COMP90106 Data Science Project Pt1 Group 12 Written Report

Haitong Gao
1068254
haitongg@student.unimelb.edu.au

Yuexin Li
959634
yuexin1@student.unimelb.edu.au

Eric Luanzon
1218692
mluanzon@student.unimelb.edu.au

Meijun Yue
1190161
meijuny@student.unimelb.edu.au

Chi Zhang
1067750
chzhang1@student.unimelb.edu.au

May 27, 2022

Abstract

The accuracy of small-area population forecasts is an important challenge in current society as it provides recommendations and instructions for the community or government to plan future infrastructure facilities in specific areas. The purpose of this capstone project is to develop a machine learning model that helps to provide accurate small area population forecasts. In this report, the introduction of the topic and difficulties from a data science perspective are presented, following a discussion of previous papers and models that provide the potential related solution to the target issue. Then, the data analysis of the provided datasets, including time-series data of the of the total populations for 351 Australian small area populations at the Statistical Area 3 (SA3) level and an additional dataset which provides the populations by age and sex for these small area, are presented in ways of tables and plots. Based on the preprocessed dataset and patterns of population trend observed from each specific small region, preliminary models based on time series forecasting are proposed. At the end of this report, a proposal for this project in Semester 2 is provided, and a timeline table is presented to illustrate the whole working process and key tasks (milestones).

Keywords: Population Forecasting, Time Series, Machine Learning, Proposal, Timeline

Contents

1 Introduction **2**

1.1 Background and Purpose of Small Region Population Forecasting 2

1.2 Difficulties of Machine Learning in this Scenario 2

1.3 Capstone Project 2

2 Related Works **3**

2.1 Literature Review 3

2.2 Long-Short Team Memory (LSTM) 3

2.3 Theta Forecasting Method (Theta) 3

2.4 Exponential Smoothing (ETS) 4

2.5 Forecast Reconciliation 4

3 Data Analysis **5**

3.1 Description of the Raw Data 5

3.2 Data Preprocessing 5

3.3 Exploratory Data Analysis 6

3.3.1 Data Structure 6

3.3.2 Descriptive Statistics 6

3.3.3 Data Sparsity 6

3.3.4 Data / Regions' Correlation 7

4 Potential Models **9**

4.1 Long Short-Term Memory (LSTM) 9

4.2 Theta Forecasting Method 9

4.3 Exponential Smoothing Method (ETS) 10

4.4 Forecasting reconciliation method 11

5 Proposal for Semester 2 **12**

6 Timetable **14**

1 Introduction

1.1 Background and Purpose of Small Region Population Forecasting

Population forecasting in small regions aids policy and decision making for businesses and governments such as the specific point-to-point policy and constructing facilities in each area, indicating its significance in the current society. The demographic data provided by the Demography & Ageing unit at the School of Population and Global Health at the University of Melbourne included the time series of age and sex cohort data points in each region, and the total aggregated population each year.

1.2 Difficulties of Machine Learning in this Scenario

Presented by the previous works, statistical models such as Hamilton-Perry model (Baker et al., 2020) or geographically-weighted regression (GWR) model (Chi & Wang, 2017) were developed for forecasting purposes. However, there is limited research on using machine learning models to produce small area population forecasts by age and sex. Since less provided (true) data can be used for training and testing the model, due to the trade-off between higher accuracy in testing and overfitting, the more complicated machine learning models will have less training data selected from the small dataset. Besides, he population of different small areas will vary depending on many factors, the integral machine learning model might not outperform the simpler model. Data sparsity caused by the natural characteristic (sex & age difference) of a human's lifespan also decreases the performance of a machine learning model. Presented in the age-sex population dataset, the elder shows a relatively small population compared to the younger groups, while having the sparse population in the based time-series data, bias within the captured growth trend would become more significant. With the limitation brought by the dataset, designing a machine learning model for general usage of small regions' population forecasting with higher performance than simpler statistical models is difficult.

1.3 Capstone Project

In this project, potential improvements in the accuracy of population forecasting for total and age-sex cohort scales with additional variables is investigated with consideration for multivariate machine learning forecasting models. A general understanding of population forecasting based on time series data is obtained by first getting access to the previous studies on population forecasting. Then, with further analysis and preprocessing of the provided 2-level hierarchical time series dataset, the preliminary model is designed for implementation based on the given benchmark or the existing models. Moreover, external variables which might affect the population trend are investigated based on the related paper from the demographic aspect.

