generating an average result of 5.5985 and 18.1360 for MAPE and MAE, respectively.

Keywords

<Machine Learning, Live Births Prediction, Univariate Time Series, Multivariate Time Series>

Table of Contents

List of Figures

CHAPTER 1

Introduction

The live birth rate is an important indicator of the population's health services, as it reflects the health and well-being of mothers and newborns. The sustainable development goals (SDGs) established by the United Nations (UN) are a global initiative to address a wide range of pressing global issues. One of the key issues addressed by the SDGs is the need to reduce global maternal mortality, which includes deaths related to complications of childbirth, pregnancy, and postpartum, to 70 deaths per 100,000 live births [WHO].

Maternal mortality can be classified into two categories: indirect and direct causes. Indirect causes are those that are related to preexisting diseases, which can be aggravated by physiological changes during pregnancy. Direct causes are related to intervention, omission, or inadequate treatments [Pícoli, Cazola e Lemos 2017]. Furthermore, as it is registered mostly in the population of developing countries, it is a global public health concern since most deaths are from direct causes and therefore preventable [BVS].

The Pan American Health Organization (PAHO) reported that globally, about 830 women die every day from direct causes related to pregnancy or childbirth [PAHO]. As the majority of deaths are from direct causes, maternal mortality is an important indicator of the quality of healthcare for women and is closely linked to access to quality services, reflecting inequalities that affect underdeveloped countries, where access to healthcare is often limited [Pinto et al. 2022].

The Brazilian Ministry of Health (MH) reports that the maternal mortality ratio (MMR) has remained persistently high. In recent years, MMR has seen a significant increase, rising from 57.9 deaths per 100,000 live births in 2019 to 74.7 in 2020, and a preliminary estimate of 107.7 in 2021. Additionally, the center-west region of Brazil has also seen a sharp increase in MMR, with a reported 59.0 deaths per 100,000 live births in 2019, 77.0 in 2020, and 123.6 in 2021 per 100 thousand live births. However, 2021's data is still preliminary and subject to change [MH, BrOO]. Additionally, the state of Goiás has also seen a sharp increase in MMR, with a reported 69.7 deaths per 100,000 live births in 2019 and 90.5 in 2020 [MH].

In Brazil, structuring a regionalized and hierarchical healthcare network such as macro-region, micro-region, and health regions, allows the characterization based on

its socioeconomic, demographic, and epidemiological profile of the population with the identification of priority health problems [LORENA]. Therefore, this structuring of a hierarchical healthcare network can improve decision-making, and ensure that the right resources and support are available at the right place at the right time to address specific health issues of each region, improving the overall health outcomes.

Improving these indicators is essential for the advancement of healthcare services provided to the population, as well as effective management is a key aspect of the improvement of healthcare services. Forecasting the number of live births in a specific area can play a crucial role in improving maternal and newborn health by enabling the proactive provision of care and resources for pregnant women, both before and after delivery, as well as for newborns.

Statistical and machine learning algorithms are widely used to predict future behaviors of health indicators, among other areas. While there are approaches focused on forecasting very specific health conditions such as ischaemic heart disease [McGregor, Watkin e Cox 2004] or ischemic stroke development [Abdullaeva et al. 2019], several works are focused on improving the overall results of the health services provided to the population [Tomašev et al. 2021, Taloba et al. 2022, Ashfaq et al. 2019, Vollmer et al. 2021]. Some works use machine learning models to predict complications in childbirth, preterm birth, and differences in childhood by birth conditions, among other factors of equal importance to public health [Zhang et al. 2022, Akazawa e Hashimoto 2022, Neamțu et al. 2021].

In Adeyinka and Muhajarine [Adeyinka e Muhajarine 2020] approach, machine learning models were used to forecast the under-five mortality rate (U5MR) in Nigeria for the next years to help perform policy actions and planning, just like Elhag and Abu-Zinadah [Elhag e Abu-Zinadah 2020] did with fertility rate in the Saudi Arabic Kingdom. Statistical methods were also used by Bravo and Coelho [Bravo e Coelho 2020] to predict births and deaths in Portugal, and by Ribeiro et al. [Ribeiro et al. 2019] who applied them specifically to forecast tuberculosis incidence in Brazil.

Traditional univariate forecasting techniques are promising for generating accurate forecasts, but artificial neural networks (ANNs) trained on all available time series data in a multivariate approach have shown superior performance [Hewamalage, Bergmeir e Bandara 2021]. Although the performance of forecasting models may decline when dealing with heterogeneous time series data, studies have shown that using clustering techniques to leverage cross-series information can substantially improve results [Bandara, Bergmeir e Smyl 2020]. This is particularly true when dealing with diverse time series data, as demonstrated in various studies [Alvarez et al. 2010, Dantas e Oliveira 2018, Hartmann et al. 2015]. This is a difficult task given the territorial dimension and socioeconomic diver-

sity present in different regions' time series data [Albuquerque et al. 2017]. Clustering strategies can improve cross-series information, which not only can improve the model performance, but also boost its generalization ability. There has been an increase in the use of these state-of-the-art methodologies for health approaches [Cassetti et al. 2008, McCloskey e Poon 2017, Orlandic, Valdes e Atienza 2021], being widely used with traditional unsupervised learning techniques, such as k-means [Aguiar et al. 2022, Imtiaz et al. 2020, Wang et al. 2018].

This work proposes the forecast of the live birth rate for the 18 health regions of the state of Goiás using both univariate and multivariate approaches. The child and maternal mortality rates of the state of Goiás for 2020 were 11.26 per 1,000 live births and 89.47 per 100,000 live births, respectively, while Brazil's rates for the same year were 11.51 child deaths per 1,000 live births and 71.94 maternal deaths per 100,000 live births, which highlights the state's necessity to implement action plans to reduce these numbers, especially MMR, which is above the national rate. The potential of these two approaches allows the study of how the territorial hierarchy can be better understood and managed.

CHAPTER 2

Related Works

Several authors have proposed research studies that utilize both statistical methods and machine learning techniques to forecast health indicators. The task of predicting different health indicators is commonly accomplished using models such as autoregressive integrated moving average (ARIMA), exponential smoothing (ETS), and artificial neural networks (ANN). Statistical and machine learning models were used to predict the under-five mortality rate (U5MR) in Nigeria for the next years to help perform policy actions and planning by Adeyinka and Muhajarine [Adeyinka e Muhajarine 2020], just like Elhag and Abu-Zinadah [Elhag e Abu-Zinadah 2020] did with fertility rate in the Saudi Arabic Kingdom. Statistical methods were also used by Bravo and Coelho [Bravo e Coelho 2020] to predict births and deaths in Portugal, and by Ribeiro et al. [Ribeiro et al. 2019] who applied them specifically to forecast tuberculosis incidence in Brazil.

Adeyinka et al. [Adeyinka e Muhajarine 2020] compared the performances of ARIMA, Holt-Winters exponential smoothing (HWETS), and group method of data handling (GMDH) type artificial neural networks (ANN) in forecasting Nigeria's under-five mortality rate (U5MR) using historical annual data from 1964 to 2017, obtained from the World Bank website. The study conducted two experiments: in-sample prediction (1964-2017) and out-of-sample prediction (2018-2030). Results indicated that GMDH-type ANN was more suitable for long-term forecasting, achieving a Root Mean Squared Error (RMSE) of 0.09. In contrast, ARIMA and HWETS had RMSEs of 0.23 and 2.87, respectively.

Elhag et al. [Elhag e Abu-Zinadah 2020] showed that using 60 years of annual data (1960-2019) from the WHO as input for a multilayer perceptron (MLP) model, it is possible to predict Saudi Arabia's fertility rate for the next five years with a mean absolute percentage error (MAPE) of 2.48%, outperforming the forecasting accuracy of single exponential smoothing (SES) and autoregressive integrated moving average (ARIMA) models.

Bravo and Coelho [Bravo e Coelho 2020] investigated the effectiveness of Seasonal ARIMA (SARIMA), HWETS, and ETS in predicting births and deaths by sex for the 25 Portuguese Territorial Units Nomenclature 3 (TUN3) regions over a 12-month ho-

rizon. The ETS model performed best in predicting births, achieving a weighted average mean absolute percentage error (MAPE) of 7.83% for females and 7.52% for males from 2014 to 2018. On the other hand, SARIMA produced the best results for deaths, with a weighted average MAPE of 8.25% for females and 7.35% for males for the same period as births.

Huang et al. [Huang et al. 2020] proposed using the Legendre Memory Unit (LMU) architecture, introduced by Voelker, Kajić, and Eliasmith [Voelker, Kajić e Eliasmith 2019], to predict mean aortic pressure. They found that the recurrent neural network (RNN) approach using LMU outperformed other models, including Long Short-Term Memory (LSTM) and Transformers, achieving a RMSE of 1.837.

Makipaa [Mäkipää 2021] employed the Facebook Prophet model to predict next-day critical patient admissions at Tampere University Emergency Department Acuta, with the aim of facilitating resource allocation. The model achieved a MAPE of 6.57%, making it the second-best model among other studies that used the same dataset. In a prior investigation, Tuominen et al. [Tuominen et al. 2021] also used the same dataset and task to compare the performances of SARIMA, Prophet, and the general linear model. Among the three models, Prophet showed the best performance, achieving a MAPE of 6.7% in the univariate approach.

Ribeiro et al. [Ribeiro et al. 2019] employed Simple Exponential Smoothing (SES), ARIMA, and HWETS to forecast tuberculosis incidence in Brazil between July 2018 and December 2018. The authors trained these models using data acquired from DATASUS, which included 210 monthly records of tuberculosis diagnoses ranging from January 2001 to June 2018. The results indicated that HWETS outperformed both ARIMA and SES, achieving a MAPE value of 4.00%, while ARIMA and SES achieved MAPE values of 4.84% and 6%, respectively.

Recent research has focused heavily on time series analysis and clustering methods to group time series based on their similarities for complex purposes. In a study by Gómez-Losada et al. [Gómez-Losada, Pires e Pino-Mejías 2018], various statistical clustering techniques were evaluated to estimate the level of background air pollution in urban areas, the characteristics of pollutant concentrations, and the duration of their presence over several years.

James et al. [James e Menzies 2020] proposed a statistical clustering method that evaluates the performance of parametric and nonparametric analyses to analyze the evolution of multivariate time series, with the goal of using a dynamic and simplified implementation of cluster analysis to scale worldwide COVID-19 infections. Meanwhile, Luczak et al. [Łuczak e Kalinowski 2022] utilized a fuzzy clustering method to identify changes in the epidemiological situation of COVID-19 across European countries.

Piryatinska et al. [Piryatinska et al. 2009] proposed a technique for analyzing the level of dysmaturity in newborns by analyzing sleep stages based on extensive EEG records. Their approach clusters nonstationary time series data produced by the EEG signal to provide a more accurate assessment.

Aguiar et al. [Aguiar et al. 2022] introduced a Deep Learning-based prediction strategy, the CAMELOT model, to cluster multivariate time series data from Electronic Health Records (EHR). The model architecture consists of three stages: first, the input time series are represented using an RNN-based encoder network with an attention layer; second, the clusters are selected and identified using an MLP; and third, the predictions are made using another MLP. The authors evaluated the performance of CAMELOT in comparison with other prediction algorithms, such as Support Vector Machine (SVM), XGBoost (XGB), and NEWS2, and found that CAMELOT outperformed them by at least 4% in interpretability metrics in cluster formation.

