THE UNIVERSITY OF
MELBOURNE

Faculty of
Engineering and
Information Technology

Group 12

Does Global Machine Learning Model Improve the Accuracy of Age-Sex Cohort’s Population Forecasts in Small Area

Industry Partner: Dr Irina Grossman

Project Supervisor: Prof Michael Kirley

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

Presentation Outline

THE UNIVERSITY OF MELBOURNE

- Project Team Introduction
- Industry Client Introduction and Requirements’ Illustration
- Challenges of the Data Science Project
- Literature Review
- Data Science Pipeline
- Conclusion & Recommendation

Team Introduction

THE UNIVERSITY OF MELBOURNE

Chi Zhang

Work Coordinating

Data Analysis

Data Reconstruction

Benchmark Model

LSTM

Model's Evaluation

Report Writing

Eric Luanzon

Data Preprocessing

Potential Model

Model Testing

LSTM

Parameters' Tuning

Report Writing

Haitong Gao

Forecast Reconciliation

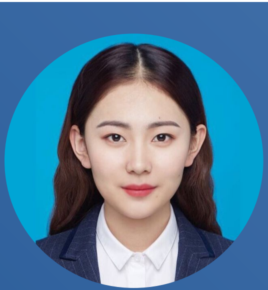
Reference Collecting

LSTM

Model's Improvement

Multivariate Implement

Report Writing



Meijun Yue

Data Visualisation

Data Reconstruction

Model Testing

Data Splitting

LSTM

Report Writing

Yuexin Li

Meeting Agenda Arranging

Data Visualisation

Data Reconstruction

LSTM

Multivariate Implement

Model's Improvement

Report Writing

Project Background

Population Forecasting: Planning, Marketing, Research etc.

Current Outstanding Model: Synthetic Migration Model

Global Machine Learning Model: Long-Short Term Memory (LSTM)

Target: Construct a LSTM Model on Forecasting the Age-Sex Cohorts’ Population in Small Area

Other Requirement: Comparison between LSTM and Synthetic Migration Model

Challenges

- 1. Data Sparsity
- 2. Short Time-Series
- 3. Less Feature Input for the LSTM Model
- 4. Lower Interpretability of Model
- 5. Computational Consumption
- 6. Error Stack Issue
- 7. Input Structure

Literature Review

- Hamilton-Perry Model

  Could be implemented without migration data, easy to implement. But less detail output

- Synthetic Migration Model

  Age-Sex Cohort population forecasting with birth, death, migration rate and total population constraints

- Long-Short Term Memory (LSTM)

  Long Term Dependencies

Synthetic Migration Model

- Constraint the forecast with ‘National Projection’ data
- Change the inward migration flows to maintain consistency
- Apply extra 4 models to create projection total population data
- Migration, Birth, and Death rates are considered
- Area’s independence

Synthetic Migration Model

THE UNIVERSITY OF MELBOURNE

- Data Source: SA3 Age-Sex Cohorts' Data
- Investigate Area: 325 Area + 1 Aggregated Remainder
- Difference from LSTM: More Features / Variables for Forecasting
- Forecast Result: 2006, 2011 Age-Sex Cohort's Population

Data Science Pipeline

- Data Collection
- Data Preparation & Description & Analysis
- Data Modelling and Validation
- Model Deployment on New Data
- Comparison & Reviewing

Data Collection & Description

- All Data are Preliminarily Cooked
- Data Scale — Statistical Area Level 3
- Data Format — Age-Sex Cohort’s Population
- Time Series Data
- Above_1000 Area
- Below_1000 Area

Data Description

Age Cohorts

0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+

Each Region’s Time-series in one year: 2 * 18 (Sex Cohort * Age Cohort)

SA3 Code	SA3 Name	m0-4	m5-9	m10-14	m15-19	m20-24	m25-29	...	m60-64	m65-69	m70-74	m75-79	m80-84	m85+
10101	Goulbum Yass	2603	2565	2517	2472	2178	2392	...	1513	1170	765	506	260	159
10102	Queanbeyan	1593	1362	1223	1406	1743	1803	...	617	478	330	198	95	43
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Partial Dataframe (Male Group in 1991)

Data Description

1991
-
2011

Year	SA3 Name	Total
1991	Goulburn - Yass	61667
1991	Queanbeyan	35281
…	…	…
1992	Goulburn - Yass	61751
1992	Queanbeyan	36409
…	…	…
2011	Goulburn - Yass	69775
2011	Queanbeyan	56051
…	…	…