2 Related Works

2.1 Literature Review

Based on the provided dataset of age-sex cohort hierarchical time series among small regions in Australia from the host, the model with hierarchical forecasting functionality is considered. Presented by Palande and Recasens (2019), the hierarchical forecasting allows time series within each group to be forecasted individually but preserve the relationship within the hierarchy. Since the target forecasting result of age-sex is obtained as a multi-level output, method of forecast reconciliation can be applied for constructing further required target result such as total population in each region (Hyndman & Athanasopoulos, 2021).

Demographers have developed several models that could be used for forecasting the age-sex and total population in small area, with only limited input data. One of them is the Hamilton-Perry model which is easy to implement and used after getting the Cohort Change Ratio. It is the ratio of the population size that often compared with itself that five years ago. Besides, Wilson (2022) introduced the synthetic migration cohort-component model. Compared with traditional migration models that are widely used in forecasting higher geographical levels such as states and national forecasts, the synthetic migration models provide a better accuracy of forecasting small area population size also in different gender and age range levels with the directional migration in local area cohort-component calculation and limited fertility and mortality rate are involved. Until now, Wilson's model is the best performance model on forecasting small area population so it is set as our benchmark model.

2.2 Long-Short Team Memory (LSTM)

Limited research adopt machine learning models, especially neuron network to solve demographic problems. For this project, we need to forecast future population based on the past years data. A Recurrent Neuron Network based model, Long Short Term Memory (LSTM), which is designed to forecast value based on a sequential history data, is considered as our main model. (Hochreiter & Schmidhuber, 1997)

2.3 Theta Forecasting Method (Theta)

The theta method by Assimakopoulos and Nikolopoulos (2000) is a method of forecasting univariate time-series however it has several limitations. The original model used two fixed theta coefficients. An optimised theta method by Fioruci et al., (2015) allowed one coefficient to have different values with appropriate av-

eraging. The model also requires that the time series does not exhibit seasonality. Several methods exist for testing seasonality such as Edwards' type statistic (Edwards, 1961), and for deseasonalizing data (Shisken et al., 1967). Being univariate, this doesn't consider how cohorts within the same SA3 are related.

2.4 Exponential Smoothing (ETS)

The exponential smoothing (ETS) proposed in the late 1950s generates forecasts equal to the averages of the exponential weights for time-series observations. The simple exponential smoothing method was extended by including trend (Holt, 1957) and seasonality (Winters, 1960) components. The ETS is widely implemented in time-series data from industry and other fields with its high efficiency and accurate forecast.

2.5 Forecast Reconciliation

Hierarchical models may be suitable for small area population forecasting, however limited research has investigated their use for this domain. It is possible to forecast not only the age sex groups but also the total population of a higher levels geographic forecasts. In order to make the upper and lower projections uniform, (i.e. the sum of the projections for the different sexes and age groups equals the total local population), forecast reconciliation method is used, so that trends in the total population are not lost which possibly effecting the age-sex population forecasting. The synthetic migration cohort-component model changes the forecasts through the adjustment of inward migration flows only (Wilson 2022) to achieve this. But in hierarchy model, bottom-up approach would be a better choice. The details of the forecast reconciliation method in hierarchy model will be discussed in the section forecasting reconciliation method.

3 Data Analysis

3.1 Description of the Raw Data

The data consists of two excel files: one containing a time series of total populations from 1971 to 2011, and the other has the counts for each age-sex cohort from 1991 to 2011 in small regions of Australia. Since the data was provided in an excel file, the formatting was removed so that the table could be read into python as a DataFrame using the pandas library in a CSV format. Both files include 351 SA3 regions' time series in Australia on which the forecasting is to be done. As one of the Main Structures of the Australian Statistical Geography Standard, SA3s are designed to provide a regional breakdown of Australia. For presenting the population size of SA3 regions in Australia, the mean value of total population as around 68577 people in year 2011 (the latest datapoint from the time-series) is calculated among all regions. The dataset of age-sex cohorts was studied first to see if predictions using just age and sex would be sufficient before adding additional predictor variables as would be done in the total population dataset. Detail analysis of age-sex cohort data is presented in section 3.3.2 of Descriptive Statistic accompany with the table of average population in each age-sex group from the anchoring year 2000.