Imtiaz et al. [Imtiaz et al. 2020] introduced a federated learning model based on Long-Short Term Memory (LSTM) for predicting user privacy preservation in a health data stream. The model uses multivariate time series data from MyFitnessPal apps, Fitbit dataset, and Fitbit-GAN dataset, and implements a clustering mechanism using a K-means streaming algorithm and pattern matching to group users with similar health and diet profiles. The LSTM model achieved results that were within 0.025% of the reference values, and the clustering approach significantly reduced computational time, resulting in up to 49% error reduction compared to the model for the entire dataset.

Gonzalez et al. [González 2019] also employed federated learning with an LSTM in a clustering mechanism using agglomerative clustering algorithms. Their study used telecom operator customer data to generate monthly time series observations. The study was effective in exploring similar time series and showed better prediction metrics when using clusters.

CHAPTER 3

Background

3.1 Time Series Forecasting Problem

A time series is a collection of values that are ordered chronologically and observed over a period of time, where the data are typically sampled at regular intervals with a fixed frequency. Approaches to time series forecasting can be either univariate when there is only one time-dependent variable, or multivariate when there are multiple time-dependent variables.

A univariate time series with L values in the historical data can be defined as:

$$y = y(t-L),\ \ldots\ , y(t-1),\ y(t),\ y(t+1),\ \ldots,\ y(t+h) \tag{3-1}$$

where, each $y = t-i$, for $i = 0, 1, \ldots, L$, represents the recorded values of y at time $t-i$. The forecasting process consists of estimating a horizon of prediction, denoted as h, which refers to the number of predicted time steps ahead of $y(t)$. Given a desired horizon of predictions h, the forecasting process involves estimating h values ahead of $y(t)$, that is $\hat{y}(t+i)$, where $i = 1, \ldots, h$. The optimal predicted values are reached when the function $\sum_{i=1}^{h} (y(t+i) - \hat{y}(t+i))$ is minimized. The univariate approach is also demonstrated in Figure 3.1.

A multivariate time series forecasting approach can be represented as a method for predicting the h values ahead of $y_i(t)$, with $i = 0, 1,\ \ldots,\ h$, of T time series simultaneously, and is described in the matrix form in Equation (3-2).

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} y_1(t-L),\ \ldots,\ y_1(t-1)\ ,\ y_1(t),\ y_1(t+1),\ ,\ldots,\ y_1(t+h) \\ y_2(t-L),\ \ldots,\ y_2(t-1)\ ,\ y_2(t),\ y_2(t+1),\ ,\ldots,\ y_2(t+h) \\ \vdots \\ y_n(t-L),\ \ldots,\ y_n(t-1)\ ,\ y_n(t),\ y_n(t+1),\ ,\ldots,\ y_n(t+h) \end{pmatrix} \tag{3-2}$$

where $n = 1, 2,\ \ldots\ , T$ represents the total number of time series that will be modeled. The same process is illustrated in Figure 3.2.

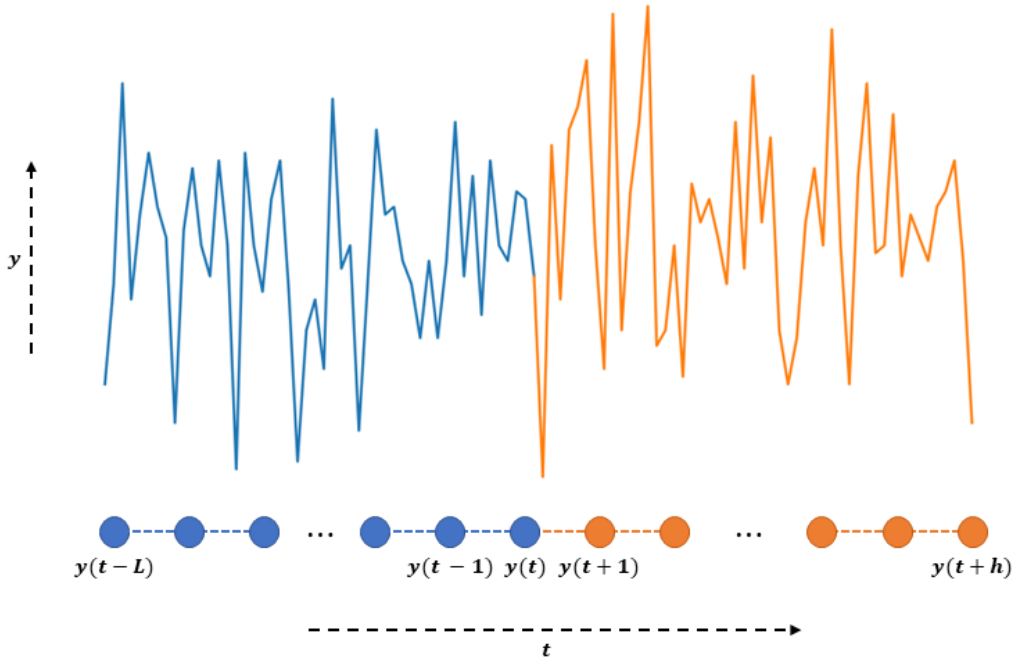

Figure 3.1: Univariate time series forecast problem

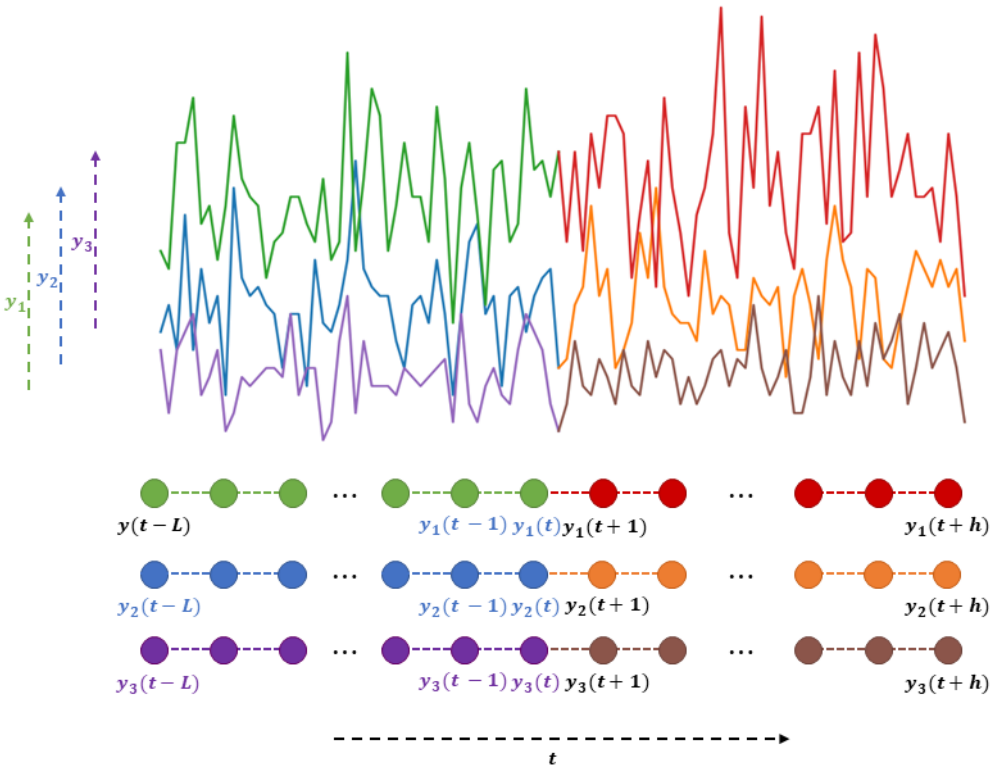

Figure 3.2: Multivariate time series forecast problem

3.2 Time Series Clustering

Time series clustering can be used to discover the distribution of patterns in data by organizing data points into paired groups based on their similarities. The challenge of

clustering time series lies in the fact that each data point is an ordered sequence, however, making clusters of similar time series may improves forecast accuracy and optimize the computational time of training. Clustering algorithms are mostly based on distance measures to find similar attributes between the data used.

3.2.1 *K*-means

According to Forsyth [Forsyth 2016] *K*-means is a technique capable of creating groups or clusters by measuring the similarity between samples in a dataset. Calculating the distance between samples can determine how similar they are to one another. This distance usually is defined by the Minkowski distance [Kamber, Pei et al. 2001], which is a generalization of Euclidean and Manhattan distances:

$$d(x_i, x_j) = \left(|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \ldots + |x_{ip} - x_{jp}|^g\right)^{1/g} \tag{3-3}$$

Given a dataset X, clustering is the process of creating multiple groups $(C_1, C_2, \ldots, C_k)$ in which the samples represented therein have a high similarity to each other and a higher dissimilarity between different groups [Kamber, Pei et al. 2001]. The points representing each group are called the centroid [MacQueen 1967]. Centroids are given by the mean vector of C_k [Maimon e Rokach 2005]:

$$\mu_{k,j} = \frac{1}{N_k} \sum_{\forall x_i \epsilon C_k} x_{i,j} \tag{3-4}$$

To obtain the groups, first, the number of groups K must be defined, then all samples are assigned to the nearest centroid and then the centroids are calculated again. This process is repeated until the algorithm finds a convergence [Maimon e Rokach 2005, Selim e Ismail 1984].

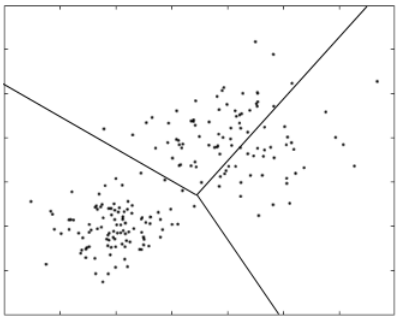

Figure 3.3: Clustering after *K*-means convergence.

The sum of squared error (SSE) is the simplest and most common method for evaluating groups formed by the *K*-means [Maimon e Rokach 2005] algorithm and is

calculated by:

$$SSE = \sum_{i=1}^{k} \sum_{\forall x_i \epsilon C_k} ||x_i - \mu_k||^2 \tag{3-5}$$

3.2.2 Dynamic Time Warping

Dynamic Time Warping (DTW) is invariant to time changes and insensitive to abnormal points, making it very important for work involving time series clustering. Some works highlight the use of DTW measurement for time series clustering cases [Keogh e Ratanamahatana 2005, Petitjean et al. 2014, Müller 2007]. The calculation of the DTW distance measure combines and maps the morphology of the time series by bending the time axis with respect to time data points, and then elastic transformations are made to find optimal nonlinear alignment between different time series sequences. An example of how it is calculated is illustrated in Figure 3.4, where these matching closer points are calculated through the distance matrix calculated by Equation (3-7).

