THE UNIVERSITY OF MELBOURNE

Faculty of Engineering and Information Technology

Group 12

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

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
- 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

Parameters' Tuning

LSTM

Report Writing

Haitong Gao

Forecast Reconciliation

Reference Collecting

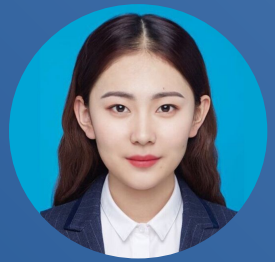
Model Testing

Model's Improvement

LSTM

Multivariate Implement

Report Writing



Meijun Yue

Data Visualisation

Data Reconstruction

Data Splitting

Model Testing

LSTM

Report Writing

Yuexin Li

Meeting Agenda Arranging

Data Visualisation

Data Reconstruction

Model's Improvement

LSTM

Multivariate Implement

Report Writing

Project Background

THE UNIVERSITY OF MELBOURNE

Small Area Age-Sex Population Forecasting: Planning, Marketing, Research etc.

Current State-of-the-art Model: Synthetic Migration Model

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

Target: Forecasts Populations by Age-Sex for Small Areas with LSTM

Other Requirement: Comparison between LSTM and Synthetic Migration Model

Challenges

THE UNIVERSITY OF MELBOURNE

- 1. Data Sparsity
- 2. Short Time-Series
- 3. Less Feature Input for the LSTM Model
- 4. Lower Interpretability of Model
- 5. Computational Time
- 6. Error Increases with Rolling Update

Literature Review

- Synthetic Migration Model (Wilson, 2022)

  Age-Sex Cohort population forecasting with mortality and fertility and calculated migration based on those two

- Long-Short Term Memory (Olah, 2015)

  Long Term Dependencies

- Machine Learning Model vs. Statistical Methods (Makridakis et al., 2018)

  Computational & Data volume requirements

Data Science Pipeline

THE UNIVERSITY OF MELBOURNE

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

Data

- Data Collection – Australian Bureau of Statistics
- Data Scale — Statistical Area Level 3
- Above1000 (Total Population > 1000) & Below1000 Areas

Data Description – Age-Sex Population

351 Areas & 18 Age Cohorts & 2 Sexs

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)

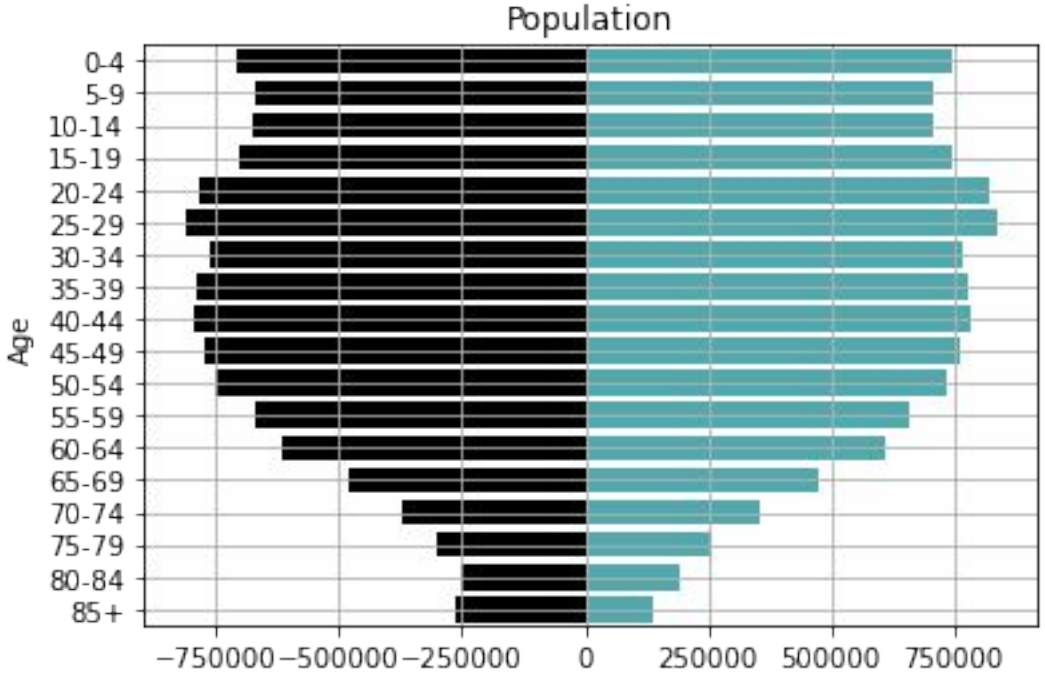
SA3 Code	SA3 Name	m0-4	m5-9	m10-14	m15-19	m20-24	m25-29	...	m60-64	m65-69	m70-74	m75-79	m80-84	m85+
10101	Goulbum Yass	2603	2565	2517	2472	2178	2392	...	1513	1170	765	506	260	159
10102	Queanbeyan	1593	1362	1223	1406	1743	1803	...	617	478	330	198	95	43
...	...	...	...	...	...	...	...	...	...	...	...	...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