3.2 Data Preprocessing

The preprocessing step includes applying exclusion criteria to remove SA3 areas with insufficient data; only the regions with at least 1000 people in total population across all cohorts are included in the model. Those excluded from the model would be aggregated into a remainder group which would be treated as its own area with its own time series. Areas that surpass the exclusion criteria are also included if their total population didn't meet the requirement in some years, such as Gungahlin. Additionally, areas with no data for some years but have recorded counts in later years are collected for the remainder. These null values are assigned zero population on the assumption that these cohorts weren't present, 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.

3.3 Exploratory Data Analysis

3.3.1 Data Structure

After the preprocessing step for collecting data into the investigated group, the data type and the descriptive statistics were first studied. A yearly frequency short time series with 21 data points is obtained from each region, indicating that models such as ARIMA, which by default requires more than 50 observations in each time series (Box and Tiao, 1975), are not preferable. Besides, as the 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), models like LSTM (Hochreiter & Schmidhuber, 1997) will be worth further investigation as the preliminary model.

3.3.2 Descriptive Statistics

Since the population trend in years is different across each small region, describing the data range for each sex-age cohort among all years does not provide informative results. Instead, as the population in all regions follows a specific trend, an anchoring year's data (The year 2000) was selected for basic analysis. Due to the unobserved difference between regions' factors not recorded in the age-sex population dataset, the mean value is preferred for illustrating the general data information. The subsample of the age-sex group's mean values in 2000 is presented in Table 1 for better visualisation.

age-sex	m5-9	m15-19	m25-29	m35-39	m45-49	m55-59	m65-69	m75-79	85+
mean	2114.1	2065.4	2199.7	2285.6	2038.2	1497.4	1014.5	671.2	237.0
age-sex	f5-9	f15-19	f25-29	f35-39	f45-49	f55-59	f65-69	f75-79	f85+
mean	2006.5	1978.8	2213.9	2310.3	2059.7	1446.3	1054.5	879.6	535.6

Table 1: Subsample of age-sex groups' average population in the Year 2000

3.3.3 Data Sparsity

Presented in Table 1, a large gap between the elder group (e.g. 85+) and younger group (e.g. 35-39) is observed, indicating the potential issue of data sparsity on the elder group during the further investigation within forecasting models. To further determine this phenomenon, the pyramid graph (Figure 1 & 2) from the sampling regions "Gladstone - Biloela" and "Melbourne City" in 2000 from the dataset are plotted. The lack of base population in elder age-sex groups is apparent, which potentially leads to the consequence of significant bias in forecasting the elder groups' population.

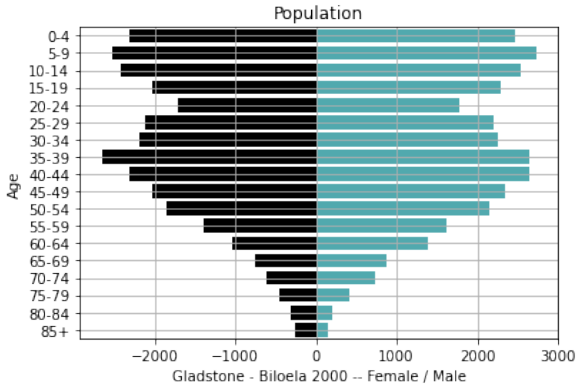

Figure 1: Gladstone - Biloela Population Pyramid

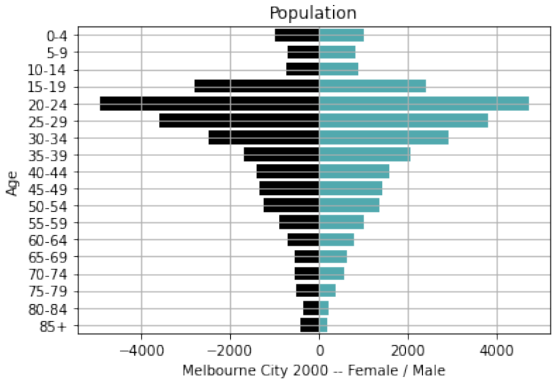

Figure 2: Melbourne City Population Pyramid

The multivariate model with hierarchical forecasting functionality can be applied to reduce the bias in forecasting the elder group's population with higher data sparsity. Since the multivariate model takes in all data (with input matrix of sex_group * age_group * timestamp), sharing data from another age-sex group will contribute to helping capture the trend of time series data with higher data sparsity.