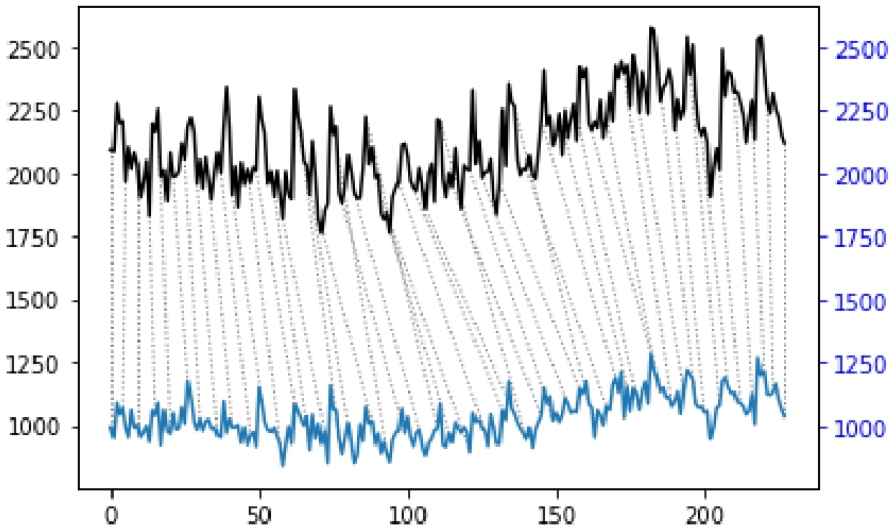

Figure 3.4: Time matching of the data based on the DTW alignment.

Considering two time series $O = o_1, o_2, \cdots, o_i, \cdots, o_n$ and $P = p_1, p_2, \cdots, p_j, \cdots, p_m$. A n-by-m matrix is allocated to store the distance points of any two pairs of values of the time series O and P. Thus, aligning two points to the (i^{th}, j^{th}) element of the matrix will contain the distance corresponding to $d(o_i, p_j) = (o_i - p_j)^2$. The home element ($i$,$j$) of the matrix is represented by an alignment between two points in the time series, as seen in Figure 3.4. The path L is a set of matrix elements that defines a mapping between O and P so that we find the best match between the two sequences. Thus, we will have:

$$L = l_1, l_2, \ldots, l_k, max(n, m) \leq k < (m + n - 1), \tag{3-6}$$

where the l^{th} element of L is being defined as $l_k = (i,j)_k$. The path that minimizes the warping cost for the aligned points is defined by:

$$DTW(O,P) = min\left(\sqrt{\sum_{k=1}^{K} l_k}\right) \tag{3-7}$$

and therefore, the DTW distance can be represented by:

$$\begin{aligned}\gamma(i,j) = d(o_i,p_j) + min\{&\gamma(i-1,j-1),\\ &\gamma(i-1,j),\gamma(i,j-1)\}\end{aligned} \tag{3-8}$$

Recursively, similarities between time series and cumulative distance patterns $\gamma(i,j)$ are measured using the current distance $d(i,j)$ and the minimum of cumulative adjacent distances. Thus, by providing the desired number of clusters, the method calculates the optimal cluster based on the shortest distance of the distance matrix n-by-m.

3.3 Artificial Neural Networks - ANNs

ANNs are inspired by our biological understanding of the human nervous system [Lippmann 1987]. ANNs have a structure that allows information to be received and then associated with responses, by means of a highly connected system based on neurons, its basic unit [Oliveira 2010]. The idea of an artificial neuron was first proposed by [McCulloch e Pitts 1943], as can be depicted in Figure 3.5. Subsequently, McCulloch's artificial neuron model culminated in the design of the *Perceptron* [Rosenblatt 1958] and *Adaline* [Widrow e Hoff 1960].

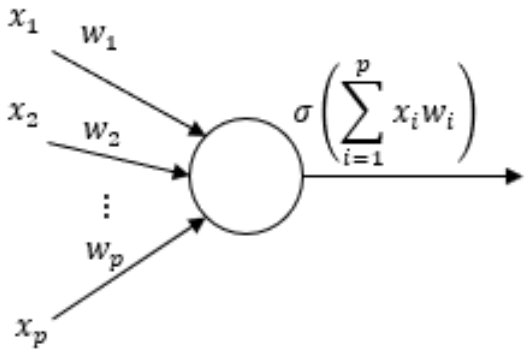

Figure 3.5: Model of an artificial neuron [McCulloch e Pitts 1943].

Each input value in the neuron is associated with a weight w_i that will reflect the importance of this input to the output y_j. The output of perceptron can be represented by Equation (3-9), where a threshold τ is subtracted from the linear combination between the

inputs and their respective weights, if the result is ≥ 0 then $y = +1$ and $y = -1$ otherwise. It is only possible to solve linearly separable problems [Lippmann 1987, Rauber 2005].

$$y_j = \sigma\left(\sum_{i=1}^{p} w_i x_i - \tau\right) \tag{3-9}$$

The potential and flexibility of the computations performed in a neural network comes from the creation of a set of interconnected neurons. An ANN can be referred to as a feedforward network because received information is propagated in a single direction, thus allowing information to flow forward only [Rauber 2005].

Neurons that receive information simultaneously are organized in layers. The most common form of an ANN is one with multi-layer perceptrons (MLP). As can be seen in Figure 3.6, an MLP network has an input layer, composed of neurons that just propagate information, an output layer, composed of neurons that will determine the output y_k for a given input, and all the layers that are between the input and output layer are described as hidden layers [Lippmann 1987, Oliveira 2010].

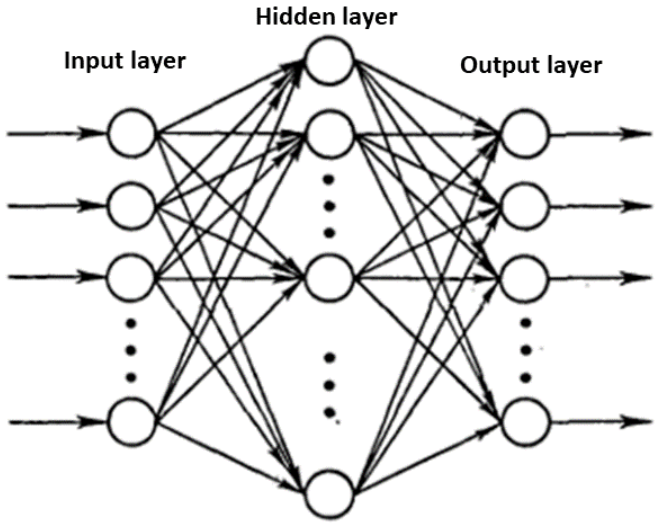

Figure 3.6: MLP Representation by [Rumelhart, Hinton e Williams 1985].

An MLP network with a fixed number of neurons in the hidden layer is capable of approximating a wide range of functions. The process of training an MLP net can be supervised, where a set of training data with corresponding outputs is required to enable learning. Training an MLP network is the process by which the values of the weights are determined so that the network can generalize to new data. In other words, the weights w_i are adjusted to minimize the error value, which is the difference between the predicted output and the true output for a given input [Gardner e Dorling 1998].

Through an backpropagation algorithm the neural network is able to adjust the weights w based on the error ε between predicted and true outputs. The algorithm

determines the direction of decreasing ε by computing the gradient defined in Equation 3-10. To minimize ε, the weights are updated in the direction of negative gradient $-\nabla\varepsilon$, with the magnitude of change determined by a predefined learning rate η, as given by Equation 3-11 [Rauber 2005].

$$\nabla\varepsilon(w) = \frac{\partial\varepsilon(w)}{\partial w} \tag{3-10}$$

$$w_{i+1} = w_i - \eta\nabla\varepsilon \tag{3-11}$$

3.4 Legendre Memory Unit - LMU

The LMU model is an recurrent neural network (RNN) architecture constituted mainly by a Linear Time-Invariant (LTI) system and a nonlinear dynamical system. The first system is a memory cell responsible for orthogonalizing a sliding window of length θ of the encoded input u_t, shown in Equation (3-12), storing this information through a linear combination of scale-invariant Legendre polynomials m_t, whose definition is presented in Equation (3-13). The second system applies a nonlinear function f to a linear combination of the input vector x_t, the memory vector m_t and the hidden state vector h_t and its respective weight matrices W_x, W_m and W_h. Then, the hidden state output of one time-step iteration of LMU is demonstrated in Equation (3-14),

$$u_t = e_x^T x_t + e_h^T h_{t-1} + e_m^T m_{t-1} \tag{3-12}$$

$$m_t = \bar{A}m_{t-1} + \bar{B}x_t \tag{3-13}$$

$$\mathbf{h_t} = f(W_x x_t + W_m m_t + W_h h_{t-1}) \tag{3-14}$$

where $\bar{A}$ and $\bar{B}$ are discretized matrices generated by Equation (3-15) and Equation (3-16), and d is the order of the system and size of the m_t:

$$\mathbf{A} = [a]_{ij} \in \mathbb{R}^{d\times d}, \quad a_{ij} = (2i+1)\begin{cases}(-1) & i < j\\ (-1)^{i-j+1} & i \geq j\end{cases} \tag{3-15}$$

$$\mathbf{B} = [b]_i \in \mathbb{R}^{d\times 1}, \quad b_i = (2i+1)(-1)^i, i, j \in [0, d-1] \tag{3-16}$$

The architecture of the LMU cell is then based on its parameters and systems. As time steps pass through, the cell functions in a self-looping process, which is the agent that keeps temporal data stored in the cell. This recurrent relation is illustrated in Figure 3.7.

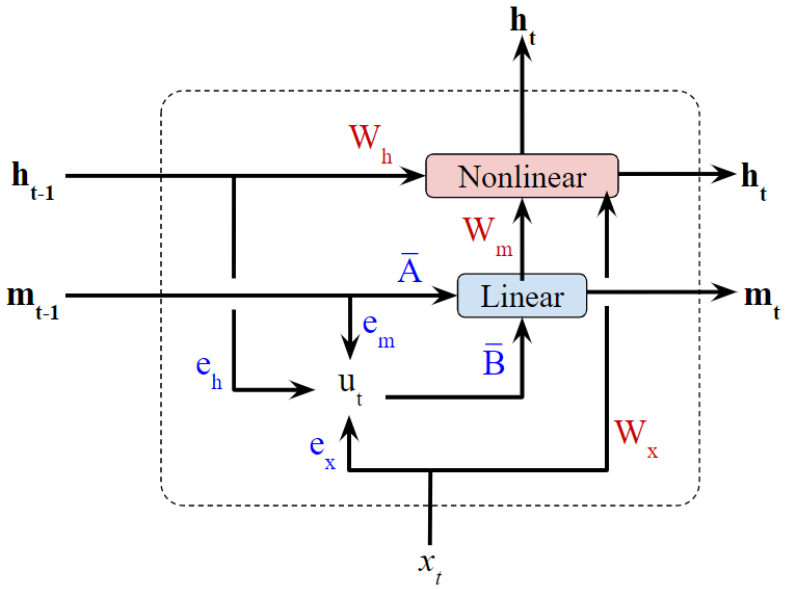

Figure 3.7: LMU cell in one-time-step according to [Voelker, Kajić e Eliasmith 2019].

The memory's parameters A, B, and θ, like in Voelker, Kajić, and Eliasmith [Voelker, Kajić e Eliasmith 2019], are not required to be learned, hence they were not trained, keeping their values frozen along the training process. On the other hand, the encoding vector of weights e_x, e_h, and e_m and the weight matrices W_x, W_m, and W_h are learned throughout the proceeding.

The fact that the linear system is decoupled from the nonlinear system ensures the possibility to trade between the order parameter (d), which when augmented improves the linear memory capacity, and the hidden state size parameter (n), which when increased enhances the memory's ability to learn nonlinear complex functions.

One of LMU's biggest advantages is that it approaches the exploding and vanishing gradients, that's usually a problem in RNN models, allowing it to handle long-range temporal dependencies, contemplating around 100,000 time steps. It also proves to converge quickly using fewer internal state variables when compared to other RNN models.

