



Can Global Machine Learning Models Improve the Accuracy of Small Area Age-Sex Population Forecast

Author

Mr. Chi Zhang
Mr. Eric Luanzon
Mr. Haitong Gao
Miss. Meijun Yue
Miss. Yuexin Li

Supervisor

Dr. Irina Grossman
Dr. Kasun Bandara
Prof. Michael Kirley

A report submitted in fulfilment of the requirements
for the degree of Master of Data Science
in the
School of Mathematics and Statistics, University of Melbourne

October 26, 2022

Student number: 1067750, 1218692, 1068254, 1190161, 959634

Project type: Industrial Project

Subject code: MAST90106 & MAST90107

Credit points: 25

Abstract

The accuracy of population forecasts, especially in the small areas with less research at the current stage, is essential in society. It provides recommendations and instructions for the community or government to plan future infrastructure facilities in specific regions. However, due to limited research, the current method might provide inaccurate forecasts. This project aims to develop a specific machine learning model, Long-Short Term Memory (LSTM), that helps to provide an accurate small-area age-sex population projection and compare it to a state-of-the-art statistical model in the demographic domain. This report firstly presents the project's background and challenges, and the research question and the relevant project goals are stated. Then, previous papers and models that provide the potential solution to the topic are presented, followed by the higher-level description of the designed pipeline of the project. The visualisation and tables are obtained based on preprocessing of the population data, which collected from the Australian Bureau of Statistic. The characteristics of the data from the visualisation and tables will be the support evidence in selecting the model for implementation. Afterwards, the LSTM algorithm is selected for the model's development, following the process of repeatedly testing and self-improvement logic by changing the hyper-parameters and the in-built logic. Besides, the Synthetic Migration Model, which is the demographic benchmark model (state-of-the-art model), is reconstructed during the project and is applied for comparison to the LSTM Model to fulfil the project goal. Following the LSTM and the Synthetic Migration Model illustration, the results are computed based on the evaluation matrix and provided with tables. Discussions are then made based on the model's outcome and the related articles to present the model's performance and limitations. In order to answer the project's goals and fulfil the client's requirements, recommendations are provided based on the result, the data, and the model's characteristics from both positive and negative perspectives. In summarising the report, the project's conclusion and the contribution to the research domain are made. Last but not least, the limitations of the research are provided, accompanied by an illustration of the future direction of the study.

Keywords: Population Forecasting, Time Series, Synthetic Migration Model, LSTM

Declaration of Authorship

We, Chi Zhang, Eric Luanzon, Haitong Gao, Meijun Yue, and Yuexin Li declare that this report titled, ‘Can Global Machine Learning Models Improve the Accuracy of Small Area Age-Sex Population Forecast’ and the work presented in it are our own. We confirm that:

- this report does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- the report is 8111 words in length (excluding text in images, table, bibliographies and appendices).

Signed: Chi Zhang

Signed: Eric Luanzon

Signed: Haitong Gao

Signed: Meijun Yue

Signed: Yuexin Li

Date: 26 / 10 / 2022

Acknowledgements

We would like to thank our client Dr. Irina Grossman of the Demography and Ageing Unit, Melbourne School of Population and Global Health. Dr. Grossman was always there for us and kept following our real-time progress. Whenever we encountered trouble or an incorrect implementation track, she would guide us in the right direction.

We would like to thank our supervisors, Prof. Michael Kirley and Dr. Kasun Bandara of the School of Computing and Information Systems at the University of Melbourne. Without their guidance and suggestions throughout the year, the industrial project could not have been completed successfully.

Thank you.

Authors: Chi Zhang, Eric Luanzon, Haitong Gao, Meijun Yue, Yuexin Li

Contents

Abstract	i
Declaration of Authorship	ii
Acknowledgements	iii
List of Figures	vi
List of Tables	vii
Abbreviations	viii
1 Introduction	1
1.1 Project Background	1
1.2 Research Question Overview	2
1.3 Methodology Outline	2
1.4 Project Challenges	2
1.5 Project Structure and Goals	3
2 Literature Review	5
2.1 Purpose	5
2.2 Challenge of Small Area Population Forecasting	5
2.3 Existing Models	6
2.4 Suitable Machine Learning Model	6
2.5 Potential Disadvantage of Machine Learning Models	7
3 Data Science Pipeline	8
4 Data	9
4.1 Data Collection and Description	9
4.2 Data Preprocessing	9
4.3 Data Analysis	10

4.4	Data Characteristic	11
4.4.1	Data Structure	11
4.4.2	Data Sparsity	11
4.4.3	Data / Regions' Correlation	12
5	Model	14
5.1	Reason of Applying Models	14
5.2	Long-Short Term Memory	15
5.2.1	Theoretical Description of LSTM	15
5.2.2	Theoretical Description of Sliding Window	16
5.2.3	Application Description of LSTM Model (Type 1, 2, 3)	17
5.3	Synthetic Migration Model	20
5.3.1	Theoretical Description	21
5.3.2	Extra Variables & Models	21
5.3.3	Project Application Description (2006, 2011)	21
6	Evaluation	23
6.1	Evaluation Matrix	23
6.2	Evaluation Result	24
6.3	Evaluation Measurement	24
6.4	Result Interpretation	25
7	Discussion	26
7.1	Discussion of Different Models	26
7.2	Limitation of LSTM and Synthetic Migration Model	27
8	Recommendation	29
8.1	Positive	29
8.2	Negative	30
8.3	Overall Recommendation	30
9	Conclusion	31
	Bibliography	33
A	Appendix	35
A.1	Code & Meeting Minutes Repository Link	35
A.2	Pseudo Code for Sliding Window	35
A.3	Pseudo Code for Rolling Update	36
A.4	Pseudo Code for the LSTM	36

List of Figures

4.1	Goulburn - Yass's Age-Sex Population Data Structure	10
4.2	Melbourne City Population Pyramid	11
4.3	Average Population Pyramid	12
4.4	Comparison of Plot Trends between Latrobe Valley & Barossa	13
4.5	Comparison of Plot Trends between Wellington & Gawler - Two Wells . .	13
5.1	Repeating module in an LSTM with four interacting layers (Colah, 2015).	15
5.2	Sliding window with window width = 3 for male's training data.	16
5.3	LSTM rolling update for male testing data with window width = 3. The black word means true data from the training set; the red word means predicted data. Each data contains 18 age groups.	18

List of Tables

4.1	Sub-sample of Age-Sex Population Data Mean among all Small Areas . .	10
5.1	Window width = 3 for training set in date between 1991 and 2001 data. .	16
6.1	Forecast Errors of Population in Age-Sex Level	24
6.2	Forecast Errors of Total Population	24

Abbreviations

ABS	A ustralian B ureau of S tatistics
APE	A bsolute P ercentage E rror
MAPE	M ean A bsolute P ercentage E rror
LSTM	L ong S hort T erm M emory
RELU	R ectified L inear U nit A ctivation F unction
RNN	R ecurrent N eural N etwork
SA3	S tatistical A rea L evel 3
VBA	V isual B asic for A pplication

Chapter 1

Introduction

1.1 Project Background

Population forecasting, a topic in demography, is valued by organisations such as the government, which plans for the industries such as healthcare and education that require a specific age-sex population. The same for companies that need guidance in their marketing decisions, as the different age-sex demographics of different regions will allow them to make different marketing strategies. Based on the research, this topic has already been investigated on the national level but less at the small-area level. In order to expand the investigation in small areas' population forecast, the tasks of model reconstruction or implementation can be proceeded to contribute to this topic's study.