3.3.4 Data / Regions' Correlation

Despite the data sparsity in the elder age-sex group, more information between each region might help improve forecasting accuracy. By plotting the regions' total population trends, the characteristic of the population in each region shows a distinct pattern. From Figure 3 and 4, the regions Latrobe Valley and Barossa present a different population growth trends, indicating the potential difference between the related factors to population exists.

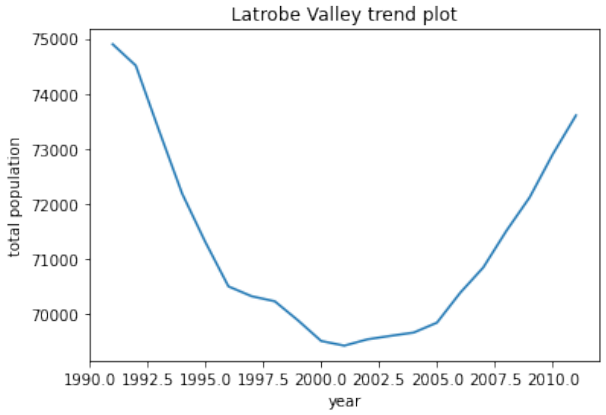

Figure 3: Latrobe Valley Total Population Trend

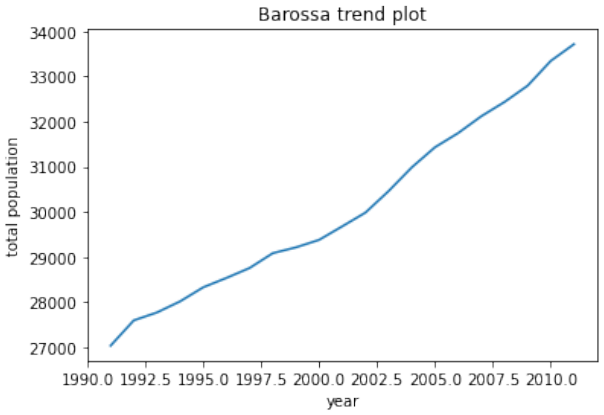

Figure 4: Barossa Total Population Trend

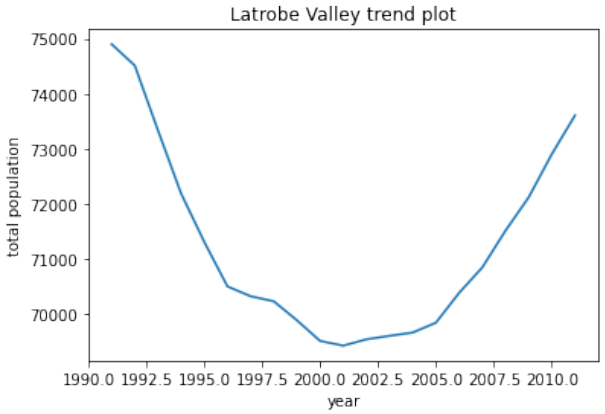

Figure 5: Latrobe Valley Total Population Trend

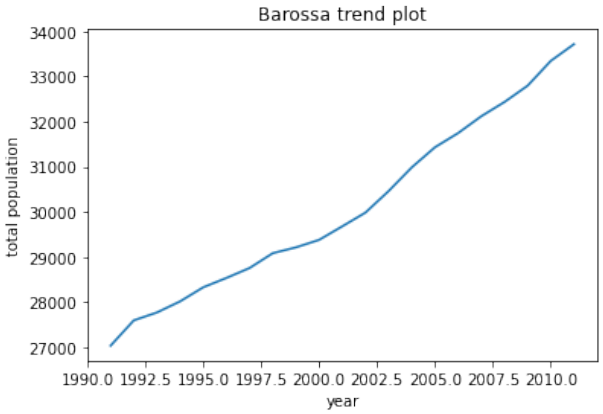

Figure 6: Barossa Total Population Trend

However, as presented in Figure 5 and 6 for the population trends of Wellington and Gawler - Two Wells, the trend's characteristic matches the areas shown in Figure 3 and 4. In this case, further analysis of the correlation between regions' time series, which is used as a grouping standard, can be conducted.