CHAPTER 4

Live Births Prediction for Health Regions of Goiás with Machine Learning

The following sections will present the methodology of the proposed approaches, comprising a detailed description of data collection and preprocessing methods, as well as an explanation of the forecasting and evaluation strategies employed in the subsequent chapters. Additionally, the methodology here presented was used in two case studies (Chapter 5 and Chapter 6), where each one is one submitted paper.

4.1 Data Collection

The Brazilian population's heterogeneity in terms of socioeconomic, demographic, and epidemiological profiles can be addressed by systematizing a hierarchical healthcare network consisting of macro, micro, and health regions [LORENA]. This approach allows for the adaptation of health services and resources to the unique needs of specific geographic areas, ensuring that each region receives the appropriate level of care based on its distinct profile [Viana et al. 2015].

Goiás, located in the mid-western region of Brazil, has a large population and is one of the states that places a strong emphasis on efficient government decision-making to improve public health indicators. As the 12th most populous state in Brazil, with an estimated population number of 7,206,589 inhabitants, as reported by the lasted demographic census published by the Fundação Instituto Brasileiro de Geografia e Estatística (IBGE), Goiás faces significant challenges in ensuring the health and well-being of its population. Effective and well-coordinated efforts are necessary to address the unique needs and requirements of such a large and diverse population, and to achieve positive health outcomes across the state.

This work analyzes the 18 health regions of the state of Goiás, shown in Figure 4.1, and utilized data on live births from 2000 to the end of 2020, with a monthly resolution. These data were provided by the information system on live births of the information department of the single health system (SINASC-DATASUS). The Brazilian DATASUS is an agency that manages health information, such as records and information

processing, and financial information referring to public resources, credits, and budgets directed to health. The dataset is organized into 252 observations distributed monthly for the number of live births from january 2000 to december 2020 for each health region in the state of Goiás, resulting in a total of 18 unique time series data. This dataset is restricted to 2020 data as the information for the subsequent years is being compiled and may change. The data history of all health region time series can be seen in Figure 4.2.

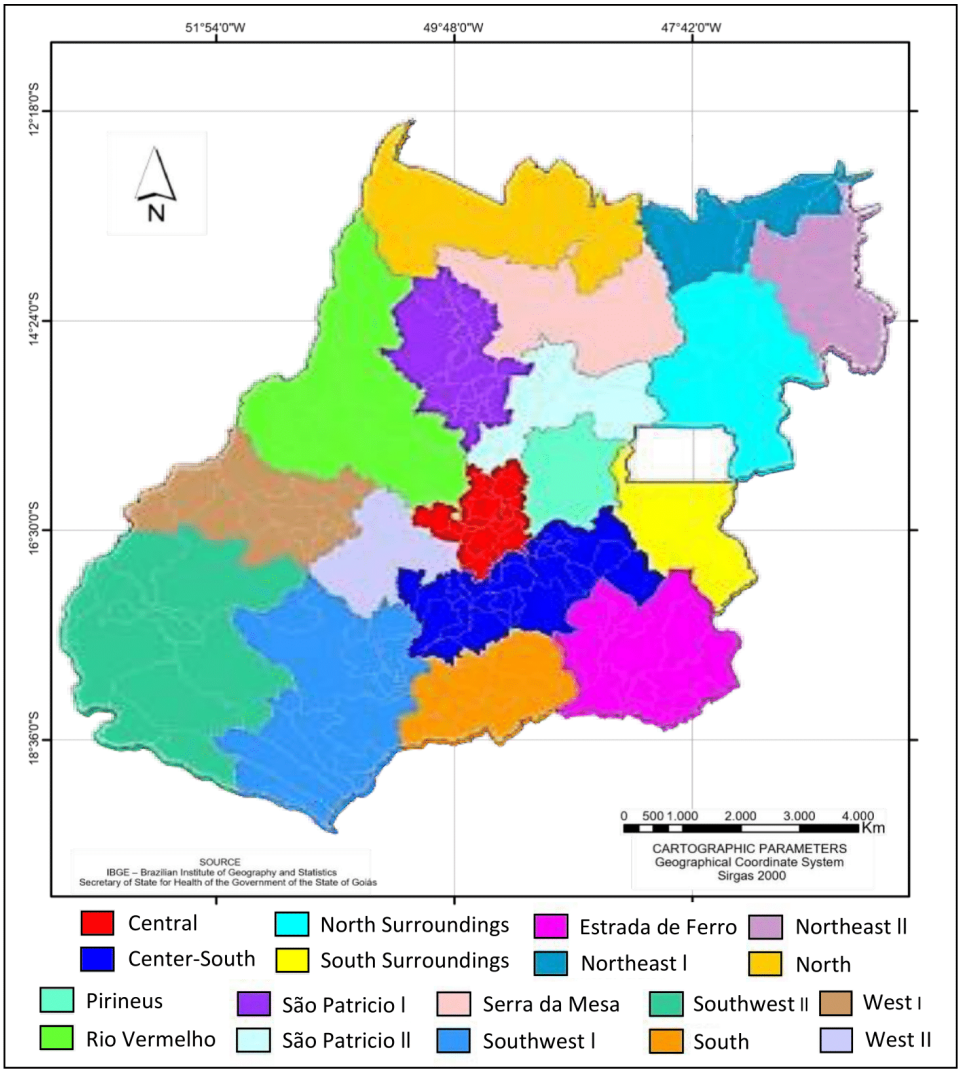

Figure 4.1: State of Goiás divided into 18 health regions.

Between 2010 and 2018, the number of live births increased in Brazil, with a slight decrease in 2017 associated with alerts caused by the Zika virus epidemic, which can cause fetal malformation [Castro et al. 2018]. However, after 2018, a reduction in live births was observed, possibly related to changes in the socioeconomic profile of the population [Duarte e Teixeira 2021]. In 2020 was a significant decrease in the total number of births, as many women opted to delay pregnancy due to the high

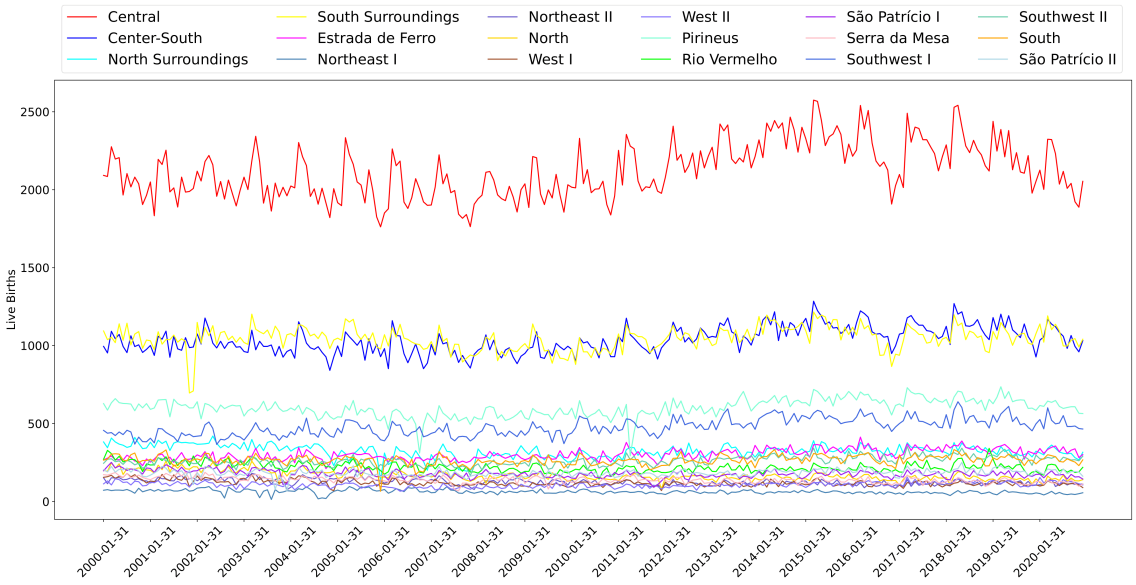

Figure 4.2: Number of live births and date in monthly time spans for all health regions of the state of Goiás, Brazil.

incidence of SARS-CoV-2 virus infections and the associated risk of maternal mortality [Souza e Amorim 2021].

4.2 Data Preprocessing

To adjust the data on a common magnitude scale and provide more effective weight adjustments for neural networks, the data is normalized and scaled. This helps ensure that the data are uniformly scaled, allowing the model to learn the underlying patterns and relationships in the data [Hewamalage, Bergmeir e Bandara 2021]. The preprocessing step occurs according to Equation (4-1)

$$X_t = \left(\frac{x_t - min(x)}{max(x) - min(x)} \right) \cdot (max - min) + min \tag{4-1}$$

where x_t represents a time serie sample at time step t, while X_t represents the same sample after preprocessing. $min(x)$ and $max(x)$ are the lowest and highest value of a series, respectively. Finally, max represents the highest desired value, which in this case is set to 1, while min represents the lowest desired value, set to -1.

4.3 Sliding Window

Given that the forecasting problem requires making predictions for multiple time steps in advance, the sliding window (SW) strategy has been adopted. This approach involves using a moving window of historical data to train the model, where the window size corresponds to the number of time steps to be predicted [Hewamalage, Bergmeir e Bandara 2021]. That means, the SW is used in order to split the samples into pairs of $x(t)$ and $y(t)$, as illustrated in Figure 4.3.

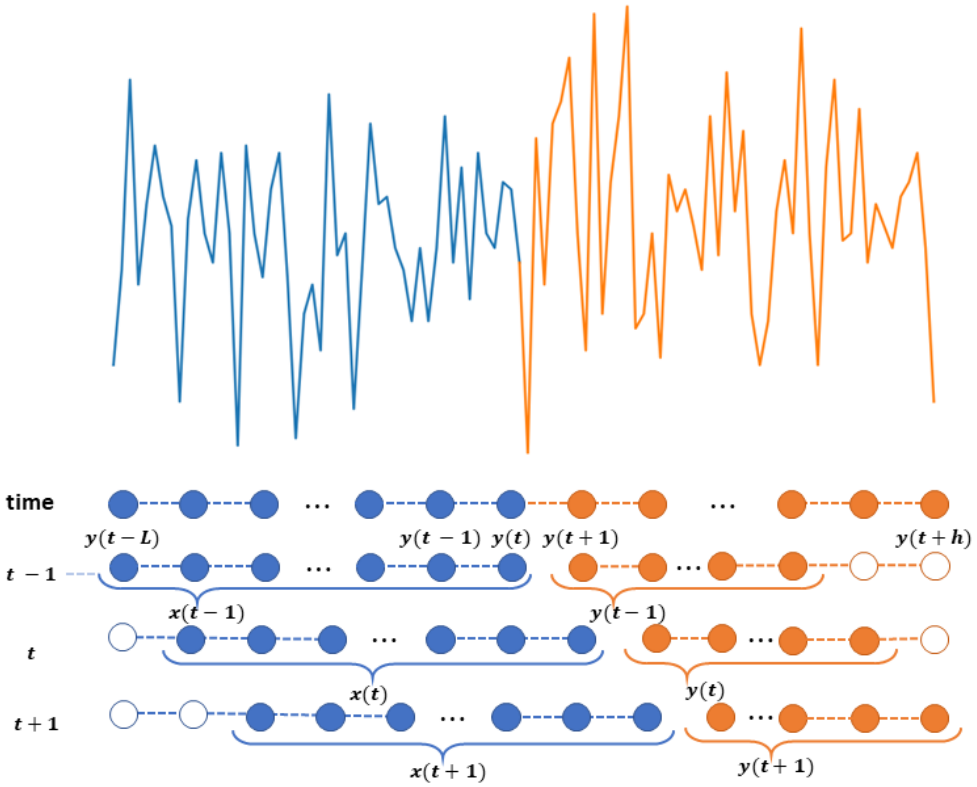

Figure 4.3: Representation of a moving window training strategy.