As stated by the client, previous research has been more concerned with predicting the total population of a region. Here we focus more on age-sex population projections for small areas. In the existing research, more external variables are required to collect for age-sex population, however, our goal is to use less input data and achieve good prediction results. In this case, it is worthwhile to construct a model with techniques from the machine learning aspect and fewer feature requirements to potentially improve or retain the prediction accuracy compared with the complicated statistical model.

1.2 Research Question Overview

In this project, based on the client's requirement and relevant research on population forecasting, the research question of 'Can Global Machine Learning Model Improve the Accuracy of Small Area Age-Sex Population Forecasts?' is obtained. In order to address this question, the core task is to follow the data science pipeline to construct the target model. This machine learning model will only use population data with the age-sex cohort, which is different from the previous demographic model that also uses fertility and mortality data.

1.3 Methodology Outline

The methodology applied for model construction follows the data science pipeline. After the preliminary understanding of the population forecasting based on the age-sex population data from the previous research, the method of data preprocessing and analysis has proceeded with a similar approach based on the data provided by the client and the Australian Bureau of Statistics (ABS). Due to the data characteristic obtained from the analysis, the machine learning method Long-Short Term Memory (LSTM) model that suits the requirement is selected for construction. By repeatedly testing the model after self-improvement, including tuning the hyper-parameters and related methods inside the algorithm, the comparison is then made with the Synthetic Migration Model (the benchmark model in this project).

1.4 Project Challenges

Several challenges were met from the data and the applied model's aspects during the project's progress.

From a data perspective, the area populations are split into age-sex groups. Consequently, there would be sparsity in the data especially when taking into account the natural lifespans of people. Also, since the age-sex cohort's data are hard to collect, it results in the issue of only having a short time series among each investigated instance.

When applying a machine learning method, especially in a Neuron Network type, insufficient training data may prevent the model from capturing the data pattern and may not result in an accurate prediction. Moreover, compared to the statistical population prediction model, the applied dataset includes fewer features (information) and might deteriorate the model's performance.

As for the model, the applied LSTM is a type of Neuron Network that cannot provide much interpretability after fitting the model. While increasing the number of epochs when fitting the LSTM model can capture the data trends better, this would lead to greater computational requirements. Other challenges include the logic behind the LSTM model's implementation, data preprocessing into a format that can be input into the model, and balancing out the data available for training against the additional error introduced from iterative forecasts. All these challenges are considered while developing the model following the data science pipeline. They will be referred to when providing recommendations based on the results from the delivered models and provide pathways for further study if some of these issues are currently unresolved.

1.5 Project Structure and Goals

In order to obtain the relevant information that states the research question, two main goals were set following the process of the project's implementation:

1. Discover the features and logic that helps improve the performance of an LSTM model on the forecasts of small area populations by age and sex.
2. Evaluate whether the designed LSTM model with fewer features could outperform the state-of-the-art Statistical Model (Synthetic Migration Model).

The project is divided into sections described in the following chapters to achieve these two goals. Chapter 2 presents the literature review about population forecasting and the relevant statistical or machine learning model developed in the previous studies. This chapter provides evidence and guidance for the model's construction. In Chapter 3, the methodology designed as a data science pipeline is explained in a high-level framework to illustrate the process of solving the research question. While in Chapter 4, the related information of the applied dataset is revealed along with the preprocessing

and analysis step. Then, the illustration of both the benchmark model (the Synthetic Migration Model) and the constructed LSTM model is presented from both theoretical and applicational perspectives in Chapter 5. Followed by Chapter 6, the evaluation of related models is computed and is discussed in Chapter 7, which will refer back to the two main goals of the research. Then, the recommendation will be provided in Chapter 8 based on the result and the discussion, which recall the issue mentioned in the 'Project's Challenges' section and answer the research question. Finally, Chapter 9 concludes the work and suggests the avenues for further research based on our implementation.

Chapter 2

Literature Review

2.1 Purpose

Small area population projections play a crucial role in the future infrastructure decisions and many other planning and development decisions by business and local governments (Wilson et al., 2021). The allocation of health care, education, housing, transportation and other resources in a small area is a decision that the government needs to make based on the projected population of different age groups in the small area. Therefore, the prediction accuracy for small regional age groups becomes particularly critical.

2.2 Challenge of Small Area Population Forecasting

Compared with national or sizeable sub-national region population projection, accurate population projections for small regions require detailed data on birth and death rates and migration, which are often missing for small regions, so it brings a significant challenge for small area forecasting (Wilson et al., 2021). Besides geographical factors like geocoding uncertainty, boundary change will also be the reason for shorter time series or data loss (Wilson et al., 2021). At the same time, from the perspective of modeling prediction, the population of a small area is more likely to be affected by natural change, population migration, births and deaths, so some special events in a certain area in a certain year lead to significant fluctuations in the population will have a misleading

effect on the learning of the model (Rees et al., 2004). These challenges will lead to a significant increase in errors in population projection.

2.3 Existing Models

Demographers have developed several models to forecast the age-sex population and total population in a small area with limited input data. One is the Hamilton-Perry model, which is easy to implement and use after getting the age and sex population Change Ratio. It is the ratio of the cohort population size with itself several years ago (The specific year parameters are set by the model user), and it will deal with the problem of the sparsity of data (Dockery et al., 2020). Besides, Wilson (2022) introduced the synthetic migration cohort-component model. Compared with the Hamilton-Perry model, the synthetic migration cohort-component model is more accurate and is not limited by the locally-specific migration data. The directional migration in local area cohort-component calculation is involved instead of requiring detailed migration data input which is often unavailable for small areas. Until now, Wilson's model is the best performing model using the given dataset and accompanied by the generated external variables' information, so it is set as our benchmark model. The Synthetic Migration Model will be illustrated for reference since it is applied as the benchmark for comparison to our implemented model. Details of the Synthetic Migration Model will be introduced in a later section.

2.4 Suitable Machine Learning Model

For the machine learning aspect, models in machine learning have been widely used in time series forecasting in recent years. In the current small area population projections, the population data is no longer limited to the total population of an area, but detailed population data by age and gender groupings for multiple areas. In this context, more complex models like neural networks might better handle large datasets (Hewamalage et al., 2021). Artificial neural networks become popular by simulating the operation of brain neurons, letting the input data go through the neurons and get the output for prediction. However, according to Grossman, Wilson and Temple's paper (2022), this early version of neural networks could only refer to the nearest neuron input; it could

not get the information from earlier neurons which means they do not have memory. Recurrent Neural Networks (RNNs) is an improved version of neural networks by adding "hidden states", which can continue to put the output of a unit back to next time step prediction after weighting it (Rumelhart et al., 1986). However, as the time series increases, the accumulation of weights may cause the problem of vanishing and exploding gradients. Long Short-Term Memory (LSTM) is a Recurrent Neural Network based architecture widely used in time series forecasting which effectively solves the problem of vanishing and exploding gradient in traditional RNNs (Graves, 2012). LSTM can add or remove information in cell states and use three gates to control that. A sigmoid layer called Input Gate controls what values to update, Forget Gate controls how much information to forget from the last memory cell, and the Output Gate controls the output based on the current cell state. This strategy allows LSTM to remember more extended data patterns and make the prediction.