Presented with the above trend plots, it is clear that unobserved / un-quantified factors in each region would significantly affect their population time series trend. A dummy variable that indicates the cluster number of each group of regions with similar trends is applied to help fit the forecasting model.

4 Potential Models

4.1 Long Short-Term Memory (LSTM)

LSTM is a Recurrent Neural Network (RNN) based architecture that is widely used in time series forecasting (Hochreiter Schmidhuber, 1997). Compared with RNN, it mainly solved two problems: capturing long-range dependency by storing them in memory cells and solving vanishing gradients with the help of 'gates' in each neuron.

There are three steps for fitting the LSTM. First, normalise the multivariate input so that it can converge faster with higher accuracy. Then, decide the time step after we split the training and deviation set. Finally, we plan to fit both unidirectional and bidirectional LSTM to compare how the past and future data can influence the prediction.

Previous work from the clients (Grossman et al., 2022) has implement univariate LSTM with SA2s dataset, according to the provided suggestions for future improvements and further research, our motivations for choosing this model are:

- Use SA3s dataset so that we have multivariate input with age-sex break down.
  Each SA3 area has a larger population.
- Expand the time range to 21 years since the NN based model requires more data for training.
- Include new variables other than 'population' into the model if possible.

However, as a "black-box" model, LSTM is not interpretable. Also, it can be hard to implement and time consuming. So we plan to compare this model with the benchmark model and decide a tradeoff, if its performance is obviously superior than it.

4.2 Theta Forecasting Method

In the Theta method two additional time-series would be fit by applying Theta coefficients to the second differences of the original time series, then the two series would be averaged using appropriate weights to give the forecasting line. This coefficient could either remove or magnify deviations from a straight line of best fit so as to take long and short-term variations respectively into consideration.

The model was fitted onto 1991-2006 data to predict 2007-2011 data, excluding areas that fall below the exclusion criteria, and the absolute percentage errors obtained. Since the data is 5-yearly, only the predictions for 2011 are considered for gauging the model's performance.

It is worth noting that the model performs worst on cohorts with very low, typically single-digit values where the prediction may easily be 2-4 times the actual as shown in the table below.

Year	SA3 Name	Cohort	Value	Fit	Residual	APE = 100 * abs(Residual) / Fit
2011	Fyshwick-Pilligo-Hume	m5-9	9	29.04	20.04	226.67
2011	East Arnhem	m80-84	3	9.799	6.799	226.63
2011	Litchfield	m85+	6	12.87	6.87	114.5
2011	Fyshwick-Pilligo-Hume	f5.9	9	20.00	11	122.22
2011	Pilbara	f80-84	26	53.80	27.8	106.92

Table 2: Notable results from Theta Forecasting based on 1991-2006 training data

4.3 Exponential Smoothing Method (ETS)

The exponential smoothing method will be used to forecast higher hierarchical level data. In the ETS, forecasts are weighted averages of past observations. Since the more recent the observation, the higher the associated weight. And as the observations get old, the weight is decaying exponentially.

The simplest of the exponential smoothing methods is called simple exponential smoothing (SES) with only a forecast equation. There is not a clear trend or seasonality in SES forecasting dataset. As the result shows in Data / Regions' Correlation, the observations are grouped by whether there is a trend. Therefore, for the observations with linear trend, Holt's linear trend method and Damped trend method are considered. Holt's linear trend method involves a forecast equation and two smoothing equations. The smoothing equations consist of a level equation and a trend equation, which shows that the function is trending rather than flatten. To capture seasonality, Holt and Winters generated a seasonal method which contains one forecast equation and three smoothing equations. The method has two different seasonal components, additive and multiplicative. When the seasonal variations remain nearly constant, the additive component is better. And when the changes of seasonal variations are proportional to the level of the data, the multiplicative component is better. The forecasts methods above display an infinite or infinitesimal trend in forecast, which tend to be

over-forecast, especially for longer forecasts. Considering this situation, a damp trend is included to generate the Additive damped trend method and Holt-Winters' damped method. The method can be classified by whether there is a trend or seasonality in the time-series data as shown in Table 3.