Partial Dataframe (Male Group in 1991)

Data Analysis – Data Sparsity

THE UNIVERSITY OF MELBOURNE

National Level Population of Australia (Year 2000)



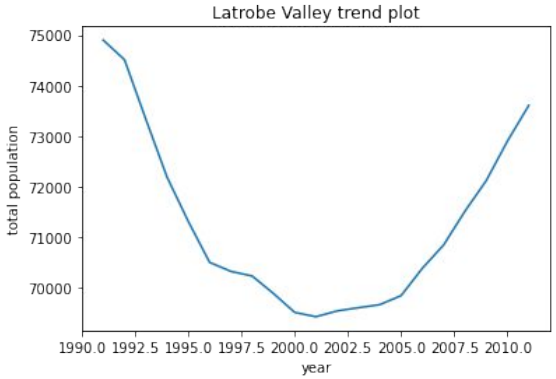
Data Description – Time Series

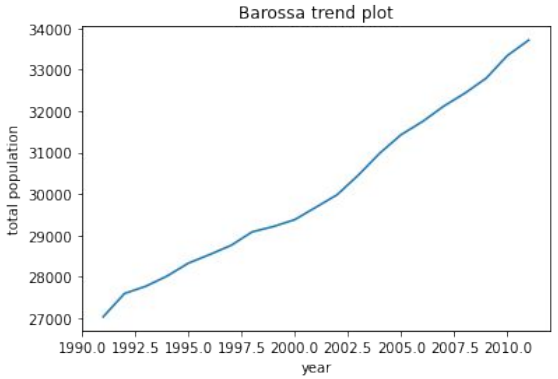
1991
-
2011

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
…	…	…

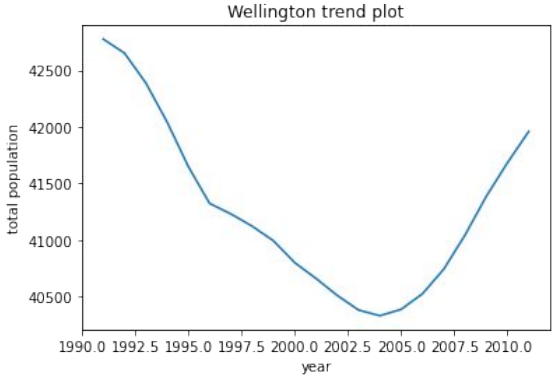
Data Analysis – Population Trends & Clustering

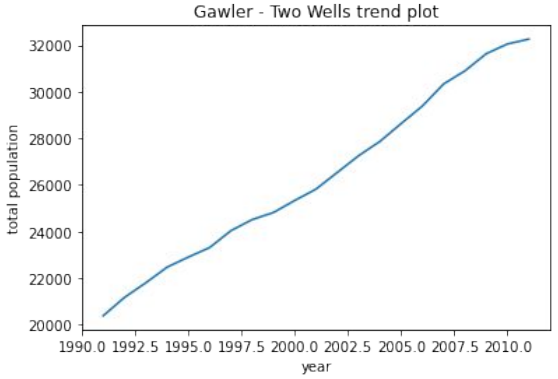
Latrobe Valley vs. Barossa





Wellington vs. Gawler Two Wells





Data Characteristics

THE UNIVERSITY OF MELBOURNE

Short Time-Series for Training (1991-2001 Training & 2002-2011 Testing)