2.5 Potential Disadvantage of Machine Learning Models

The LSTM model is suitable for time series forecasting. However, Makridakis et al.'s research (2018) suggests it may not help to increase the forecast accuracy compared with the statistical model for its computational requirement is more significant and higher data volume requirements as well. Another critical issue is the inability to see the process of predicting the output of the neural network model, which is not acceptable to "practitioners" (Makridakis et al., 2018). Researchers need to know how the prediction result is generated and how the accuracy could be improved and optimised instead of an invisible black box and auto-generated hyperparameters, which could not be explained. Similarly, the estimation of the future must take into account the uncertainty interval. Nevertheless, according to Makridakis's paper (2018), lack of the prediction of the model uncertainty interval is precisely what is missing in machine learning models. Another essential factor is that the time series may be too short to be modelled by machine learning or neural networks requiring large data volumes (Hewamalage et al., 2021). Last but not least, Ahmed et al. (2010) observation of how to do the data preprocessing may differ the model performance significantly. A further discussion about the implementation result will be shown at the end.

Chapter 3

Data Science Pipeline

According to the data science pipeline, the project is designed under the following outline. Starting with data collection, then data analysis, preprocessing, model construction and evaluation, ending with the recommendation and conclusion based on the result. We first collected the age-sex population dataset from the client, a preliminary analysis is done to understand the data's characteristics by computing and visualising the descriptive statistics. Then, we select the machine learning model (LSTM) that best matches the data characteristics, and further preprocessing of data will be proceeded based on the selected model. With the selected method (LSTM), the model is constructed and is kept improving following an iterative process. Evaluation and benchmark comparison is required when developing or improving the model. By changing the internal logic of the model each time, we could obtain a new model and compute its performance. Once it becomes hard to achieve significant improvement, we will stop the model improvement process. Based on the delivered model(s) and the relevant research, discussions are presented, and recommendations will be provided based on the results. In this industrial project, all the implementation is delivered in response to the research question and fulfils the client's requirements. The data analysis and the model's implementation steps answer the research question, while the recommendation step is stressed, contributing to the client's needs. These steps are mainly focused on the whole pipeline and will be interpreted in more detail in the following chapters.

Chapter 4

Data

4.1 Data Collection and Description

Based on the dataset from ABS (ABS, 2021), the population in small areas of Australia is divided into 36 groups according to sex and age. Since the data was provided in Excel, the formatting was removed so the table could be read into python as a DataFrame using the pandas library in a CSV format. The DataFrame includes 351 SA3 regions' time series in Australia on which the forecasting will be done. As one of the levels of geographic areas Structures of the Australian Statistical Geography Standard, SA3s are designed to provide a regional breakdown of Australia. The dataset of age-sex groups was studied primarily in this project. Then, accompanied by the application of the age-sex dataset, extra variables such as the total population from the same dataset would be introduced later when improving the model to see if any improvement can be obtained.

4.2 Data Preprocessing

The preprocessing step includes applying exclusion criteria to fit the model on SA3s with sufficient data points; only the regions with at least 1000 people in the total population across all age and sex groups are included in the model. Those excluded from the model would be aggregated into a remainder group which would be treated as its area with its time series. Areas surpassing the exclusion criteria, such as Gungahlin, are also included if their total population did not meet the requirement in some years. Additionally, areas

with no data for some years but have recorded counts in later years are included in the remainder. These null values are assigned zero population, and all these regions are placed into the remaining group. Furthermore, the areas which only contain null values in their time series across the years are discarded. After preprocessing, the dataset applied in this project contains 325 regions at the SA3 level.

4.3 Data Analysis

After the preprocessing step, the data type and the descriptive statistics were first studied. The dataset used for the study includes the total population and 36 sex-age cohorts populations for 325 SA3 regions from 1991 to 2011. To illustrate the data structure of each region, the data of Goulburn - Yass, one of the small areas, is shown in Figure 4.1 for reference.

Year	SA3 Code	male			female			Total
		0-4	...	85+	0-4	...	85+	
1991	10101	2603	...	159	2400	...	353	61667
1992	10101	2605	...	163	2392	...	383	61751
...								
2011	10101	2350	...	490	2225	...	942	69775

FIGURE 4.1: Goulburn - Yass's Age-Sex Population Data Structure

Since the time series data relies more on sequential logic but not descriptive statistics, the maximum and minimum become less informative. We selected the sample, including 36 age-sex cohorts in the year 2000 from regions with the same yearly trend for fundamental analysis. Presented below, Table 4.1 records a sample of mean values of the age-sex population for reference as the preliminary investigation of the data sparsity.

age-sex	m5-9	m25-29	m35-39	m55-59	m65-69	m85+
mean	2114.1	2199.7	2285.6	1497.4	1014.5	237.0
age-sex	f5-9	f25-29	f35-39	f55-59	f65-69	f85+
mean	2006.5	2213.9	2310.3	1446.3	1054.5	535.6

TABLE 4.1: Sub-sample of Age-Sex Population Data Mean among all Small Areas

4.4 Data Characteristic

4.4.1 Data Structure

The dataset is a yearly frequency short time series with 21 data points in each region, having a short time series with 1991-2001 for the training set but requires prediction of 2002-2011. In this case, models such as ARIMA, which by default requires more than 50 observations in each time series (Box & Tiao, 1975), are not preferable. Besides, forecasting is based on a two-level hierarchical structure that requires multivariate input with a 3-dimensional input time series matrix, also known as a hierarchical time series.

4.4.2 Data Sparsity

As shown in Table 4.1, there is a large gap between the older group (e.g., 85+) and the younger group (e.g., 35-39), suggesting a data sparsity in the older group. To further identify this characteristic, the pyramid plot Figure 4.2 is plotted based on the total population in the year 2000 Melbourne City for reference. The pyramid diagram shows that the population size of the older group is significantly smaller, which may lead to bias in predicting the older group when modelling.

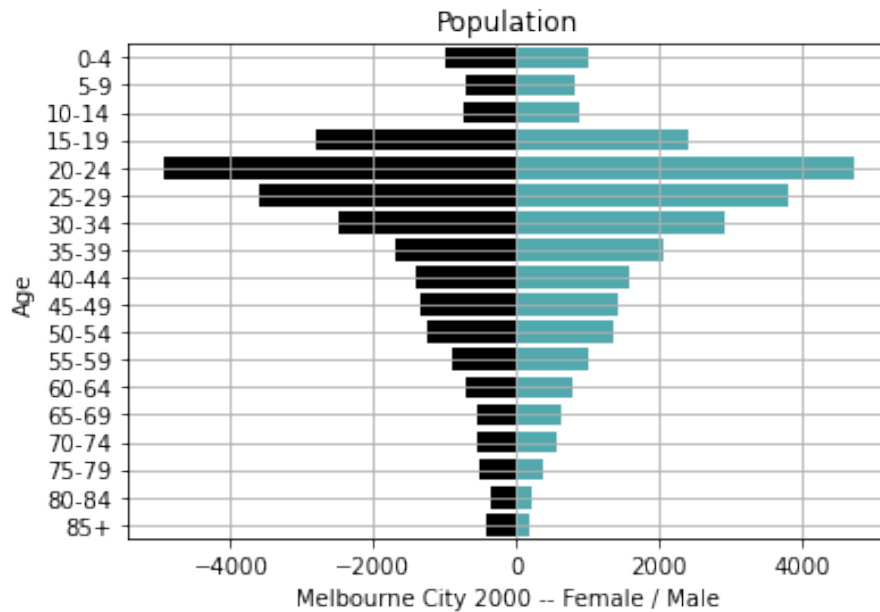


FIGURE 4.2: Melbourne City Population Pyramid

Not only from a local small-area perspective investigated, but the aggregated population in Australia (global perspective) is also computed and plot Figure 4.3 for reference. A similar pattern has been observed from this pyramid plot by having sparse data in the elder-age groups, indicating that the natural constraint of the population exists and leads to challenges of constructing the machine learning method to capture the population changing trend.