Trend	Seasonality	Method
N	N	Simple exponential smoothing
A	N	Holt's linear method
A	N	Additive Holt-Winters' method
A	M	Multiplicative Holt-Winters' method
Ad	A	Additive damped trend method
Ad	M	Holt-Winters' damped method

Table 3: Classification of ETS Method

4.4 Forecasting reconciliation method

It forecasts each series at the bottom level, and then summing these to all upper levels in the structure. Different aggregation levels may contain important features of the data to be modelled, and simply forecasting every level independently would lose these important features of the data. A better approach may be to independently forecast all time series at all levels before optimally reconciling these forecasts with a regression model. And then a forecast reconciliation method will be used to help for aggregation. In this project, the structure of the hierarchy model is shown as figure1. It contains the total population of one district, separated by gender to male and female and the bottom level would be the age range data from 0-4, 5-9 continue to 85+. In the hierarchy model, data reconciliation method allows to use different models when forecasting in different levels, which means model that is good for project population totals like Extrapolative and Comparative Methods (Wilson et al., 2021) could be used at higher level, and synthetic migration cohort-component model (Wilson, 2022) might be used at the lower level for forecasting age-sex groups.

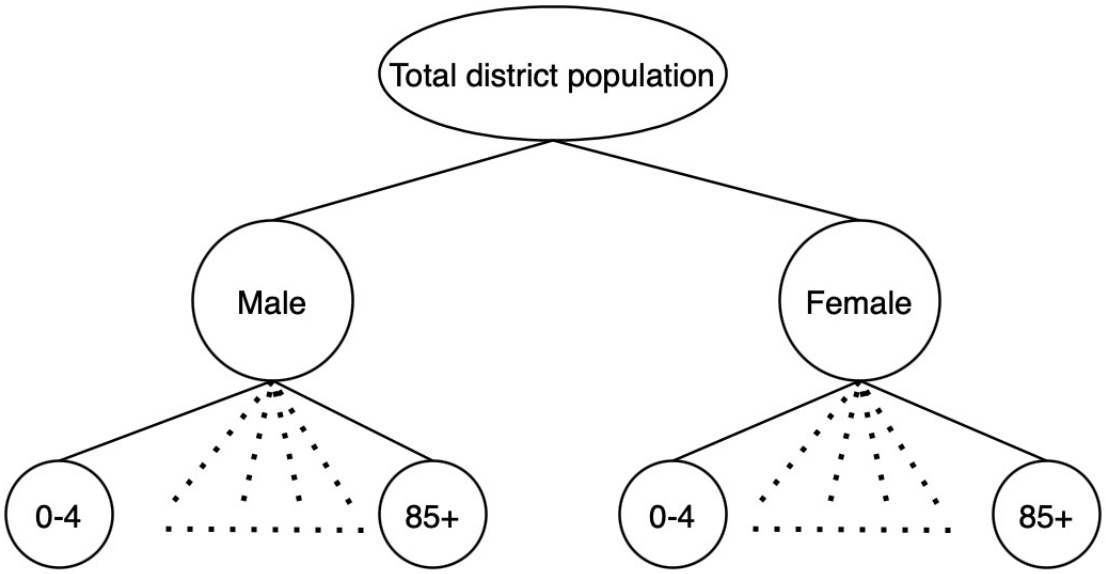

Figure 7: Hierarchical Model Diagram

5 Proposal for Semester 2

By the end of this project, we need to delivery an LSTM model and a report which describe the performance of the small area population forecasts by age and sex. According to the project scope, what we have done for the first stage of this project are:

- Studying the previous work (benchmark model and related papers) from clients.
  Literature review on small area population forecasting and hierarchical time-series forecasting.
- Preprocessing and analysis of the age-sex data .
- Summarise suitable models suggested in related papers.
- Implement ETS and Theta models for total population forecasting to have a look of its performance.