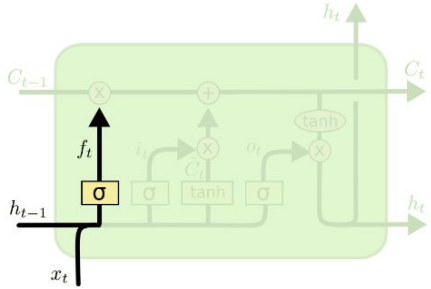
Sparse — Less Population in Elder Age Cohorts

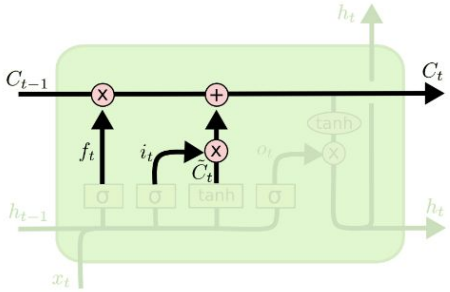
Different Trends — Hard to fit the Model with the Same Parameters

Dependency between Age and Sex Groups

Long-Short Term Memory (LSTM)

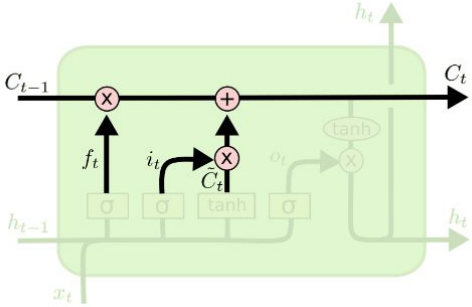
Forget Gate

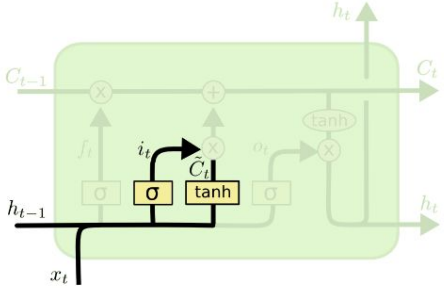




Input Gate

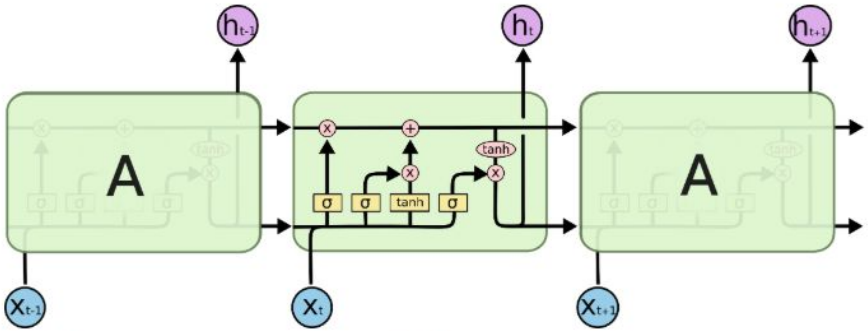
Update Memory Cell





Output Gate

Long-Short Term Memory (LSTM)



The repeating module in an LSTM contains four interacting layers.