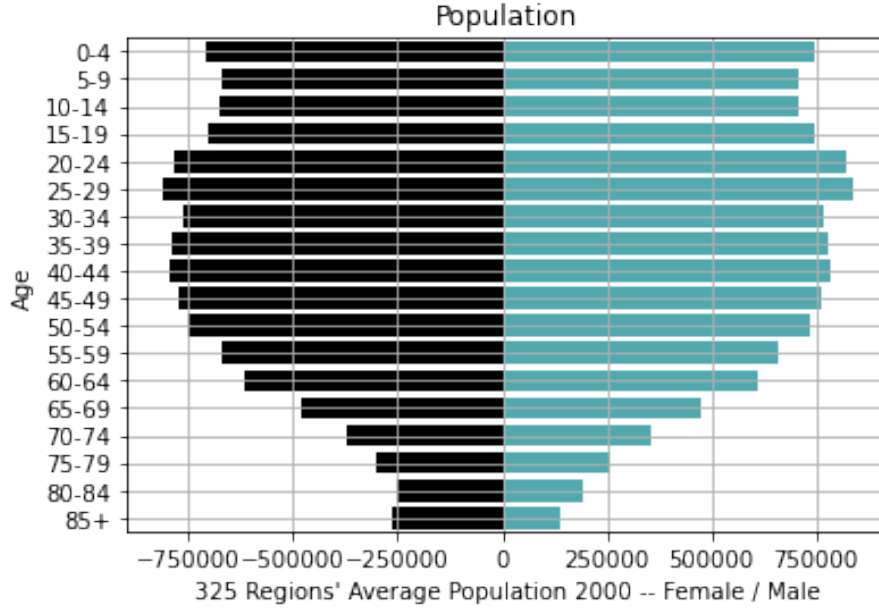


FIGURE 4.3: Average Population Pyramid

4.4.3 Data / Regions' Correlation

In addition to the data sparsity, the difference between each region's population changing pattern is investigated by observing the trend plots. From Figure 4.4, the trend plots of Latrobe Valley and Barossa are plotted for reference and present a distinct population trend pattern among the selected time series. In this case, differences among latent variables such as birth rates and population base exist in regions, which might lead to an inaccurate result. An identical statistical model constructed for forecasting with the same parameters might fail to capture the potential difference between areas.

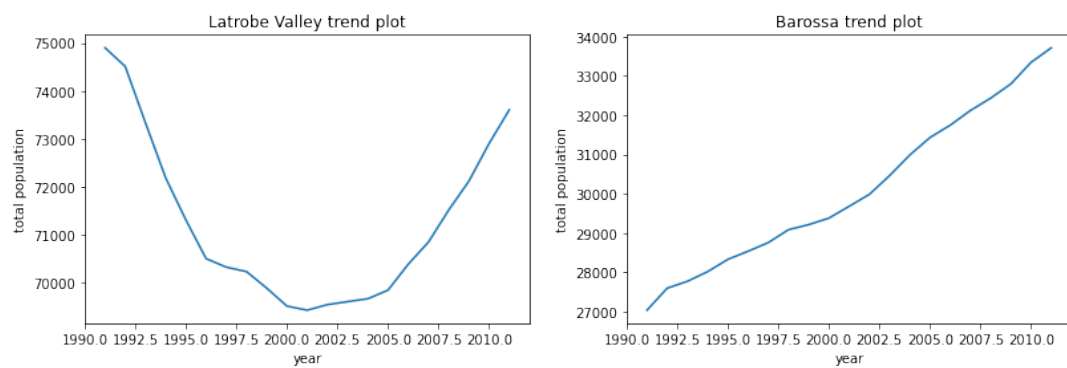


FIGURE 4.4: Comparison of Plot Trends between Latrobe Valley & Barossa

Similarly, presented in Figure 4.5, a distinct pattern can be observed between Wellington and Gawler-Two Wells. However, by comparing the trend plots from Figure 4.4 and Figure 4.5, respectively, a similar pattern is obtained. In this case, it is reasonable to apply a grouping method before fitting the implemented model, potentially improving forecasting accuracy.

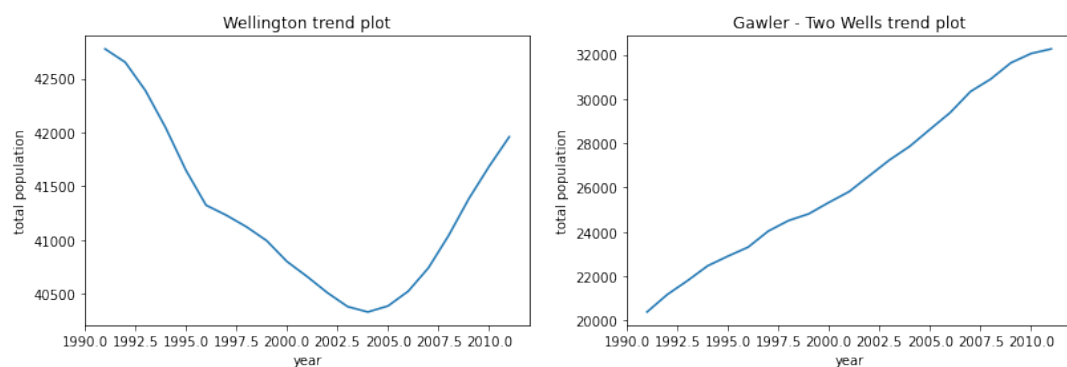


FIGURE 4.5: Comparison of Plot Trends between Wellington & Gawler - Two Wells

Chapter 5

Model

5.1 Reason of Applying Models

Based on the data's characteristics, with the provided sparse data, especially in the older age cohorts, the regular statistical model requires more parameters that fit the sparse feature and decrease the model's performance due to the zero values in a specific instance's group. Besides, since the population trend among each region is different based on the visualisation of the example areas in the 'Data' chapter, it is hard to obtain an identical statistical model that captures all the variety by only having the same set of parameters. Since the investigated dataset is in a time-series format, the correlation between any of the two years' data (predictors) in the same instance should be considered and increase the complexity when applying a statistical model. Moreover, due to the natural logic, the future population depends on the previous population, indicating that the algorithm which uses its reasoning of previous data to inform the future is considered.

With the above facts, a machine learning model which captures the correlation between all predictors (age-sex population in each year) and considers the natural sequential relationship between predictors (previous year) and response (forecasted year) is chosen. Besides, applying extra variables (e.g. population group trend) to cluster the instance before fitting the model could improve the performance. The machine learning method based on the Recurrent Neural Network (RNN) will be chosen since it considers the time series logic between predictors and captures the changing trend of the data (Rumelhart

et al., 1986). However, as mentioned in the related work part, a standard RNN cannot handle the long-term dependency once the time series is expanded due to the gradient vanishing or exploding. LSTM, which could solve the gradient issue, is finally picked in this case. The LSTM could also potentially outperform in this task since it applies different 'gates' in each neuron to control the previous state's information and avoids the issue of recent input occupying more weight in prediction and squeezing the contribution from an earlier input. It is also shown in Makridakis's paper (2018) that the machine learning model (which is used here as LSTM) could catch the potential trend that is invisible in the dataset.