Both input samples, denoted as $x(t)$, and output samples, denoted as $y(t)$, has been defined to be of size 24. This means, that for each time step the model has 24 consecutive data points to predict the next 24 data points. It was defined 24-month outputs to provide a sufficient planning horizon for health managers and accommodate potential delay of official information in SINASC-DATASUS. That means that the trained model will be able to generate a 2-year prediction. By predicting outcomes up to two years in advance, the model can support proactive decision-making and mitigate the impact of any delays or data reporting issues. The input size was set to match the output size to enable the model to learn the complexity required for long-range prediction.

The SW is also used to generate two years of predictions. As depicted in Figure 4.4, at step zero the most recent 24 months of recorded data is gathered, which will be

model's first input. Since the model is expected to produce 24 outputs at each step, only the first prediction is taken. In this way, an accumulative error from the last predictions is considered at each step to generate the next predictions and so on, until $h = 24$. Resulting predictions are then evaluated against the ground truth data using MAPE and MAE metrics, which have been selected as they provide reliable evaluation strategies for forecasting models.

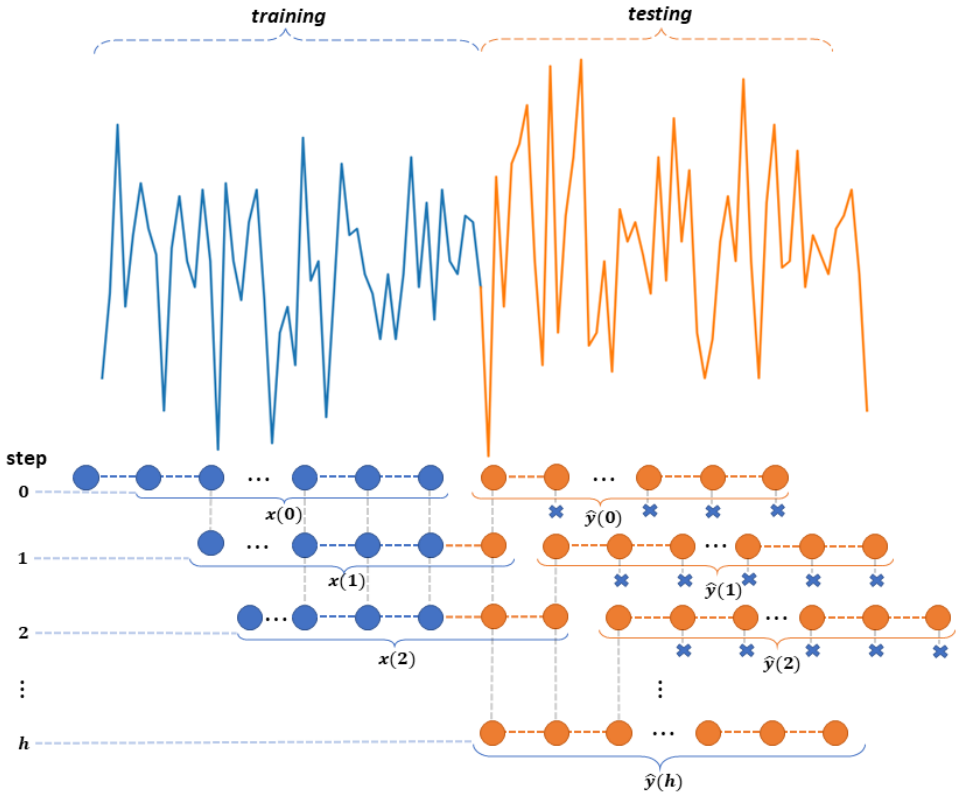

Figure 4.4: Representation of a prediction over prediction strategy.

4.4 Performance Measures

The proposed methodologies presented in the next chapters use the same evaluation strategies, namely, mean absolute percentage error (MAPE), described in Equation (4-2), and mean absolute error (MAE), described in Equation (4-3), as they have been widely used in the literature and provide a comprehensive evaluation of the forecast performance.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} * 100 \tag{4-2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4-3}$$

CHAPTER 5

Case study I: Live Births Prediction using Legendre Memory Unit: A case study for the health regions of Goiás

Although traditional univariate techniques such as ETS and ARIMA may not always perform well in capturing the complexity and dynamics of real-world data, they are still widely used due to their simplicity of implementation and understanding. Additionally, they provide a straightforward approach for modeling time series data, which can be useful in situations where computational resources or data availability is limited. Despite the recent successes of RNNs in forecasting, there are currently no established guidelines on when traditional statistical methods will outperform RNNs, even when developing and adapting complex RNN models [Hewamalage, Bergmeir e Bandara 2021].

In this context, the LMU model is explored in this chapter. This novel recurrent architecture has achieved state-of-the-art memory capacity tasks and performs more efficiently than other models [Voelker, Kajić e Eliasmith 2019]. Additionally, results for statistical models such as ARIMA, ETS, and Prophet are reported, these models are optimized in the same way that LMU. Chapter 3 presents the model description, while Chapter 4 demonstrates how the data is collected and processed to fit into the model.

Furthermore, the model was optimized using a grid search strategy, where the hyper-parameters considered in the process are mentioned in Table 5.1. The selection of these parameters was determined by prior experiments and available resources capability. This optimization procedure is essential for identifying the hyperparameters that best suit each region and for preventing overfitting of the model.

The search space for batch size consisted of 8, 16, 32, and 64, while the number of epochs was set to 500. Setting the number of epochs to 500, and a patience of 50 epochs provides a sufficient amount of training time for the model to learn the underlying patterns and relationships in the data without overfitting. This approach allowed the model to stop training when performance improvement began to plateau. Additionally, since some parameters of LMU are randomly initialized, such as kernel matrices and encoders, the seed was also changed along the optimization. It was done to guarantee that the model

Table 5.1: LMU's hyper-parameters and its search space.

Parameter	Search Space and frozen values
Order (d)	[32, 64, 128, 256]
Hidden state size (n)	[32, 64, 128, 256]
SW size (θ)	[24]
Batch size	[8, 16, 32, 64]
Epochs	[500]
Patience	[50]
Seeds	[10, 42, 58, 3407]

was stable independently of the stochastic initialized parameters.

Finally, the optimizer used was the Continuous Coin Betting (COCOB), which is a recent stochastic gradient descent algorithm proposed by Orabona and Tommase [Orabona e Tommasi 2017], who affirm in their work that the optimization process is reduced to a game of betting on a coin. The main advantage of this optimizer is that it self-tunes the learning rate, which is why it is not included in the grid search procedure. COCOB has gained popularity once it eliminates the need to optimize the learning rate and it proves to perform better in time series forecasting in state-of-the-art papers [Hewamalage, Bergmeir e Bandara 2021, Bandara et al. 2021].

5.1 Experimental Results

The total number of optimization tryouts for LMU was 64 for each region, resulting in 1,152 combinations considering the 18 regions. Figure 5.1 shows results for all 18 health regions, where the best model configuration is used to generate the predictions. Additionally, the results of all evaluated models are represented in Table 5.2.

In Table 5.2 it is evidenced MAE and MAPE results for LMU, ARIMA, Prophet, and ETS models considering each health region separately and the state's mean. Additionally, there is also the live births mean for each region from 2019 to 2020. From that, it is noticeable the correlation between LMU's performance and the live births mean. The model achieved better error metrics in health regions with higher live births mean and only reasonable performance in regions with lower live births mean. This behavior was also seen in Bravo and Coelho [Bravo e Coelho 2020] when predicting births and deaths for Portugal's TUN3 using only statistical models.

The results for the state of Goiás showed that ARIMA achieved the best overall MAPE. Accordant to the observed in the related works section, ARIMA is one of the most used models for time series forecasting and the results presented in this work prove that it also fits well in the scenario of live births. While LMU achieved a slightly lower MAPE

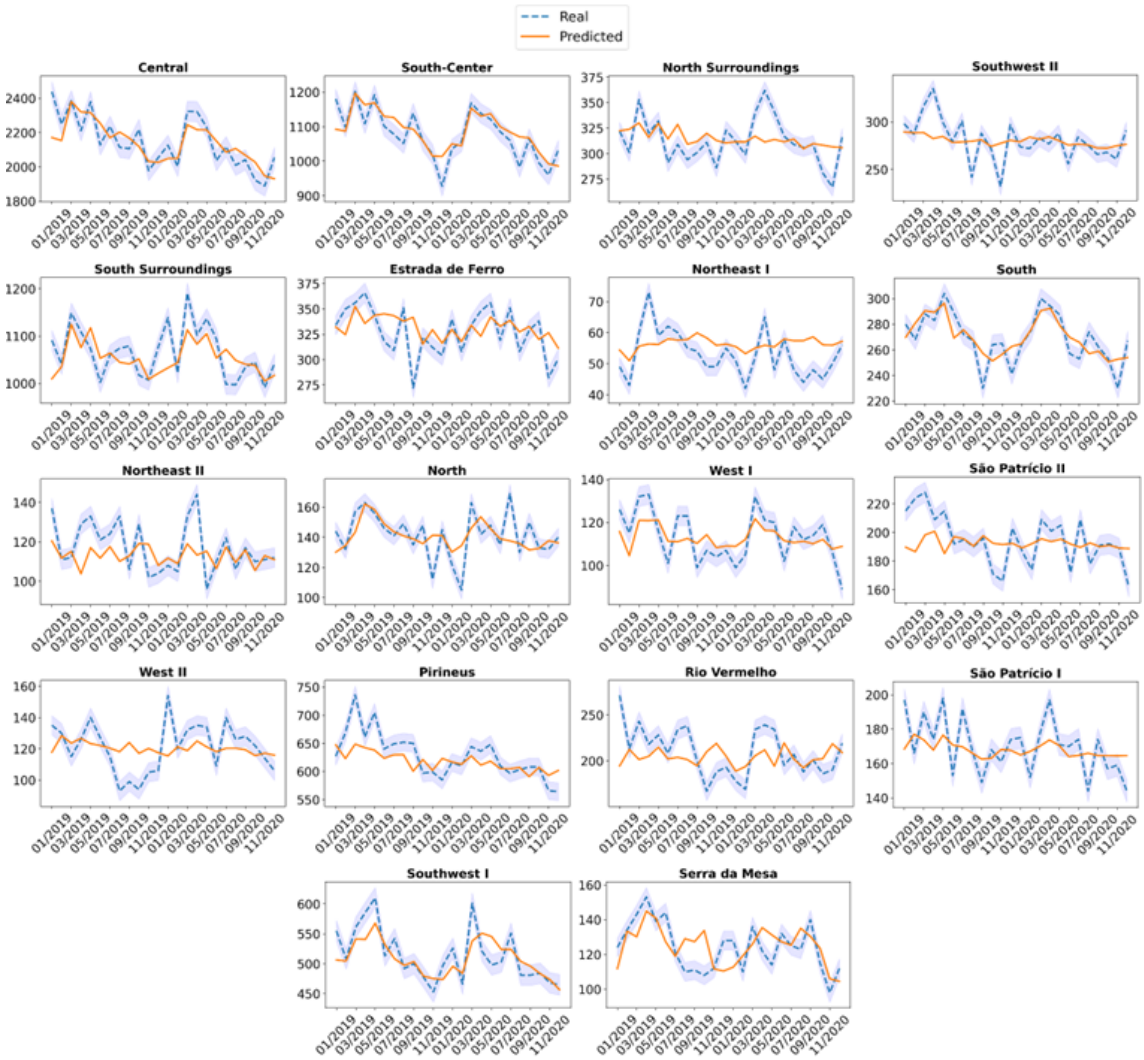

Figure 5.1: Results of the univariate LMU applied to all health regions of Goiás.

than ARIMA, LMU reached the best overall MAE. These results suggest that LMU is a highly effective approach for modeling such data and offers a competitive alternative to traditional statistical models such as ARIMA.