Above_1000 & Below_1000 Area — Preliminary Preprocessing

Above_1000	
SA3 Code	SA3 Name
10101	Goulburn - Yass
10102	Queanbeyan
10103	Snowy Mountains
10104	South Coast
10201	Gosford
10202	Wyong
10301	Bathurst
10302	Lachlan Valley
10303	Lithgow - Mudgee
10304	Orange
…	…

Below_1000	
SA3 Code	SA3 Name
10702	Illawarra Catchment Reserve
10803	Lord Howe Island
12402	Blue Mountains - South
19797	Migratory - Offshore - Shipping (NSW)
19999	Special Purpose Codes SA3 (NSW)
29797	Migratory - Offshore - Shipping (Vic.)
29999	Special Purpose Codes SA3 (Vic.)
39797	Migratory - Offshore - Shipping (Qld)
39999	Special Purpose Codes SA3 (Qld)
49797	Migratory - Offshore - Shipping (SA)
…	…

(Aggregate to Remainder Area)

Data Analysis – Descriptive Statistic

Time-Series Data (21 Years)

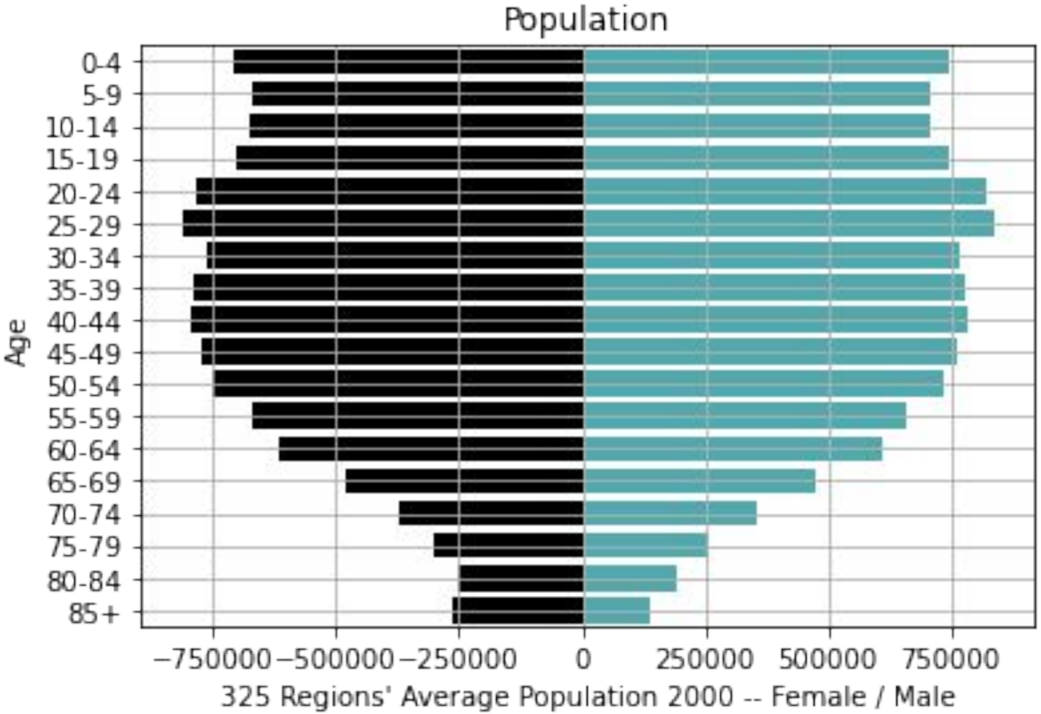
Maximum & Minimum Value of Total Population — Min = 0 ; Max = 190621

Population Distribution among each Age-Sex Cohort

Population Growth Trend in Each Area

Data Analysis – Data Sparsity

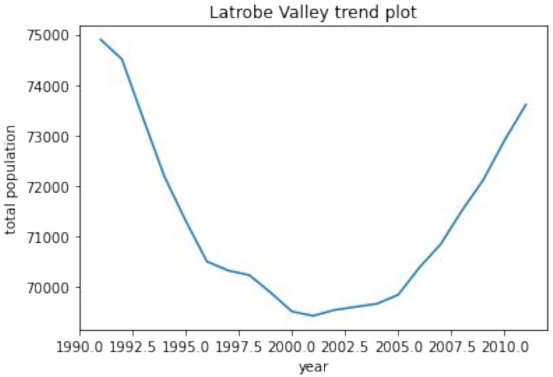
Elder Age-Sex Cohorts Population's Lacking (Visualisation of Anchor Year 2000)