5.2 Long-Short Term Memory

5.2.1 Theoretical Description of LSTM

LSTM solves vanishing gradients by adding or removing information in cell states and using three gates to control that. Forget Gate controls how much information to forget from the last memory cell C_{t-1} . Using a sigmoid layer that outputs a value between 0 and 1, 1 means completely keep it, and 0 means completely forget it. Then, LSTM will decide what new information to add to the cell states. A sigmoid layer called 'input gate' controls what values to update, and a tanh layer creates new candidate values. Combine the forget and input gates to update the old cell state to the new cell state C_t . Finally, the output gate controls the output based on the current cell state. First, run a sigmoid layer decides which parts of the cell state to output. Then a tanh layer will push the output to -1 and 1 and only output the parts we need for the next state h_t .

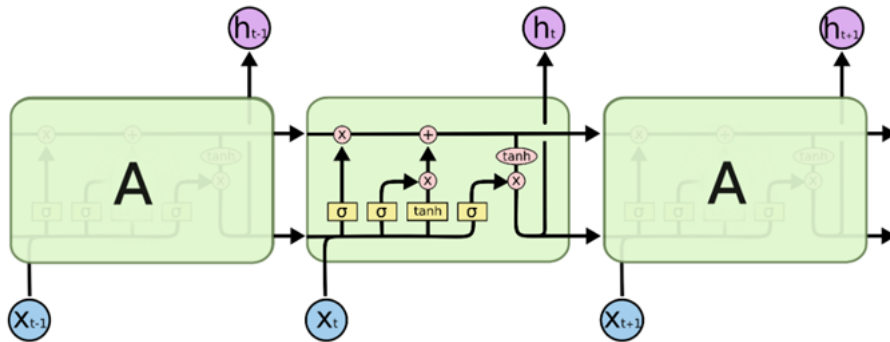


FIGURE 5.1: Repeating module in an LSTM with four interacting layers (Colah, 2015).

5.2.2 Theoretical Description of Sliding Window

In this project, we put time-series data into the LSTM model to perform supervised learning. Adding a sliding window can help us control how much previous time step(s) will be used for the next time step's prediction (Tomar et al., 2022).

For example, Table 5.1 explains how sliding window with window width equal to three works in training set. Data from 1991 to 1993 will be the multivariate input for training and 1994 is the output for this window, next window starts from 1992-1994, etc.

X (input)	Y (output)
1991, 1992, 1993	1994
1992, 1993, 1994	1995
...	...
1998, 1999, 2000	2001

TABLE 5.1: Window width = 3 for training set in date between 1991 and 2001 data.

While figure 5.2 below shows a general logic of sliding window for reference:

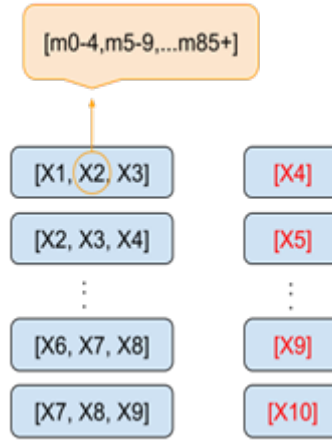


FIGURE 5.2: Sliding window with window width = 3 for male's training data.

Different window widths are applied when implementing LSTM. Since the time series is only eleven years, to have more training data, we mainly use the window width to be one or three and compare the results between different types of LSTM models.

5.2.3 Application Description of LSTM Model (Type 1, 2, 3)

We build four LSTM models in total. Each time, we try to introduce different approaches to the model. If it does not improve the performance and makes the model more complex, we will not adopt it.

Type 1: Basic LSTM with Multivariate Input and Rolling Updates Logic

During preprocessing, we separate the data set into eleven years of training data (1991 - 2001) and ten years of testing data (2002 - 2011). Data is further divided by gender, and we build two LSTM models for each and predict them separately.

For LSTM model building, we use the LSTM packages from Keras. As a neural network model, we need to add LSTM layers following our request:

1. Initialise a sequential model using the inbuilt function '`Keras.models.Sequential()`', and the layers can be defined and passed through the Sequential as arrays.
2. Add a hidden layer that defines the dimension of the hidden states, activation function used and input shape.
3. Add a fully-connected dense layer with the size of 18 to output 18 age groups for each gender.

After completing the model, use the command '`model.compile()`' to run the model; also specify the optimiser, loss function and metric used.

To improve the performance of this model, the Keras-Tuner (O'Malley et al., 2019) is applied for tuning hyperparameters and trying different combinations for LSTM units, activation function, optimisers and loss function. The best model is the one with the Minimum Absolute Percentage Error (MAPE). It ends with using mean square error as the loss function that the optimum prediction will be the mean target value; Rectified Linear Unit (RELU) as the activation function, which is computationally efficient and provides state-of-the-art results. For optimisers, we chose between 'adam' and 'adagrad'. Adaptive Gradient Algorithm (AdaGrad) uses the second moment, and there is no decay to deal with sparse features. Adaptive Moment Estimation (Adam) includes the properties in AdaGrad. It computes the adaptive learning rate by updating weights iteratively according to the training data.

Since we set different window widths, the model will consider more than one year of data when predicting. Thus we generate a step-by-step time series forecasting method called rolling update that will refit the predicted results into the LSTM model for the next prediction. So the model can predict year by year with fixed and full window width, which is the same as the training process.

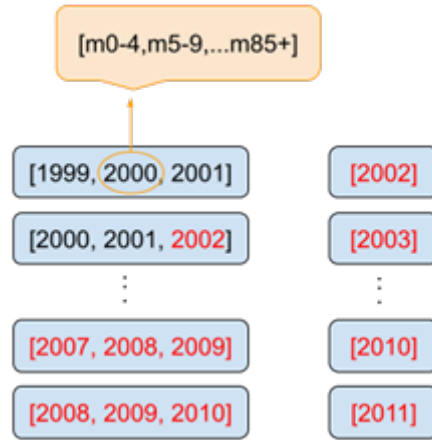


FIGURE 5.3: LSTM rolling update for male testing data with window width = 3. The black word means true data from the training set; the red word means predicted data. Each data contains 18 age groups.

Type 2: Based on Type 1, Add Scaling and Random Splitting

For the type 2 model, we consider scaling the input data to a range of 0 - 1. It is vital to do scaling in the neural network since we will receive smaller input variables with a mean closer to 0, and the model can learn from them faster and easier. In most cases, scaling will improve the results, but it also depends on different situations.

When splitting the train and test set, we use random splitting rather than directly splitting by index, with 20% of the data assigned to the test set. Although data will not be sequential, it prevents the model from overfitting and potentially improves its performance in unseen data sets. The random state will affect the results; to reproduce the 'best' result, we try multiple random states and confirm a fixed one to be 2020.

After we get the prediction results, we notice that some areas receive negative population predictions, which is irrational. It occurs in some small areas with low populations and a decreasing population growth trend. For the final results, we replaced the negative values with 0. However, the previously predicted results will affect the next prediction

round according to rolling predictions. We decided to replace the negative value after each prediction term so that it would not increase the error or lead the trend in the wrong direction.

Type 3: Based on Type 2, Add Early-Stopping and Learning Rate Reduction

A large epoch number may lead to overfitting, but a small epoch may underfit. We set the epochs to be 1000; until now, all the models took hours to compile. In order to save some computational time, early stopping is adopted. Early stopping means stopping the model earlier if the performance stops improving.

We monitor the validation loss, which indicates how we measure the model performance, and we want it to be as small as possible. In this case, the model will stop automatically if the validation loss stops decreasing. However, the first sign of no improvements will not guarantee the best time to stop. The model may be stuck at the local optimum, so the results may stop improving or even worsen before reaching the global optimum. Thus, we set the argument `patience` to be 50 according to different trials to delay the stopping by 50 epochs.