Furthermore, it can also be seen that most of the health regions where ETS and ARIMA overcame LMU were regions with lower live births mean. This happens due to the fact that statistical models need less amount of data to train compared to machine and deep learning methods, such as LMU. For data from low live births regions, it is harder to express the health attributes in comparison to high live births regions. Hence, the 252 samples of these regions may have been enough for statistical models to learn better than LMU, which may need more data to fit better the live birth rate for these locations.

Table 5.2: Live births mean and error metrics for each health region and state with all models evaluated on 2019-2020 forecast.

Health region	Live births mean	ARIMA		Prophet		ETS		LMU	
		MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Central	2149	82.4479	3.7940	119.4458	5.4865	128.6836	5.7422	**81.3121**	**3.7630**
Estrada de Ferro	327	**17.3135**	**5.3292**	18.4363	5.8751	19.3445	5.6765	17.6482	5.6295
North	142	**9.1326**	**6.4177**	12.1325	8.5205	9.5333	6.6195	9.7638	7.1731
North Surroundings	313	19.8791	6.0833	18.3211	5.8478	26.8542	7.9628	**15.5576**	**5.0142**
Northeast I	53	6.1999	11.1708	5.8825	11.5876	**4.8953**	**8.9647**	6.5262	12.8202
Northeast II	117	10.6430	9.3465	10.2323	8.3717	9.3436	8.0597	**9.5880**	**7.8834**
Pirineus	628	**23.3633**	**3.7147**	32.2141	5.0747	29.7174	4.6125	24.4367	3.7843
Rio Vermelho	210	27.4492	11.7839	**21.4760**	**9.8399**	47.2651	18.5988	22.0541	10.2840
São Patrício I	170	**11.6243**	**6.6171**	13.4162	7.5975	12.0453	6.9699	12.2117	7.1695
São Patrício II	195	12.3725	6.3453	16.1416	8.1900	**10.9664**	**5.7232**	13.7203	7.0622
Serra da Mesa	124	10.9499	8.7926	11.4358	9.5512	11.0501	9.0740	**10.3866**	**8.5684**
South	269	11.0334	4.0559	13.1251	5.0129	11.5999	4.2193	**9.9294**	**3.8154**
South Surroundings	1065	35.4194	3.3777	50.0397	4.6028	**32.9006**	**3.1557**	36.5511	3.3959
South-Center	1077	33.5097	3.1078	58.1816	5.4811	55.0561	4.9244	**31.2450**	**2.9627**
Southwest I	514	23.4386	4.4926	34.1443	6.5823	27.0962	5.0913	**23.2684**	**4.3728**
Southwest II	281	**12.4981**	**4.4788**	18.8340	6.8762	18.3881	6.2626	13.9953	5.0357
West I	113	6.4814	5.7596	9.4666	8.4516	**6.2235**	**5.4391**	8.0232	7.2046
West II	121	12.2170	9.7636	13.3485	11.9695	**11.8608**	**9.5868**	12.2266	10.3658
Goiás**	437	20.3318	**6.3572**	26.4597	7.4955	26.2680	7.0380	**19.9136**	6.4614

***** *- Best result of five random seed initialization*

****** *- Average of the results of the health regions*

CHAPTER 6

Case study II: Live Births Forecasting Across Health Regions of Goiás using Artificial Neural Networks: A Clustering Approach

Although univariate forecasting has been shown to be highly effective for modeling time series data, many researchers have also explored the potential benefits of leveraging cross-series information in their forecasts. This approach has gained popularity in recent years and has been used in a wide range of forecasting applications [Aguiar et al. 2022, Imtiaz et al. 2020, Wang et al. 2018, Imtiaz et al. 2020], rather than developing separate local models for each time series in a dataset, a global model is trained by leveraging data from multiple time series simultaneously [Hewamalage, Bergmeir e Bandara 2021]. Although global models can be applied to a set of time series, this does not mean that the series' forecasts are dependent on one another. Instead, this method entails estimating parameters for all of the available time series simultaneously. [Januschowski et al. 2020, Hewamalage, Bergmeir e Bandara 2021].

The idea of building global models was enhanced in this study by incorporating the concept of clustering similar time series, which allows the model to explore temporal similarities between cross-series information. In this way, all 18 health regions' data were clustered based on their similarities, which means that each cluster had the most similar health regions' time series.

A global model was trained for each cluster, and different numbers of clusters were tested to see the benefits of clustering similar time series. The K-means was used for clustering the most similar health regions' data, with $K = 1, 2, \ldots, 6$, using DTW as a distance metric. The proposed ANN-based approach is multivariate [Reinsel 2003], where K models are trained to generate results for each of the health regions. The input and output layer corresponds to the number of clustered time series and the length of the prediction horizon. For the training set, the SW strategy is used in order to sample sequence values into a pair of input and output data, using a principle of multi-input multi-output [Hewamalage, Bergmeir e Bandara 2021]. Table 6.1 shows the hyperparameters and the range of values used for performing grid search greedy optimization. As an

optimizer for the training algorithm the continuous coin betting optimizer (COCOB) [Orabona e Tommasi 2017] is used.

Table 6.1: Hyper-parameters Optimization of ANN.

Parameter	Search Space
Hidden Layers	1,2,3
Number of Neurons	50, 150, 200
Batch Size	32, 64
Epoch	500
Seed	10, 3407

6.1 Experimental Results

The proposed methodology uses K-means for clustering time series of health regions in the state of Goiás. The clusters generated by K-means will be used as input to train and test MLP model. In this case, a study was carried out with time series of all health regions in the state of Goiás to determine values of K for better prediction using MLP. Figure 6.1 shows the relationship of K with MAPE generated by MLP prediction. Note that the best MAPE and K relationship was obtained for $K = 2$ on average. Thus, $K = 2$ was adopted to generate all clusters using K-means for MLP input for all health regions. Additionally, the silhouette score is calculated for each value of K to measure the similarity of each time series to its own cluster in comparison to other clusters. The optimal number of clusters can be chosen by selecting the value of K that maximizes the silhouette score. The silhouette score for each value of K is shown in Figure 6.2, and it can be observed that the highest score of 0.7622 was obtained for $K = 2$.

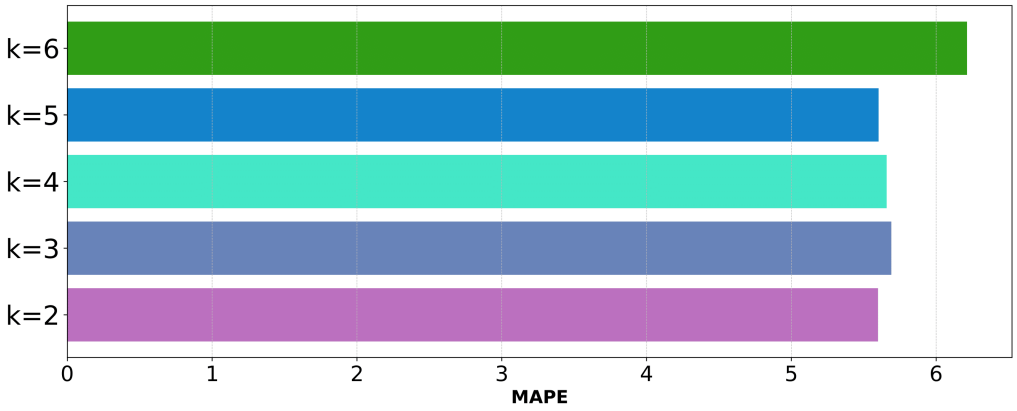

Figure 6.1: MAPE versus K for health regions of the state of Goiás.

Figure 6.3 shows the prediction results obtained from two global models trained on two clusters of time series, for 18 health regions in the state of Goiás, using an MLP.

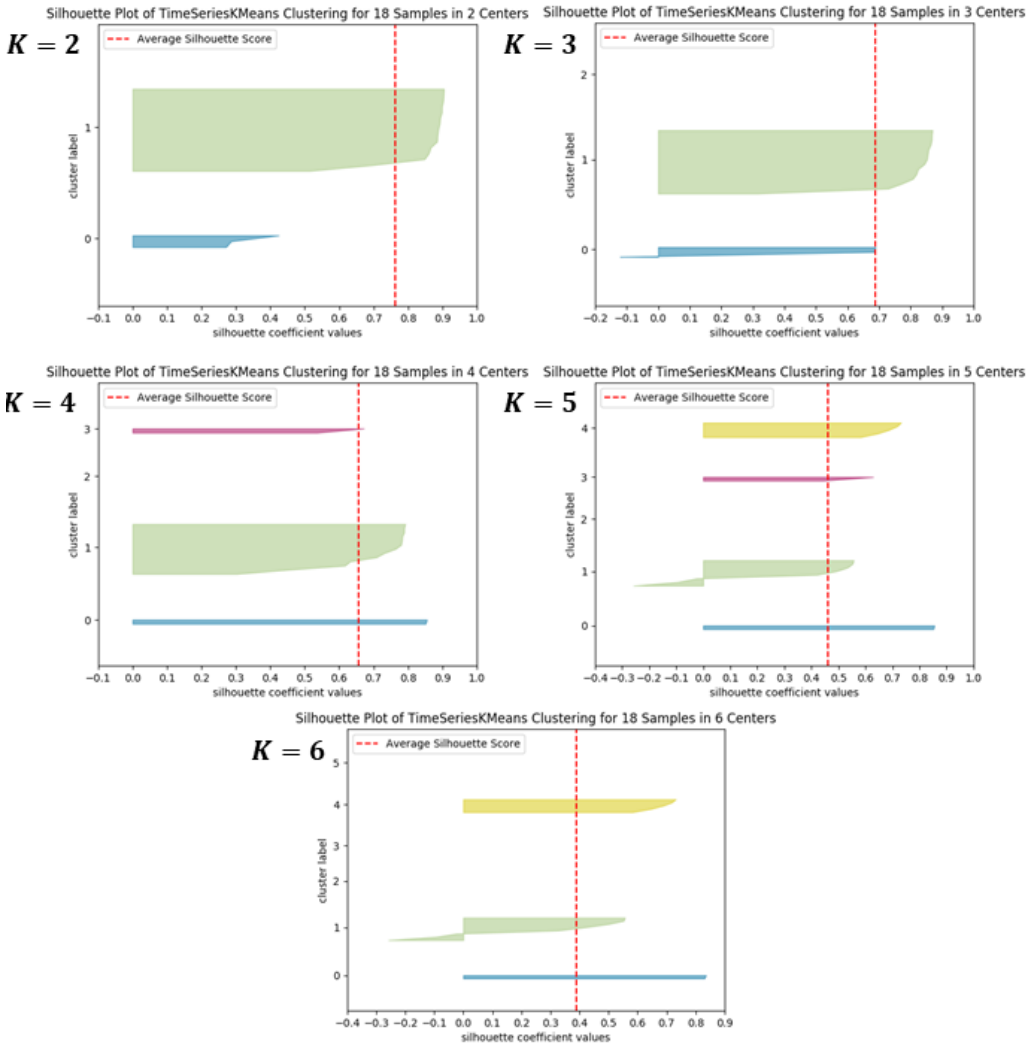

Figure 6.2: Silhouete scores for $K = 1, 2, \ldots, 6$.