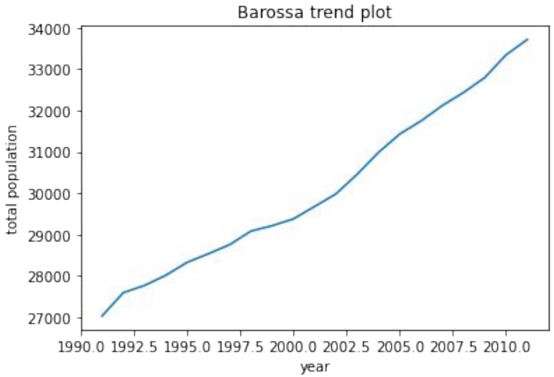


Data Analysis – Population Trends & Clustering

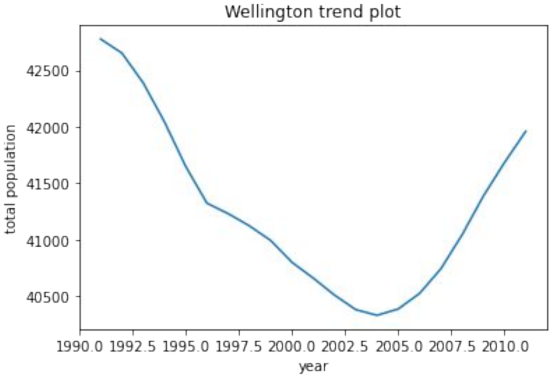
Data Trends – Region's Population Growth Trends' Difference / Similarity

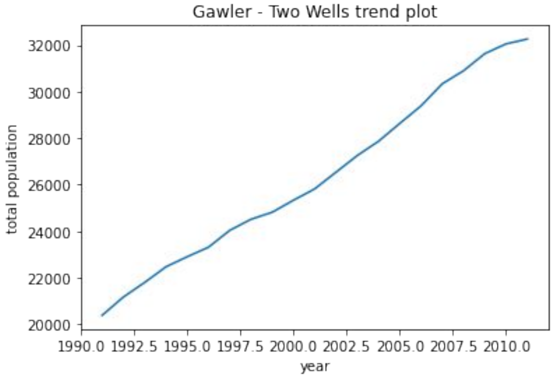
Latrobe Valley vs. Barossa





Wellington vs. Gawler Two Wells

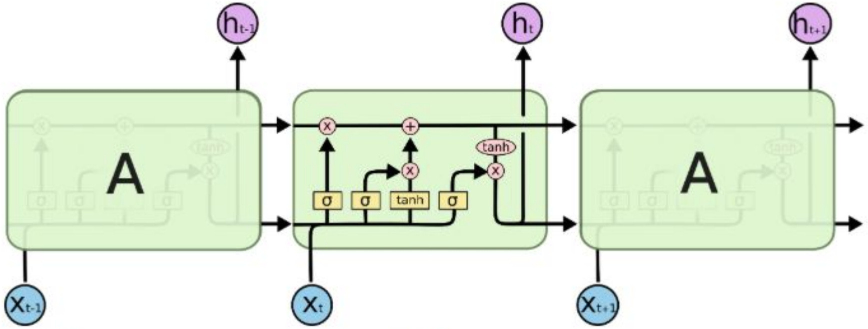




Characteristic of Data

- Short Time-Series for Training
- Sparse — Less Population in Elder Age Cohorts
- Different Trends — Hard to fit the Model with the Same Parameters
- Dependency among Age or Sex Cohorts

Long-Short Term Memory (LSTM)



The repeating module in an LSTM contains four interacting layers.

★ **LSTM Basic Structure**

- Input Gate
- Update Memory Cell
- Forget Gate
- Output Gate

★ **Age & Sex Prediction**

- Scaled input data
- Decide time step
- Unidirectional
- Multivariate Input

Python Package(s)

- TensorFlow — LSTM Package — Model's Framework
- Tuning / Learning Rate & Auto-Stop — Model’s Hyper-parameter(s)
- Random Seed — Model’s Reproducibility

Training & Validation & Test in LSTM

- Training Set — 1991 - 2001 (Varies Splitting of Training & Validation Set)

  (i) LSTM Model Basic (Type 1)

  (ii) LSTM Model Implemented (Type 2 & 3)

  (iii) LSTM Model Extra Implemented (Type Extra)

- Test Set — 2002 - 2011

LSTM Model Basic (Type 1)

- Sliding Window: 1, 3 Window Gap
- Multivariate Input
- Validation Set: 1999 - 2001
- Fitting & Forecasting

  Rolling Update with Fixed Length Training Set & Fitted Model
- Computational Consumption

LSTM Model Implementation (Type 2 & 3)