Another important point related to the model performance is the learning rate. The learning rate controls the pace of learning. A small learning rate requires more epochs and takes more time to converge since each update has only a small change. It has the risk of getting stuck in the local optimum. A significant learning rate will converge faster, but it may miss the global optimum result and reach a local optimum.

With the help of `'ReduceLROnPlateau'` in Keras, we can schedule the learning rate by decreasing it when the model performance meets a plateau. We monitor the same metric `'val_loss'`, which is the validation loss and set a patience number to 10. The argument `factor` indicates how we will reduce the learning rate. Based on multiple tests, we decided to use `factor = 0.2`. In summary, if the validation loss stops decreasing for ten epochs, the model will decrease the learning rate by 0.2.

Finally, we introduce external variables `'State'` and `'Average total population'` to the model and apply LSTM to each data group divided by those two features. This logic is similar to separating data by gender and grouping similar data together so the model can learn the pattern better. We first generate the quantile information for the average

total population and store them in lists. Separate data by state first, then use population quantile to separate data into three groups and predict each group separately.

Type Extra: Based on Type 1, 2, 3 Model, Implemented for Special Offer

Compared to previous implementations, the type-extra model trains on 5-yearly data such as 1991, 1996, and 2001. Consequently, this model can only make 5-yearly predictions, such as 2006 and 2011.

For training, the true values are used as though they were known in advance. When doing a rolling forecast, the true values rather than previous predictions are fed into the model to create predictions.

To predict the test data however, the model uses only the true value of the last training data point, in this case 2001, then subsequent predictions are fed back into the model to predict five years after it. This simulates the scenario wherein the 1991-2001 values are known and will be used to forecast 2006 and 2011, which would occur in the future.

The idea behind this implementation was to reduce the amount of error that accumulates when forecasts, which are estimates themselves, are used to predict later values. Since the recorded census data is obtained every five years with in-between values being interpolated, the model likewise only requires 5-yearly predictions.

However, this approach has also introduced potential drawback that might reduce the prediction accuracy. It further reduces the amount of training data, as it effectively leaves only three data points per cohort-area to be divided for training and validation.

5.3 Synthetic Migration Model

According to Wilson (2022), the Synthetic Migration Model, a state-of-the-art algorithm, can be applied and perform well in the small-area age-sex population forecasting. To achieve the project's goal, the comparison between the performance of the implemented LSTM and the benchmark, the Synthetic Migration Model, is re-constructed in this project and evaluated based on the same provided SA3 dataset of Australia.

5.3.1 Theoretical Description

The Synthetic Migration Model is a domain-specific statistical model designed based on the age-sex population in each area, accompanied by the pre-calculated fertility, mortality and migration rate or the turnover rate. By incorporating the pre-calculated rate factors, each region's age-sex population changes are obtained by considering the new births amount, records of death amount in the based period, and the net migration number in each age-sex group. The projection result is then calculated based on the current population and the related changes. Since the model is designed to predict a five-years ahead's population, the data is only prepared at an interval of every five years.

5.3.2 Extra Variables & Models

Unlike the self-constructed LSTM model in this project, the Synthetic Migration Model has introduced new variables and models that assist the prediction and improve performance. By including the fertility, mortality and inward and outward migration rate, the population changing pattern is captured more accurately than the LSTM model, which only relies on the population difference between each investigated two years. Besides, an extra feature, 'National Projection', is applied in the Synthetic Migration Model. It constrains the forecast by ensuring the consistency of the projected total population generated by the Average-of-4-Model's logic provided by Wilson (2015). Besides, the adjustment of inward migration value is applied to help avoid the negative predicted results during the prediction, which fulfil the natural constraint of the population.

5.3.3 Project Application Description (2006, 2011)

Initially, the Synthetic Migration Model is only constructed in VBA, which might be unavailable to some researchers who are unfamiliar with it. In this project, in addition to comparing to the result from the Synthetic Migration Model, we are also required to rebuild its code into the R version, allowing more users to apply it for forecasting the age-sex population. The Synthetic Migration Model is firstly reconstructed into an RMarkdown version which accepts the pre-designed CSV data files. By reading in the age-sex population dataframe, the code generates a pre-construct of the data vector that stores the related factors' value. Then, the age-sex estimated base period's birth and

death rates are computed, and the inward and outward migration data, along with their corresponding scaling factors, are obtained as well. After checking the availability of the generated input data, the model could then be applied for projection. Based on the jump-off year's data (the year set to trigger the projection algorithm), the forecasting is processed among each small area's age-sex group. An iterative logic that obtains the forecast result when the process converges is applied to reduce the prediction uncertainty. A stopping criterion is set and will be triggered once all the age-sex group's predicted population among all areas does not significantly change compared to the last iteration. A rolling update prediction logic is implemented outside the convergence prediction logic, similar to the LSTM Model. The last forecasted result (The year 2006 in this project) is used for projecting the next future result (The year 2011 in this project) iteratively until all the target year's projection values (2006 & 2011's results) are obtained. Once the target years' results are obtained, the error rate is calculated based on the same formula applied for evaluating the LSTM, which will be illustrated in the Evaluation Chapter.

Chapter 6

Evaluation

6.1 Evaluation Matrix

For this project our primary metric would be the Absolute Percentage Error (APE):

$$APE = \frac{|F - A|}{A} 100 \quad (6.1)$$

This metric compares the forecasted age-sex population F with the actual population A . However, after further investigation and discussion with the host, the traditional APE metric was considered less informative since it does not reflect the significant error difference across each age-sex cohort. Since the population of younger groups in the dataset is larger than those in older groups, the error may cancel out when over-prediction and under-prediction occur in the prediction process of different age groups. To provide a more detailed response to the accuracy of our model for age-sex prediction, we used an alternative method from Wilson (2022),

$$APE_{age-sex} = \frac{\sum_s \sum_a |F_{s,a} - A_{s,a}|}{A} 100 \quad (6.2)$$

Based on the formula 6.2, two evaluation standards are applied. The difference between the forecasted population $F_{s,a}$ and the actual population $A_{s,a}$ among each small area is firstly calculated. Then, summation of all age-sex cohort's APE in each area is done to obtain $APE_{age-sex}$ for reference. In this project, the mean, median, and the 90_{th} percentile values of APE are computed for evaluation and recorded in the below tables.

6.2 Evaluation Result

The APE for each model measured in the age–sex level are reported in Table 6.1:

		Synthetic Migration Model	LSTM Type 1	LSTM Type 2	LSTM Type 3	LSTM Type Extra
5 years (2006)	mean	7.05	16.17	13.40	13.75	11.89
	median	4.77	12.47	9.45	9.61	8.63
	90th percentile	15.42	33.16	26.65	27.31	23.60
10 years (2011)	mean	11.49	25.62	22.39	24.56	18.42
	median	8.13	19.90	16.24	16.30	14.30
	90th percentile	25.47	51.84	44.41	44.65	37.00

TABLE 6.1: Forecast Errors of Population in Age-Sex Level

The APE for each model measured in the total population are reported in Table 6.2:

		Synthetic Migration Model	LSTM Type 1	LSTM Type 2	LSTM Type 3	LSTM Type Extra
5 years (2006)	mean	6.60	14.89	12.20	12.52	10.63
	median	5.26	13.05	9.79	9.95	9.07
	90th percentile	11.07	21.89	19.56	19.47	15.43
10 years (2011)	mean	11.43	24.49	21.22	23.36	17.50
	median	9.41	20.52	16.59	16.70	15.06
	90th percentile	18.89	39.72	33.83	33.35	24.98

TABLE 6.2: Forecast Errors of Total Population

6.3 Evaluation Measurement

The model’s evaluation standard is defined along with the APE results obtained in the above two tables. A model with a greater value of APE among each measurement statistic (e.g. mean value) performs worse than one with less APE. Besides, specific measurement statistics would be preferred for evaluation based on investigated data characteristics, and the selection of evaluation statistics will be illustrated in next section.