A visual inspection of the graphs in Figure 6.3 reveals that the proposed methodology closely aligns with the actual series for all health regions, with no significant discrepancies being visually evident. Table 6.2 presents MAE and MAPE values for the proposed approach, ETS, and Prophet models. In most health regions, the proposed approach is better in terms of MAE and MAPE. The ETS is better than the proposed approach for Northeast I, North, and São Patricio II health regions. Prophet does not show better results for any health region. The hypothesis for ETS to overcome the proposed methodology is that Northeast I, North, and São Patrício II health regions are health regions with a smaller population than other regions. ETS is a time series forecasting method that does not require as much data for training as an MLP model. Smaller populations may not have all the health characteristics that larger regions do, which could negatively impact the MLP's learning process. The graphs in Figure 6.3 show that the time series of the Northeast I health region ranges from 46 to 70, while that of North ranges from 100 to 160,

and São Patrício II ranges from 160 to 220. On the other hand, the Central Health Region, one of the largest, ranges from 1800 to 2400, while the South Surrounding region ranges from 1000 to 1200. The analysis suggests that MLP models, which require a large volume of data (features) for training and testing, are more suitable for larger health regions, that have a greater diversity of features.

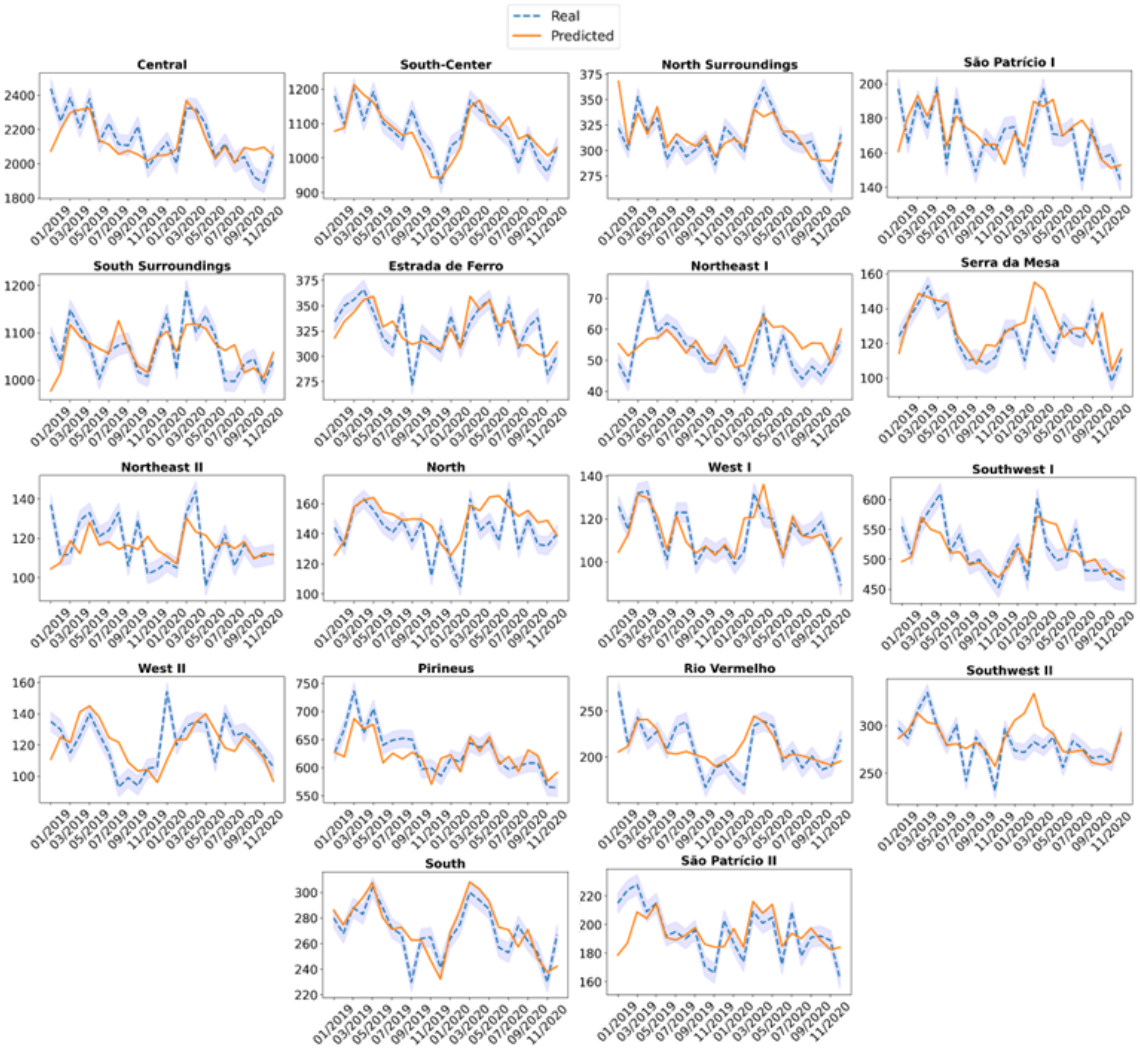

Figure 6.3: Results of the multivariate MLP applied to all health regions of Goiás.

Table 6.2: Results of MAE and MAPE for the prediction of the test set for the 18 health regions of Goiás

Health Regions	Proposed approach *		ETS		Prophet	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
Central	**77.3910**	**3.6490**	128.6836	5.7422	119.4458	5.4865
South-Center	**35.3885**	**3.3310**	55.0562	4.9244	58.1816	5.4811
North Surroundings	**10.6213**	**3.3179**	26.8542	7.9628	18.3211	5.8478
South Surroundings	**32.1666**	**3.0313**	32.9006	3.1557	50.0397	4.6028
Estrada de Ferro	**14.3650**	**4.4351**	19.3445	5.6765	18.4363	5.8752
Northeast I	5.2874	9.5220	**4.8953**	**8.9647**	5.8825	11.5876
Northeast II	**9.2735**	**8.0387**	9.3436	8.0598	10.2323	8.3717
North	11.7480	7.9385	**9.5333**	**6.6195**	12.1325	8.5205
West I	**6.0515**	**5.2985**	6.2235	5.4391	9.4667	8.4516
West II	**11.0499**	**9.4046**	11.8608	9.5868	13.3485	11.9695
Pirineus	**20.3933**	**3.2715**	29.7174	4.6125	32.2141	5.0748
Rio Vermelho	**14.6045**	**7.0024**	47.2651	18.5988	21.4760	9.8399
São Patrício I	**10.7698**	**6.2888**	12.0453	6.9699	13.4162	7.5975
Serra da Mesa	**9.4776**	**7.2025**	11.0501	9.0740	11.4359	9.5512
Southwest I	**21.4869**	**4.1051**	27.0962	5.0913	34.1443	6.5823
Southwest II	**14.5199**	**4.9459**	18.3881	6.2627	18.8340	6.8762
South	**10.0706**	**3.7820**	11.6000	4.2193	13.1251	5.0129
São Patrício II	11.7832	6.2087	**10.9664**	**5.7232**	16.1416	8.1900
Goiás **	**18.1360**	**5.5985**	26.2680	7.0380	26.4597	7.4955

** - Best result of five random seed initialization*

*** - Average of the results of the health regions*

CHAPTER 7

Conclusion

In the upcoming sections, the conclusions resulting from the analysis of the case studies presented in Chapters 5 and 6 will be discussed, along with suggestions for future research.

7.1 Conclusions regarding Case study I

This work proposed the use of the LMU model in a 24-month ahead prediction of live births in 18 health regions of the state of Goiás, in a univariate way. This proposal proved to be an interesting approach because, compared to Prophet and ETS models, LMU prevailed, achieving an average MAPE of 6.4614 and MAE of 19.9136, results that were very similar to those obtained with ARIMA.

Although ARIMA performed slightly better than LMU, LMU showed superior performance in predicting health trends in larger regions. Therefore, it would be worthwhile to test the LMU model on macro-regions that encompass multiple health regions within larger territories, or even entire states, to evaluate its performance at a larger scale. Additionaly, as it can see in Chapter 6 an neural network approach can be enchaced by the cross-series information, in this way, a deep learning approach such as LMU can be improved by training it as global models for health regions or other territory hierarchy.

7.2 Conclusions regarding Case study II

This work proposed an ANN with a clustering approach using a K-means algorithm, leveraging the time series behavior across health regions improves forecast accuracy while reducing model training time. The models were tested in different scenarios where the number of clusters changed, and the best results were reported with 2 clusters, which means, two models were trained to predict 18 health regions.

The predictive capacity of the models was tested in terms of MAE and MAPE and compared to statistical models like ETS and Prophet, the proposed ANN showed an improvement for almost all health regions. It is concluded that the proposed model

combining clustering using K-Means and ANN model is a good strategy, generating an average result of 5.5985 and 18.1360 for MAPE and MAE.

Moreover, unlike traditional univariate models that train local models for each time series, the proposed ANN model trains global models for all health regions based on its clusters, leveraging cross-series information while reduces the computational cost of training and optimizing each local model. Although the computational costs, in terms of time, of the proposed ANN model may be higher than some statistical models, it provides a practical and efficient solution for forecasting health trends across different regions by generating more accurate predictions.

References

[Abdullaeva et al. 2019]ABDULLAEVA, M. et al. Issues of multipurpose forecasting of ischemic strokes development. *Global journal of Medicine and Medical science*, v. 7, n. 8, p. 505–510, 2019.

[Adeyinka e Muhajarine 2020]ADEYINKA, D. A.; MUHAJARINE, N. Time series prediction of under-five mortality rates for nigeria: comparative analysis of artificial neural networks, holt-winters exponential smoothing and autoregressive integrated moving average models. *BMC medical research methodology*, Springer, v. 20, p. 1–11, 2020.

[Aguiar et al. 2022]AGUIAR, H. et al. Learning of cluster-based feature importance for electronic health record time-series. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2022. p. 161–179.

[Akazawa e Hashimoto 2022]AKAZAWA, M.; HASHIMOTO, K. Prediction of preterm birth using artificial intelligence: a systematic review. *Journal of Obstetrics and Gynaecology*, Taylor & Francis, v. 42, n. 6, p. 1662–1668, 2022.

[Albuquerque et al. 2017]ALBUQUERQUE, M. V. d. et al. Desigualdades regionais na saúde: mudanças observadas no brasil de 2000 a 2016. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 22, p. 1055–1064, 2017.

[Alvarez et al. 2010]ALVAREZ, F. M. et al. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 23, n. 8, p. 1230–1243, 2010.

