THE UNIVERSITY OF MELBOURNE

Faculty of Engineering and Information Technology

Group 12

Do Additional Variables Improve the Accuracy of Total Population Forecasts Generated by Global Machine Learning

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 of Population Forecasting & Long-Short Term Memory
- Data Science Pipeline (Data, Model, Result)
- Conclusion & Recommendation

Team Introduction

THE UNIVERSITY OF MELBOURNE

Chi Zhang

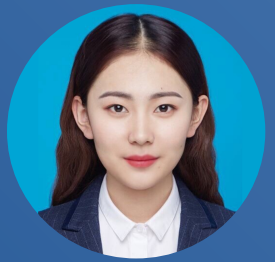
Work Coordinating
Data Analysis
Data Reconstruction
Benchmark Model
LSTM
Model's Evaluation
Report Writing

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

THE UNIVERSITY OF MELBOURNE

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 Small Area (SA3) Population in Age-Sex Cohort

Other Requirement: Comparison between LSTM and Synthetic Migration Model

Challenges

THE UNIVERSITY OF MELBOURNE

- 1. Data Sparsity (Unstable Trend)
- 2. Short Time-Series (Insufficient amount of Data)
- 3. Less Feature Input for the LSTM Model (Compare with the Benchmark)
- 4. Lower Interpretability of Model (Weights during training are not interpretable)
- 5. Computational Consumption (Epoch)
- 6. Error Stack Issue (Retraining & Rolling Update)
- 7. Input Structure (Format)

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 Forecast with the total population (National Projection) in each area
- Change the inward migration flows to maintain consistency
- Apply extra 4 models for creating projection total population data
- Migration Rate, Birth Rate, Death Rate are considered
- Area’s independence

Synthetic Migration Model

- Data Source: Based on SA3 Age-Sex Cohort Data (2 * 18 of Age-Sex Cohorts)
- Investigate Area: 325 Area (Above 1000) + 1 Remainder (Aggregated Below 1000)
- Difference from LSTM: More Features / Variables for Forecasting

  Fertility Rate, Mortality Rate, Migration Rate

  SmallAreaTotal Population from Average-4-Model's Result
- Forecast Result: 2006, 2011 Age-Sex Cohort's Population

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 Collection & Description

THE UNIVERSITY OF MELBOURNE

- All Data are Preliminarily Cooked by Client from the Australia Bureau of Statistic
- Data Scale — SA3 (Statistical Area Level 3)
- Data Format — Age-Sex Cohort's Population in Each Area (351 Areas)
- Time Series Data — Record from 1991 to 2011
- Above_1000 Area — Area with Above 1000 Total Population in Every Year
- Below_1000 Area — Area with Less than 1000 Total Population in One Year

Data Description

THE UNIVERSITY OF MELBOURNE

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	Goulburn 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

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
…	…

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

THE UNIVERSITY OF MELBOURNE

Time-Series Data (21 Years) — 11 Years for Training & 10 Years for Testing

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

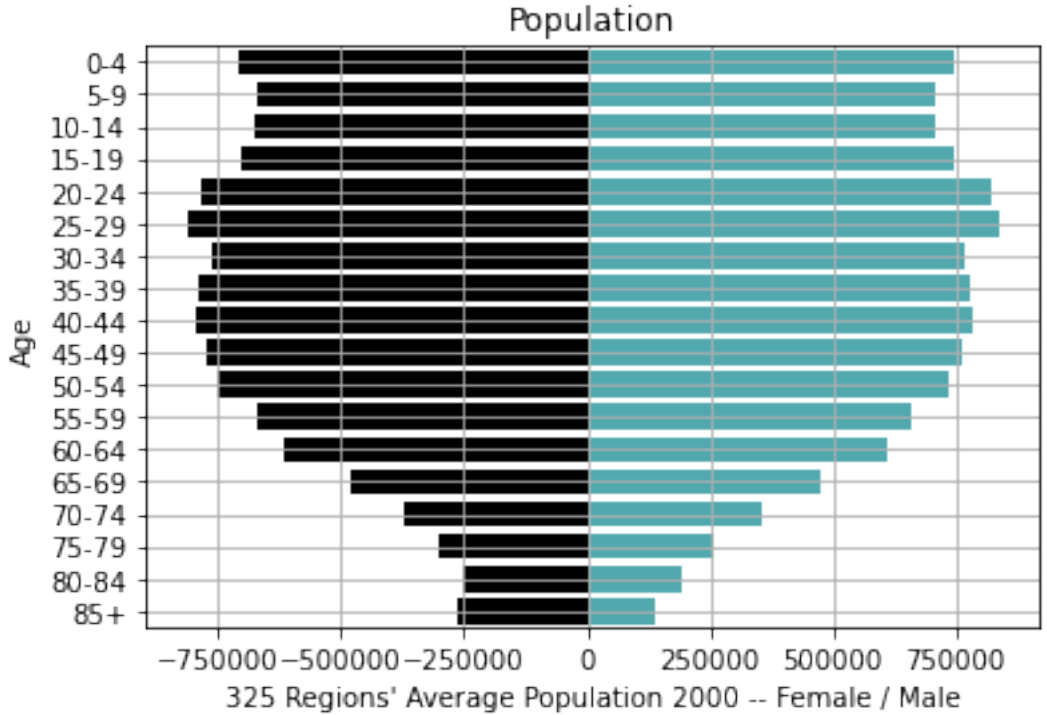
Population Distribution (Pyramid Plot) among each Age-Sex Cohort