6.4 Result Interpretation

The three metrics recorded are the mean, median and 90th percentile of the APE as mentioned in Wilson (2022). The use of APE as a metric allows areas with small populations to contribute to the error as much as larger areas. Compared to the mean, the median can provide a better overall estimate of the model performance if the errors are skewed. Furthermore, the 90th percentile is at least as important because aggregated metrics may cause accurate predictions to conceal significant errors for individual areas; the accuracy for any one area is just as important, considering that the information would be applied for each small area. In this project, since there is data sparsity among specific age groups, the model performance on each age-sex cohort's prediction might include bias, so the median value of APE is preferred for evaluation by reducing bias in measuring the model's performance.

Compared to the benchmark model, the LSTM implementations, at best, get an additive 4% higher error. Type 2 LSTM provides some improvement compared to Type 1 LSTM, while Type 3 LSTM may be less accurate than Type 2 although its difference is negligible. Finally, the Type Extra LSTM provides the best accuracy, albeit slightly more limited. In all cases, however, the LSTM model has higher mean and quantile error rates and takes much longer to create forecasts than the benchmark.

Interestingly, all models show a larger mean error than the median for the total and age-sex errors. This suggests there are areas that contain relatively large errors that pull the mean up.

Chapter 7

Discussion

7.1 Discussion of Different Models

The Type 2 implementation makes some additions, most notably random splitting of the training data for fitting and validation. Although the proportions of data used in fitting and validation are the same, the data used in fitting Type 1 was derived from the earliest years only. This leaves the latest years for validation and creates a gap of several years between the fitted data and the first year of the test set. While the model was meant to forecast as far as 15 years ahead, the gap unnecessarily extends how far ahead the predictions will be made. Hence, with the random splitting used in Type 2, forecasts can be made with more “recent” data, as it is possible to include one of the last few training data points for fitting.

Another addition to Type 2 is the implementation of learning rate reduction and early stopping. Early stopping would not necessarily improve the accuracy; rather, it would only reduce computation time. Learning rate reduction on the other hand allows the model to react to the validation results by adjusting the learning rate. This can slow the convergence of optimization but would prevent it from overshooting the global minimum. The results indicate that improvement from learning rate reduction is minuscule, and this is likely due to limitations on the number of epochs for which the optimizers can run. It should be noted, however, that too low a learning rate can prevent convergence or cause the optimizer to get stuck at local minima.

For Type 3, the only difference is that areas were grouped by quantiles according to the average total population and birth rate from an external dataset. Then the model is fit on each group separately but otherwise counts areas within each group as the same. The results indicate little if any, improvement. Since the areas are being clustered, it is likely that the model overlooks subtle differences between areas in the same group, hence the reduction in accuracy. However, fitting on the grouped data can save computation time.

Lastly, there is the Type Extra model which fits and predicts data 5-years apart. This achieves the best accuracy out of the three models, likely due to the rolling window of the LSTM implementations. With less estimated data being fed to the model, prediction error does not accumulate as quickly, allowing further years to be predicted with close-to-true values as input.

None of them matches the benchmark model in terms of accuracy. There are several factors that may explain this. For one, the LSTM model was fit using fewer predictor variables. Whereas the Synthetic Migration Model estimated predictors that are known to be important such as birth, death, and migration rates, the LSTM model only uses the population as both the response and predictor. Furthermore, the LSTM implementation does not consider the effects of one group on another, such as the obvious case of one age group being correlated with the older age groups in later years due to ageing. There is also the problem with gradients vanishing or exploding for other layers due to the ReLU activation function, hindering optimizers that make use of gradients especially when long time series are involved.

7.2 Limitation of LSTM and Synthetic Migration Model

There are limitations to applying the models previously mentioned. The LSTM model does not use external factors and so as compensation, it would require a large amount of data compared to the benchmark. The current implementation also loops across the different areas and fits on each one separately. As a consequence, the model fails to consider how the different small areas in Australia affect each other such as through migration. Like other neural networks, this model has hidden layers with their own estimated parameters, and that makes it difficult to see trends and correlations between different age-sex groups; at best parameters obtained from fitting the model can only

be used within the model itself. Lastly, using the model on different sets of data would require that said data is formatted similarly to the SA3 training data, having the same 18 age-sex groups for each of males and females and having an ‘SA3 Code’ column for each area. Furthermore, LSTM has heavy computational requirements which may prevent cross-validation or other methods of improving measures of generalization. The Keras package also lacks some metrics which provide penalized measures of loss available to traditional statistical learning methods such as AIC. Instead, these frameworks rely on callbacks or more computationally expensive methods such as cross-validation to mitigate overfitting (Makridakis et al., 2018). This can lead to a greater possibility of overfitting such models.

The Type Extra implementation would need at least six consecutive years of data or two that are five years apart, since the points it trains on are five years apart. Likewise, the forecasts it produces are limited to a 5-year interval. It is also limited to using only one data point to predict the next as opposed to Types 1-3 which can use windows of consecutive years.

On the other hand, the Synthetic Migration Model uses external factors such as fertility and mortality data (Wilson 2022). Initial values are not necessary, but in their place would have to be sufficient data for estimating the factors, and these values would then be derived from the initial populations. The migration rate estimated by the benchmark also does not separate different types such as international migration or local migration between small areas.

For both models, predictions are performed with a rolling update, using the fitted model to predict a fixed time period ahead, then feeding that prediction back into the model to predict twice that period ahead. This introduces some error into predictions due to the use of estimates, as opposed to using just the true values to create a prediction within any timeframe. For example, it may be better to fit data using 1991-2001 to predict ten-year targets 2001-2011 directly, as opposed to using 2000 to predict 2001 then using the estimate for 2002 and so on.

Chapter 8

Recommendation

8.1 Positive

From the above applications and practices, we summarise the advantages and disadvantages of LSTM models based on our experiment result to make the recommendation models for population prediction in small areas.

From a positive perspective, compared with traditional statistical models, applying the LSTM model does not require deep domain knowledge and complex mathematical logic. As mentioned in the related work part, the training and fitting process of the LSTM model is similar to a black box. However, the LSTM model may capture hidden patterns and trends in the data with its in-built logic. Meanwhile, the LSTM model only uses age-sex population data for training and fitting the model. Compared with the Synthetic Migration Model, it does not need the mortality and fertility data. A final, extra variable implementation would be much easier for the LSTM model. When adding additional data features to a statistical model, domain expertise is needed to consider the relationship with the predicted values and build the matrix calculation based on the relationship. From the point of view of the LSTM model, it is only necessary to match the input format and dimension of the extra features with required by the model.

8.2 Negative

On the contrary, applying the LSTM model to our theme also has some drawbacks. One of the most direct and obvious is that the accuracy of the prediction results is not as good as the statistical model. Among all types of LSTM models tried, the best performance still delivers 4% more error rate than the Synthetic Migration Model. In addition, because the process is a black box, there is a significant bottleneck in the optimisation of the LSTM model, and it is almost impossible to explain the reasons for the prediction result. It is challenging to change the model from its operation, except for tuning the initial parameters of it and increasing the number of hidden layers. Computational time is also a serious issue, all types of LSTM models implemented above cost 7 - 8.5 hours of running time. While the Synthetic Migration Model only takes approximately 6 minutes to complete on the same device (MacBook Pro with Apple M1 Pro Chip, and Memory = 16GB). It makes it more difficult to optimise the tuning process and potentially larger data volumes in the future. If the local computer's performance is exceeded, additional costs and learning costs are associated with using the server to model.