[Ashfaq et al. 2019]ASHFAQ, A. et al. Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, Elsevier, v. 97, p. 103256, 2019.

[Bandara, Bergmeir e Smyl 2020]BANDARA, K.; BERGMEIR, C.; SMYL, S. Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert systems with applications*, Elsevier, v. 140, p. 112896, 2020.

[Bandara et al. 2021]BANDARA, K. et al. Improving the accuracy of global forecasting models using time series data augmentation. *Pattern Recognition*, Elsevier, v. 120, p. 108148, 2021.

[Bravo e Coelho 2020]BRAVO, J. M.; COELHO, E. Modelling monthly births and deaths using seasonal forecasting methods as an input for population estimates. *Demography of Population Health, Aging and Health Expenditures*, Springer, p. 203–222, 2020.

[BrOO]BROO. https://observatorioobstetricobr.org/. Accessed: 2023-01-11.

[BVS]BVS. https://bvsms.saude.gov.br/28-5-dia-nacional-de-reducao-da-mortalidade- Accessed: 2023-01-11.

[Cassetti et al. 2008]CASSETTI, T. et al. Cancer incidence in men: a cluster analysis of spatial patterns. *BMC cancer*, BioMed Central, v. 8, n. 1, p. 1–9, 2008.

[Castro et al. 2018]CASTRO, M. C. et al. Implications of zika virus and congenital zika syndrome for the number of live births in brazil. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 115, n. 24, p. 6177–6182, 2018.

[Dantas e Oliveira 2018]DANTAS, T. M.; OLIVEIRA, F. L. C. Improving time series forecasting: An approach combining bootstrap aggregation, clusters and exponential smoothing. *International Journal of Forecasting*, Elsevier, v. 34, n. 4, p. 748–761, 2018.

[Duarte e Teixeira 2021]DUARTE, H. F. F. L.; TEIXEIRA, E. C. Efeito do nível de escolaridade sobre a fecundidade no brasil. *Economia & Região*, v. 9, n. 1, p. 167–185, 2021.

[Elhag e Abu-Zinadah 2020]ELHAG, A. A.; ABU-ZINADAH, H. Forecasting under applying machine learning and statistical models. *Thermal Science*, v. 24, n. Suppl. 1, p. 131–137, 2020.

[Forsyth 2016]FORSYTH, D. *Probability and statistics for computer science*. [S.l.]: Springer, 2016.

[Gardner e Dorling 1998]GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

[Gómez-Losada, Pires e Pino-Mejías 2018]GÓMEZ-LOSADA, Á.; PIRES, J. C. M.; PINO-MEJÍAS, R. Modelling background air pollution exposure in urban environments: Implications for epidemiological research. *Environmental Modelling & Software*, Elsevier, v. 106, p. 13–21, 2018.

[González 2019]GONZÁLEZ, F. D. *Federated learning for time series forecasting using LSTM networks: Exploiting similarities through clustering*. 2019.

[Hartmann et al. 2015]HARTMANN, C. et al. Exploiting big data in time series forecasting: A cross-sectional approach. In: IEEE. *2015 IEEE international conference on data science and advanced analytics (DSAA)*. [S.l.], 2015. p. 1–10.

[Hewamalage, Bergmeir e Bandara 2021]HEWAMALAGE, H.; BERGMEIR, C.; BANDARA, K. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, Elsevier, v. 37, n. 1, p. 388–427, 2021.

[Huang et al. 2020]HUANG, E. et al. Aortic pressure forecasting with deep learning. In: IEEE. *2020 Computing in Cardiology*. [S.l.], 2020. p. 1–4.

[Imtiaz et al. 2020]IMTIAZ, S. et al. Privacy preserving time-series forecasting of user health data streams. In: IEEE. *2020 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2020. p. 3428–3437.

[James e Menzies 2020]JAMES, N.; MENZIES, M. Cluster-based dual evolution for multivariate time series: Analyzing covid-19<? a3b2 show [editpick]?>. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, AIP Publishing LLC, v. 30, n. 6, p. 061108, 2020.

[Januschowski et al. 2020]JANUSCHOWSKI, T. et al. Criteria for classifying forecasting methods. *International Journal of Forecasting*, Elsevier, v. 36, n. 1, p. 167–177, 2020.

[Kamber, Pei et al. 2001]KAMBER, M.; PEI, J. et al. *Data mining: Concepts and techniques*. [S.l.]: Morgan Kaufmann Publishers San Francisco, 2001.

[Keogh e Ratanamahatana 2005]KEOGH, E.; RATANAMAHATANA, C. A. Exact indexing of dynamic time warping. *Knowledge and information systems*, Springer, v. 7, p. 358–386, 2005.

[Lippmann 1987]LIPPMANN, R. An introduction to computing with neural nets. *IEEE Assp magazine*, IEEE, v. 4, n. 2, p. 4–22, 1987.

[LORENA]LORENA, D. A. R. P. Indicadores de mortalidade materna em goiás no período de 1999 a 2005: implicações para a enfermagem. *Oeste*, v. 244, n. 4, p. 9.

[Łuczak e Kalinowski 2022]ŁUCZAK, A.; KALINOWSKI, S. Fuzzy clustering methods to identify the epidemiological situation and its changes in european countries during covid-19. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 24, n. 1, p. 14, 2022.

[MacQueen 1967]MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967. p. 281–297. Disponível em: <https://projecteuclid.org/euclid.bsmsp/1200512992>.

[Maimon e Rokach 2005]MAIMON, O.; ROKACH, L. Data mining and knowledge discovery handbook. Springer, 2005.

[Mäkipää 2021]MÄKIPÄÄ, A.-J. *Forecasting Emergency Department arrivals with Facebook Prophet library*. Dissertação (B.S. thesis), 2021.

[McCloskey e Poon 2017]MCCLOSKEY, R. M.; POON, A. F. A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 13, n. 11, p. e1005868, 2017.

[McCulloch e Pitts 1943]MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

[McGregor, Watkin e Cox 2004]MCGREGOR, G. R.; WATKIN, H. A.; COX, M. Relationships between the seasonality of temperature and ischaemic heart disease mortality: implications for climate based health forecasting. *Climate Research*, v. 25, n. 3, p. 253–263, 2004.

[MH]MH. https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2022/boletim-epidemiologico-vol-53-no47/view. Accessed: 2023-01-13.

[Müller 2007]MÜLLER, M. Dynamic time warping. *Information retrieval for music and motion*, Springer, p. 69–84, 2007.

[Neamțu et al. 2021]NEAMȚU, B. M. et al. A decision-tree approach to assist in forecasting the outcomes of the neonatal brain injury. *International Journal of Environmental Research and Public Health*, MDPI, v. 18, n. 9, p. 4807, 2021.

[Oliveira 2010]OLIVEIRA, A. B. *Usando redes neurais para estimação da volatilidade: redes neurais e modelo híbrido GARCH aumentado por redes neurais*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

[Orabona e Tommasi 2017]ORABONA, F.; TOMMASI, T. Training deep networks without learning rates through coin betting. *Advances in Neural Information Processing Systems*, v. 30, 2017.

[Orlandic, Valdes e Atienza 2021]ORLANDIC, L.; VALDES, A. A.; ATIENZA, D. Wearable and continuous prediction of passage of time perception for monitoring mental health. In: IEEE. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.], 2021. p. 444–449.

[PAHO]PAHO. https://www.paho.org/pt/node/63100. Accessed: 2023-01-12.

[Petitjean et al. 2014]PETITJEAN, F. et al. Dynamic time warping averaging of time series allows faster and more accurate classification. In: IEEE. *2014 IEEE international conference on data mining*. [S.l.], 2014. p. 470–479.

[Pícoli, Cazola e Lemos 2017]PÍCOLI, R. P.; CAZOLA, L. H. d. O.; LEMOS, E. F. Mortalidade materna segundo raça/cor, em mato grosso do sul, brasil, de 2010 a 2015. *Revista Brasileira de Saúde Materno Infantil*, SciELO Brasil, v. 17, p. 729–737, 2017.

[Pinto et al. 2022]PINTO, K. B. et al. Panorama de mortalidade materna no brasil por causas obstétricas diretas. *Research, Society and Development*, v. 11, n. 6, p. e17111628753–e17111628753, 2022.

[Piryatinska et al. 2009]PIRYATINSKA, A. et al. Automated detection of neonate eeg sleep stages. *Computer methods and programs in biomedicine*, Elsevier, v. 95, n. 1, p. 31–46, 2009.

[Rauber 2005]RAUBER, T. W. Redes neurais artificiais. *Universidade Federal do Espírito Santo*, p. 29, 2005.

[Reinsel 2003]REINSEL, G. C. *Elements of multivariate time series analysis*. [S.l.]: Springer Science & Business Media, 2003.

[Ribeiro et al. 2019]RIBEIRO, R. C. M. et al. Forecasting incidence of tuberculosis cases in brazil based on various univariate time-series models. *International Journal for Innovation Education and Research*, v. 7, n. 1, p. 894–909, 2019.

[Rosenblatt 1958]ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.

[Rumelhart, Hinton e Williams 1985]RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. *Learning internal representations by error propagation*. [S.l.], 1985.

[Selim e Ismail 1984]SELIM, S. Z.; ISMAIL, M. A. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, n. 1, p. 81–87, 1984.

[Souza e Amorim 2021]SOUZA, A. S. R.; AMORIM, M. M. R. Mortalidade materna pela covid-19 no brasil. *Revista Brasileira de Saúde Materno Infantil*, SciELO Brasil, v. 21, p. 253–256, 2021.

[Taloba et al. 2022]TALOBA, A. I. et al. Estimation and prediction of hospitalization and medical care costs using regression in machine learning. *Journal of Healthcare Engineering*, Hindawi, v. 2022, 2022.

[Tomašev et al. 2021]TOMAŠEV, N. et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols*, Nature Publishing Group UK London, v. 16, n. 6, p. 2765–2787, 2021.

[Tuominen et al. 2021]TUOMINEN, J. et al. Forecasting daily arrivals and peak occupancy in a combined emergency department. 2021.

[Viana et al. 2015]VIANA, A. L. D. et al. Tipologia das regiões de saúde: condicionantes estruturais para a regionalização no brasil. *Saúde e Sociedade*, SciELO Public Health, v. 24, p. 413–422, 2015.

[Voelker, Kajić e Eliasmith 2019]VOELKER, A.; KAJIĆ, I.; ELIASMITH, C. Legendre memory units: Continuous-time representation in recurrent neural networks. *Advances in neural information processing systems*, v. 32, 2019.

[Vollmer et al. 2021]VOLLMER, M. A. et al. A unified machine learning approach to time series forecasting applied to demand at emergency departments. *BMC Emergency Medicine*, BioMed Central, v. 21, n. 1, p. 1–14, 2021.

[Wang et al. 2018]WANG, K. et al. Deep belief network based k-means cluster approach for short-term wind power forecasting. *Energy*, Elsevier, v. 165, p. 840–852, 2018.

[WHO]WHO. https://brasil.un.org/pt-br/sdgs/3. Accessed: 2023-01-11.

[Widrow e Hoff 1960]WIDROW, B.; HOFF, M. E. *Adaptive switching circuits*. [S.l.], 1960.

[Zhang et al. 2022]ZHANG, Y. et al. The prediction of preterm birth using time-series technology-based machine learning: Retrospective cohort study. *JMIR Medical Informatics*, JMIR Publications Toronto, Canada, v. 10, n. 6, p. e33835, 2022.