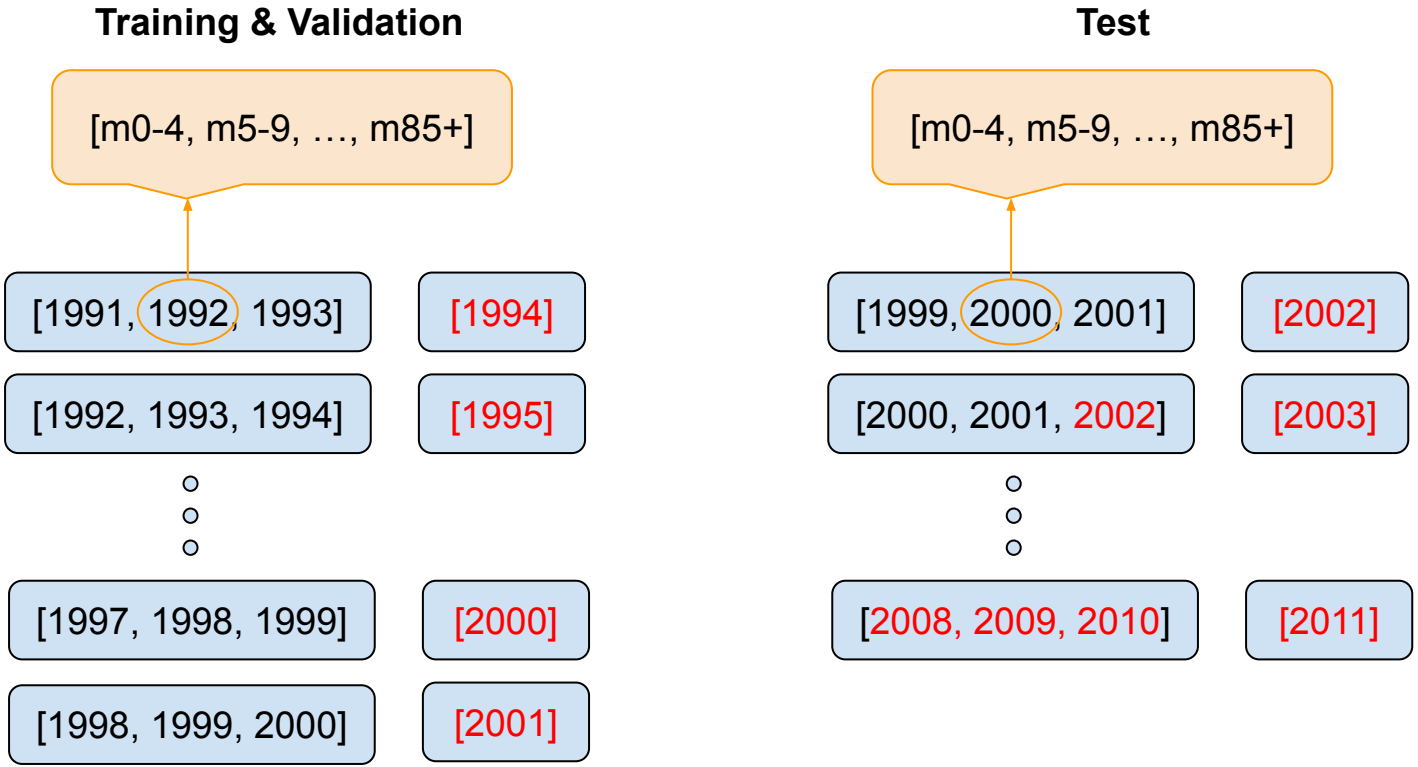
★ **LSTM Basic Structure**

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

★ **Age & Sex Prediction**

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

Sliding Window (Step = 3)



LSTM Model Basic (Type 1)

- Sliding Window & Multivariate Input
- Validation Set: 1999 - 2001
- Fitting & Forecasting

  Rolling Update with Fixed Length Training Set & Fitted Model

LSTM Model Implementation (Type 2 & 3)

THE UNIVERSITY OF MELBOURNE

- Further Implementation

  (i) Scaling

  (ii) Non-negative (change the negative population to zero for every iteration)

  (iii) Random-Splitting train and test sets

  (iv) Learning Rate / Early-Stopping

  (v) Extra Features

LSTM Model Extra (Type Extra)

THE UNIVERSITY OF MELBOURNE

- Sliding Window (Step = 1)
- Inherit Implementation from the Standard Model
- Predict the Population in 2006 & 2011 (Make less error stacking)