8.3 Overall Recommendation

Overall, the LSTM model does not achieve the same accuracy as the Synthetic Migration models and is very difficult to optimise and interpret because the process is black-box. In addition, the LSTM model is very demanding on computer hardware. We recommend using advanced GPU to drive the model when dealing with small area age-sex demographic data, reducing the model running time. As mentioned above, implementing LSTM would be easier for more extra features than statistical models, as it does not require as much expertise. Besides, based on the current result, we recommend continuing to use the Synthetic Migration Model based on our reconstructed R version, which can generate the forecast more accurately and efficiently and can be widely applied by the researcher unfamiliar with the VBA.

Chapter 9

Conclusion

In conclusion, to solve the research question of ‘Can Global Machine Learning Models Improve the Accuracy of Small Area Age-Sex Population Forecast’, the project is designed to solve two main goals by following the data science pipeline. Implementing the three standard and one special offer types of LSTM models help predict the age-sex population in small areas and achieves the goal of discovering the key aspects that can improve the forecast accuracy by self-comparison between the implemented LSTM models. Besides, after comparing the result from the best-performed LSTM (i.e. the Special Offer Type) with the Synthetic Migration model, the LSTM model cannot outperform with fewer variables and less mathematical logic. Based on the result from the two main goals, the research question is answered as follows: the Global Machine Learning Models cannot improve the accuracy of the small area’s age-sex population projection compared to the state-of-the-art statistical model in the demographic domain at the current implementation stage.

For the following limitations, the result of underperforming is obtained. The amount of training data, which is key to every machine learning model, is insufficient in the project. The external information applied to support the LSTM’s fitting in capturing the data pattern is ignored due to the limited predictors. The low interpretability of the neural network increases the difficulties in tuning for the exact hyper-parameters and number of layers that best suit this forecasting. Furthermore, from the computational time perspective, the LSTM takes an average of six hours to compute the result, making repeated implementation and testing harder due to the project’s time limit.

Although the project proceeded in an online style which restricted the communication between the client and us due to the pandemic issue, we could still complete the task following the client's instructions from each meeting without a barrier. Summarising our contribution to the client and the research domain, we not only answer the project goal stated in this report throughout the year but also resolve the task of reconstructing the Synthetic Migration Model. Moreover, the current result from our developed LSTM model is still available for contributing to the client's domain research by testing the feasibility of the machine learning method in the small area's age-sex population projection and providing recommendations to the client as well as the other researcher in this domain at the current stage. Besides, based on our model and research, a potential direction for future study would be pointed out.

In future studies, applying and improving the LSTM model based on the self-implemented package from this project is still recommended once there is sufficiently long time series data for fitting a model, and a powerful computation resource can be accessed for repeatedly testing and tuning the model. Besides, further investigation can be done by considering the extra features used as criteria to separate the data into different groups to improve accuracy, following the logic of the algorithm's concatenation, such as adding a layer after training that enables the model to capture the latent variables.

Bibliography

- [1] Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric reviews*, 29(5-6), 594-621.
- [2] Australian Bureau of Statistics (ABS) (2021). Estimated resident population. Canberra: Australian Bureau of Statistics. <https://explore.data.abs.gov.au/>.
- [3] Box, G. E., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, 70(349), 70-79.
- [4] Dockery, A. M., Harris, M. N., Holyoak, N., & Singh, R. B. (2021). A methodology for projecting sparse populations and its application to remote Indigenous communities. *Journal of Geographical Systems*, 23(1), 37-61.
- [5] Grossman, I., Wilson, T., & Temple, J. (2022). Forecasting small area populations with Long Short-Term Memory Networks.
- [6] Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37-45.
- [7] Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1), 388-427.
- [8] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3), e0194889.

-
- [9] Olah, C., (2015, August 27). Understanding lstm networks.
<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [10] O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., & et al. (2019). Keras Tuner. Retrieved from <https://github.com/keras-team/keras-tuner>.
- [11] Rees, P., Norman, P., & Brown, D. (2004). A framework for progressively improving small area population estimates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(1), 5-36.
- [12] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533-536.
- [13] Tomar, D., Tomar, P., Bhardwaj, A., & Sinha, G. R. (2022). Deep learning neural network prediction system enhanced with best window size in sliding window algorithm for predicting domestic power consumption in a residential building. *Computational Intelligence and Neuroscience*, 2022.
- [14] Wilson, T. (2015). New evaluations of simple models for small area population forecasts. *Population, Space and Place*, 21(4), 335-353.
- [15] Wilson, T. (2022). Preparing local area population forecasts using a bi-regional cohort-component model without the need for local migration data. *Demographic Research*, 46, 919-956.
- [16] Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2022). Methods for small area population forecasts: State-of-the-art and research needs. *Population Research and Policy Review*, 41(3), 865-898.

Appendix A

Appendix

A.1 Code & Meeting Minutes Repository Link

Project's Github Link: <https://github.com/RelaxZhang/Data-Science-Project.git>

A.2 Pseudo Code for Sliding Window

Algorithm 1 Sliding Window

- 1: Initialise the sliding window size $\text{step_size} = a$
- 2: Initialise the data (Predictor) collector window_X
- 3: Initialise the data (Response) collector window_Y
- 4: **for** $i = 1, \dots, (n - a)$ **do**
- 5: extract time-series data with range $[i_{th}, (i + a - 1)_{th}]$ as predictor X
- 6: extract the time-series data with index $(i + a)_{th}$ as response Y
- 7: collect the predictor X with window_X for further usage
- 8: collect the response Y with window_Y for further usage
- 9: **end for**

Notice that this algorithm has converted the whole time-series into a sliding window style data, including the training and the test set, which will be further split for model implementation and evaluation usage.

A.3 Pseudo Code for Rolling Update

Algorithm 2 Rolling Update

- 1: Load in the step size a
- 2: Initialise the first predictor for the rolling update
- 3: Set the number of iteration in rolling update b
- 4: **for** $i = 1, \dots, b$ **do**
- 5: Forecast target year result with the LSTM and the predictor (window vector)
- 6: Put the the forecasted result at the last position of the window vector
- 7: Remove the first element from the window vector (maintain the window length)
- 8: Start the next iteration with the updated predictor (window vector)
- 9: **end for**

Notice that the rolling update prediction result depends on the sliding window size defined for the Sliding Window function. This generic algorithm could automatically generate the result based on varies of requirement for prediction.

A.4 Pseudo Code for the LSTM

Algorithm 3 Type 1 LSTM

- 1: Read in the full age-sex population time-series data
- 2: Split the data into male & female groups
- 3: Initialise data-frame for storing the projection results
- 4: Apply the 'Sliding Window' function to obtain the sliding window style data
- 5: Pre-define the LSTM structure and LSTM tuning for model fitting and projection
- 6: **Start** Fit the Model and Obtain the Projection Result for each Sex Group Data
- 7: Select the best LSTM with the training and validation sets with auto-tunning
- 8: Apply the best LSTM and the '**Rolling Update**' algorithm to forecast results
- 9: Store the forecasting result for evaluation usage
- 10: **End** Finish the Age-Sex Population Forecasting with the best LSTM

Notice that this pseudo code is the basic structure (Type 1) of the implemented age-sex population forecasting model with the LSTM method. Detail of the further implementations or special design (Type 2, 3, & Extra) are included in the report main text.