During the winter break, we will try to apply the potential models with part of the data set to prepare for complex model fitting in next semester.

Since what we have done in semester 1 mainly focus on data engineering using the provided data and research to have a preliminary understanding of multivariate machine learning forecasting models. In next

semester, We will complete more on model building and include some potential external variables. There are two main steps:

- Modelling: Working with synthetic migration cohort-component model provided by client as benchmark. Build LSTM for age-sex forecasting to see if there are any improvements. We pan to use 10 years for training, 5 years for validation and the rest 5 years for testing, then compare the results from LSTM and benchmark model using the test set.

- Potential external variables: Investigate some external variables such as fertility rate, population size and density, and remoteness and migration rates. Include them into the training process after we have a stable model.

According to the clients, there is no guarantee that machine learning algorithms will outperform existing models, so our goal is to implement several models (LSTM and other suitable models) with sufficient inputs (age and sex) and possible external data to make reasonable predictions. The evaluation of success focus more on the process of models building and evidence support, three main aspects will be considered:

- Interpretable: The model can be easily interpreted and compared with existing models.

- Rational: The model can output reasonable predictions that correspond to reality.

- Stable: The model can be reused for future studying by clients.

Limited models can be used since there is only a short time series data set as input. The challenge for the team is to implement some advanced deep learning algorithms to solve a demographic problem, which is a new subject to us. So we spend a lot of time reading the related papers online and from the client.

6 Timetable

Task	Start Date	End Date	People Allocation	Additional Notes
Review (End 29/7)				
Feedback for Sem 1	Week 1	Week 2	all	Discuss the feedback given from the clients and supervisor.
New Findings	Week 1	Week 2	Haitong Gao,Chi Zhang	Report any findings during winter break.
Model Development (End 23/9)				
Benchmark Model	Week 1	Week 12	all	Use benchmark model from client.
Model Building	Week 1	Week 4	Eric Luanzon,Yuexin Li, Meijun Yue	Population forecast in AUS.
Model Tuning	Week 4	Week 6	Same as above	
Model Evaluation	Week 6	Week 7	Chi Zhang,Meijun Yue Eric Luanzon,	
Model Interpretation	Week 6	Week 8	all	Result discussion, finalise code and tidy up.
Model Extension (End 7/10)				
External Variables	Week 8	Week 10	Haitong Gao, Yuexin Li	Improve the model and make it more universal.
Report (End 21/10)				
Draft Report	Week 6	Week 11	all	
Final Report	Week 11	Week 12	all	

Figure 8: Semester 2 Timeline

References

[1] Assimakopoulos, V. & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16(4), pp. 521-530.

[2] 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.

[3] Baker, J., Swanson, D., & Tayman, J. (2021). The accuracy of Hamilton–Perry population projections for census tracts in the United States. *Population Research and Policy Review*, 40(6), 1341-1354.

[4] Chi, G., & Wang, D. (2017). Small-area population forecasting: A geographically weighted regression approach. In *The frontiers of applied demography* (pp. 449-471). Springer, Cham.

[5] Edwards, J. H. (1961). The recognition and estimation of cyclic trends. *Annals of human genetics*, 25(1), 83-87.

[6] Fioruci, J. A., Pellegrini, T. R., Louzada, F., & Petropoulos, F. (2015). The optimised theta method. *arXiv preprint arXiv:1503.03529.*

[7] Grossman, I., Wilson, T., & Temple, J. (2022). Forecasting small area populations with Long Short-Term Memory Networks.

[8] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

[9] Hyndman, R.J., & Athanasopoulos, G. (2021) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on <2022, May 6>

[10] Palande, C., & Recasens, J. (2019, October 12). Hierarchical Time Series 101. Medium. https://medium.com/opex-analytics/hierarchical-time-series-101-734a3da15426

[11] Shiskin, J. (1967). *The X-11 variant of the census method II seasonal adjustment program* (No. 15). US Department of Commerce, Bureau of the Census.

[12] Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2021). Methods for small area population forecasts: state-of-the-art and research needs. *Population research and policy review,* 1-34.

[13] 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.