Synthetic Migration Model

THE UNIVERSITY OF MELBOURNE

- VBA Version ⟶ R Version
- Requirement: Two Base-year Data
- Difference from LSTM: More Features / Variables for Forecasting
- Forecast Result: 2006, 2011 Age-Sex Cohort's Population

Synthetic Migration Model

THE UNIVERSITY OF MELBOURNE

- Fertility, Mortality data
- Apply extra 4 models to create projected total population
- Constraint the forecast with ‘National Projection’ data
- Change the migration flows to maintain consistency
- Area independence

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 (e.g. Forecast Area 10101 Male 0-4)

  $A_{s,a}$ = True Population of the Age-Sex Cohort (e.g. True Area 10101 Male 0-4)

  A = Total of True Population of the Selected Area (e.g. True Area 10101)

Evaluation (Error Measures)

- Absolute Percentage Error (APE)

$$APE = |F - A|/A * 100\,\%$$

- F = Forecast Population of the Selected Area

  A = True Population of the Selected Area

Result Table (Age-Sex Level)

Result Table (Age-Sex Level)					
Forecast Horizon	Statistic	Type 1	Type 2	Type Extra	Synthetic Migration Model
5 Years (2006)	Mean	16.17	13.40	11.89	7.05
	Median	12.47	9.45	8.63	4.77
	90 Percentile	33.16	26.65	23.60	15.42
10 Years (2011)	Mean	25.62	22.39	18.42	11.49
	Median	19.90	16.24	14.30	8.13
	90 Percentile	51.84	44.41	36.99	25.47

Higher Error Rate than the Synthetic Migration Model (Benchmark)

Result Table (Age-Sex Level)

Result Table (Age-Sex Level)					
Forecast Horizon	Statistic	Type 1	Type 2	Type Extra	Synthetic Migration Model
5 Years (2006)	Mean	16.17	13.40	11.89	7.05
	Median	12.47	9.45	8.63	4.77
	90 Percentile	33.16	26.65	23.60	15.42
10 Years (2011)	Mean	25.62	22.39	18.42	11.49
	Median	19.90	16.24	14.30	8.13
	90 Percentile	51.84	44.41	36.99	25.47

Decrease Around 3% of Error Rate from the Median Perspective

Result Table (Age-Sex Level)

Result Table (Age-Sex Level)					
Forecast Horizon	Statistic	Type 1	Type 2	Type Extra	Synthetic Migration Model
5 Years (2006)	Mean	16.17	13.40	11.89	7.05
	Median	12.47	9.45	8.63	4.77
	90 Percentile	33.16	26.65	23.60	15.42
10 Years (2011)	Mean	25.62	22.39	18.42	11.49
	Median	19.90	16.24	14.30	8.13
	90 Percentile	51.84	44.41	36.99	25.47

Decrease Around 4% of Error Rate from the Median Perspective

Recommendation

- Not Recommend

  (i) Performance

  (ii) Data Characteristic

  (iii) Interpretation

  (iv) Computational Time

Recommendation

- Recommend

  (i) External Variable Improvement

  (ii) Easy Application without Complicated Coding

  (iii) Less Information is Required

- Overall Recommendation & Future Direction

Conclusion & Report

THE UNIVERSITY OF MELBOURNE

- Conclusion of Work

  (i) Data & Benchmark Model

  (ii) Three + Extra Types of LSTM Model Implementation

  (iii) Result and Recommendation

- 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

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

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.

Above_1000 & Below_1000 Area — Preliminary Preprocessing

THE UNIVERSITY OF MELBOURNE

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
…	…

(Valid)

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)
…	…

(Excluded)

Result Table (Total Level)

Result Table (Total Level)					
Forecast Horizon	Statistic	Type 1	Type 2	Type Extra	Synthetic Migration Model
5 Years (2006)	Mean	14.8950	12.2033	10.6295	6.5987
	Median	13.0478	9.7856	9.0743	5.2553
	90 Percentile	21.8927	19.5631	15.4323	11.0650
10 Years (2011)	Mean	24.4915	21.2186	17.4967	11.4337
	Median	20.5182	16.5910	15.0642	9.4145
	90 Percentile	39.7218	33.8293	24.9798	18.8916