Relax - Introduction :

Hi everyone, we are group 12 working on the topic: ‘do additional variables improve the accuracy of total population forecasts generated by global machine learning’. It is our honour to have Dr Grossman from Demography and Ageing, School of Population Health as our host.

Here is our team, consisting of five students who have taken different tasks in this project. We would start talking about the background and the aim of this project and followed by presenting the data descriptions and the findings to date in this semester.

Eric - Background : 1m14s

So, a quick background. Population forecasts are relevant to organisations such as the government and businesses for planning, marketing, research and so on. There is much research on national level forecasts, however the same could not be said at the small-area level.

Many of the methods used on the former do not work as well since they require large amounts of data. Smaller areas naturally have smaller populations and factors such as internal migration and boundary changes add further variation that are not captured by national level forecasting methods. This is exacerbated when forecasting for specific age-sex groups is needed; certain groups tend to be much smaller, leading to sparse data.

The multitude of forecasts required also increases data requirements. For example, methods that base forecasts on migration would need to get these values for every small-area, as opposed to just the whole country.

While methods have been developed that find a way around these problems, these rely on estimates to compensate for a lack of input data. We hope to find a better solution in the form of machine learning methods.

Move to next slide

Meijun - Data Description :

Before introducing the datasets, we first need to understand the statistical calibre of geographic data.

Here we focus on SA3. As one of the Main Structures of the Australian Statistical Geography Standard, SA3s are designed to provide a regional breakdown of Australia. There are 358 spatial SA3 regions covering the whole of Australia.

Move to next slide

For this project we have two time-series datasets from 1991 to 2011. One contains the total populations in SA3 regions.

Move to next slide

Another divides the total population into male and female, then each into 18 different age-groups for a total of 36 cohorts per region.

Move to next slide

By doing the preprocessing of the raw cohort data, we extracted SA3 regions with the total population greater than 1000 in all 21 years for further analysis while the remaining areas are stored into a remainder dataframe.

Move to next slide

Relax - Data Analysis :

After presenting the background of this project and how the raw data looks like, I will then show the current findings by firstly presenting the data analysis part.

Move to next slide

As potential factors exist in the dataframe, differences between region's time-series are not quantified which makes the maximum, minimum value become less informative. Instead, we aggregate the total population in an anchoring year 2000 of each age-sex group for visulisation analysis purpose.

Move to next slide

Present in this slide, we have a pyramid plot of age and the aggregated population within each cohort in the year 2000 and obtain that a clear data sparsity exists in the elder person especially the 80-84 and 85+ male groups. For reducing the forecasting bias caused by the data sparsity, we research on models that require multivariate input which enables us to capture a more accurate trend with share-data from other input time-series.

Move to next slide

However, since there is 325 valid SA3 regions with time series in each age-sex cohort, it might reach an inaccurate result if processing all regions' forecasting as a whole. In this case, the trend of the population change in each region was studied. Present in the trend plot for Latrobe Valley vs. Barossa, a clear difference is observed.

Move to next slide

However, we also obtained a similar trend plot from the graph of Wellington vs. Gawler Two Wells. We conclude that a strong correlation exists between some regions that could be used as a standard for clustering processing before fitting the forecasting models for improving its performance.

Haitong-related work ：2mins 5s

Demographers have developed several models that only use limited input data. Simple model such as the Hamilton-Perry model is easy and cheap to implement, but it has less details of output data. In the Hamilton-Perry model, it's easier to get the result because it forecasts the population by multiplying the Cohort Change Ratio, which is the ratio of the cohort population size often compared with itself that five years ago. Based on this principle it is difficult to explain some of the changes in the total population growth of the area that appear in some of the assumptions. Considering about our topic to forecast small area time series age-sex population,

Wilson provied a synthetic migration cohort-component model. According to Wilson, the directional migration in local area cohort-component calculation is involved in the model but not the detailed migration data and it came out with a more accurate result. It fits well with our dataset which has no mortality rate, birth rate, or net migration rate. Wilson's model right now is the best performance model, so we set it as our benchmark model. However Wilson said in order to achieve consistency with the total population, it changes the forecasts through the adjustment of inward migration flows.
Move to next slide

So for our decision to create a coherent result, a forecast with hierarchical functionality is considered. The model starts from age-range population at the bottom level, different gender population in the middle and aggregates to the total population in the district on top level. Different levels in the hierarchy may contain important features, so if we simply aggregate the bottom level by summing them to obtain the total population in the district, they are actually not a coherent result. Wickramasuriya introduced the Minimum Trace approach, which is the optimal reconciliation method to get the minimum variance of the coherent result. I will let my teammate introduce the model we have tried so far.
Move to next slide

Meijun - potential methods : 20s

To forecast higher hierarchical level data, we have tried two models, Theta and ETS.

In Theta, two alternate time-series are produced from the original by applying a theta coefficient to deseasonalized data. Each line is forecasted, then averaged to get the final forecast.

ETS includes 6 methods as shown. And forecasts are weighted averages of past observations.
(Switch to next slide)

Yuexin - LSTM : 1min 30s

The objective model for this project is LSTM which is used to make age&sex prediction. It is a Recurrent Neural Network (RNN) based architecture that is widely used in time series forecasting. Comparing with RNN, it mainly solved two problems. Use memory cell to capture long-range dependency. And use 'gates' in each neuron to solve vanishing gradients

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 splitting 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 has implement univariate LSTM with SA2s dataset, to further improve that, our motivations for choosing this model are: To use SA3s dataset as input that each area has larger population. Expand the input data by setting time range to 21 years since

NN bases model need lots of data to perform better and for each time series, we have 36 data. Also, use multivariate dataset with age-sex break down as input. And to include new variables other than ‘population’ if possible.
(Switch to next slide)

Eric - Evaluation:

To evaluate our models, our primary metric would be a modified absolute percentage error that sums the absolute values of the per-cohort percentage errors.

Then our models would be compared in a manner as used by the benchmark synthetic migration model.

(Switch to next slide)

Yuexin - Proposal : 45s
During this semester, we mainly focus on background research, data engineering and have some preliminary understanding of potential models. Next semester, our key tasks are building the model and interpreting its performance. Since we only have a short time series data, there is a limitation in selecting models. According to clients, the evaluation of success is to implement LSTM for this age-sex dataset. The key challenge for the team is we are going to work with some new algorithms and in the demography area, which is a new subject to us. So we spend a lot of time reading related papers.

(Switch to next slide)
For semester 2, we will follow this timeline, which includes review, model development, model extension and report sections.
(Switch to next slide)

That is the end of this presentation, the appendix is the detailed introduction of ETS and Theta, thank you for your time!