- Extra Implementation

  (i) Scaling

  (ii) Non-Negative

  (iii) Random-Splitting

  (iv) Learning Rate / Auto-Stop

  (v) Extra Features

LSTM Model Extra (Type Extra)

- Sliding Window with Step = 1, Gap = 5
- Inherit Implementation from the Standard Model (Type 2)
- Predict the Population in 2006 & 2011 (Special Offer)
- Reduce Computational Consumption (Big O)
- Increase Prediction Accuracy

Evaluation (Error Measures)

- Absolute Percentage Error (APE) among each Age-Sex Cohort

$$APE_{age-sex} = \sum_{s} \sum_{a} (F_{s,a} - A_{s,a})/A * 100\,\%$$

- $F_{s,a}$ = Forecast Population of the Age-Sex Cohort

  $A_{s,a}$ = True Population of the Age-Sex Cohort

  A = Total of True Population of the Selected Area

Result Table (Basic vs. Benchmark)

LSTM Type 1 (Step = 3)		
	Age-Sex Level	**Total Level**
mean_2006	16.1709	14.8950
median_2006	12.4672	13.0478
percentile_90_2006	33.1622	21.8927
mean_2011	25.6209	24.4915
median_2011	19.9009	20.5182
percentile_90_2011	51.8375	39.7218

Synthetic Migration Model		
	Age-Sex Level	**Total Level**
mean_2006	7.0493	6.5987
median_2006	4.7671	5.2553
percentile_90_2006	15.4181	11.0650
mean_2011	11.4899	11.4337
median_2011	8.1259	9.4145
percentile_90_2011	25.4696	18.8916

Higher Error Rate than the Synthetic Migration Model (Benchmark)

Result Table (Basic vs. Implementation)

LSTM Type 1 (Step = 3)		
	Age-Sex Level	**Total Level**
mean_2006	16.1709	14.8950
median_2006	12.4672	13.0478
percentile_90_2006	33.1622	21.8927
mean_2011	25.6209	24.4915
median_2011	19.9009	20.5182
percentile_90_2011	51.8375	39.7218

LSTM Type 2 (Step = 3)		
	Age-Sex Level	**Total Level**
mean_2006	13.3995	12.2033
median_2006	9.4548	9.7856
percentile_90_2006	26.6544	19.5631
mean_2011	22.3879	21.2186
median_2011	16.2436	16.5910
percentile_90_2011	44.4078	33.8293

Decrease Around 3% of Error Rate from the Median Perspective

Result Table (Basic vs. Extra + Implementation)

LSTM Type 1 (Step = 3)		
	Age-Sex Level	**Total Level**
mean_2006	16.1709	14.8950
median_2006	12.4672	13.0478
percentile_90_2006	33.1622	21.8927
mean_2011	25.6209	24.4915
median_2011	19.9009	20.5182
percentile_90_2011	51.8375	39.7218

LSTM Type Extra (Step = 1, Gap = 5)		
	Age-Sex Level	**Total Level**
mean_2006	11.8885	10.6295
median_2006	8.6265	9.0743
percentile_90_2006	23.5991	15.4323
mean_2011	18.4228	17.4967
median_2011	14.2963	15.0642
percentile_90_2011	36.9972	24.9798

Decrease Around 4% of Error Rate from the Median Perspective

Negligible Difference between the Unscaled and the Scaled Version of the Extra Model

Recommendation

- Not Recommend

(i) Performance — Does not Outperform than the Benchmark

(ii) Data Characteristic — Too Short Time-Series

(iii) Interpretation — Black-Box Model

(iv) Computational Consumption — Greater Consumption

Recommendation

- Recommend

  (i) External Variable Improvement

  (ii) Easy Application without Complicate Coding

- Overall Recommendation

Conclusion & Report

- Conclusion of Work

  (i) Data Preprocessing + Reconstruct Benchmark Model in R

  (ii) Three + Extra Types of LSTM Model Implementation

  (iii) Result and Recommendation

- Introduce of the Report

  Illustration of Related Work, Model Interpretation, Result & Discussion

THE UNIVERSITY OF MELBOURNE

Faculty of Engineering and Information Technology

Thank you!

Reference

Grossman, I., Wilson, T., & Temple, J. (2022, February 8). Forecasting small area populations with Long Short-Term Memory Networks. https://doi.org/10.31235/osf.io/3k79d

Olah, C. (2015, August 27). Understanding LSTM Networks -- colah's blog. Github.io. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Wilson, T. (2015). New evaluations of simple models for small area population forecasts. Population, Space and Place, 21(4), 335-353.

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.

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.