Population Growth's Trend (Trend Plot) in Each Area

Data Analysis – Data Sparsity

THE UNIVERSITY OF MELBOURNE

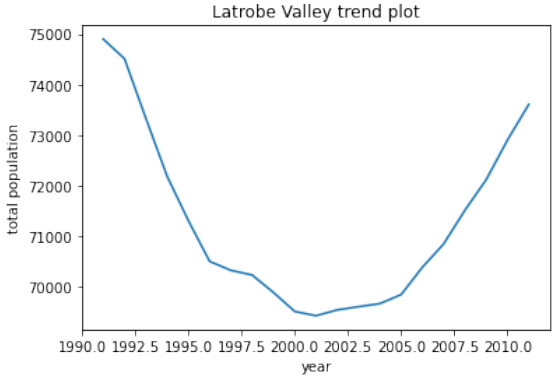
Elder Age-Sex Cohorts Population's Lacking (Visualisation of Anchor Year 2000)

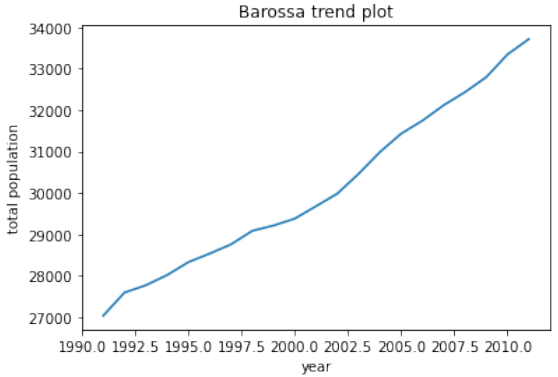


Data Analysis – Population Trends & Clustering

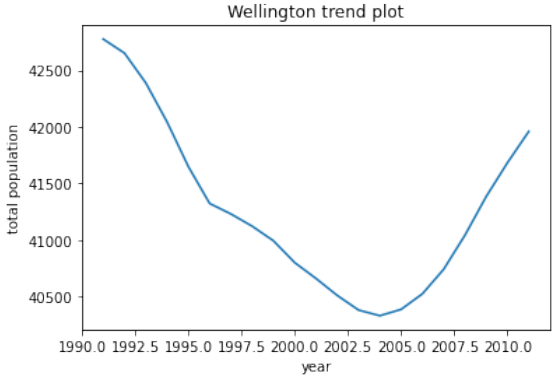
Data Trends – Region's Population Growth Trends' Difference / Similarity

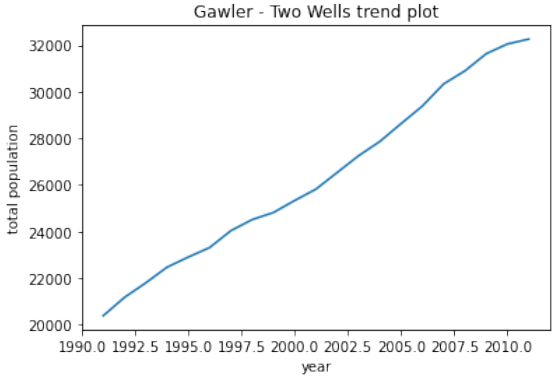
Latrobe Valley vs. Barossa





Wellington vs. Gawler Two Wells



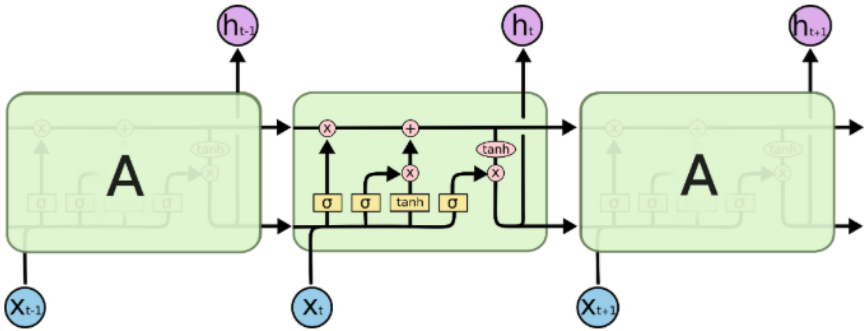


Characteristic of Data

THE UNIVERSITY OF MELBOURNE

- Short Time-Series for Training
- Sparse — Less Population in Elder Age Cohorts for both Male & Female
- Different in Changing Trend — Hard to fit Model with Same Fixed 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 (3 Gates)**

- Input Gate
- Forget Gate
- Output Gate

★ **Age & Sex Prediction**

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

Python Package(s)

THE UNIVERSITY OF MELBOURNE

- 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

THE UNIVERSITY OF MELBOURNE

- Training Set — 1991 - 2001

  Training & Validation Split Varies During Model's Tuning

  (i) LSTM Model Basic (Type 1) — 1991-1998 Training, 1999-2001 Validation

  (ii) LSTM Model Implemented (Type 2 & 3) — Random Splitting

  (iii) LSTM Model Extra Implemented (Type Extra) — Random Splitting

- Test Set — 2002 - 2011

  Models' Evaluation Comparison only Processed in 2006, 2011

LSTM Model Basic (Type 1)

- Sliding Window: 1, 3 Window Gap
- Multivariate Input (2 LSTM Models * 18 Age-Cohort List Input)
- Validation Set: 1999 - 2001
- Fitting & Forecasting

  Rolling Update with Fixed Length Training Set & Fitted Model Epoch = 1000
- Computational Consumption (Large)

LSTM Model Implementation (Type 2 & 3)

THE UNIVERSITY OF MELBOURNE

- Extra Implementation

  (i) Scaling — No Improvement / Introduce Risk

  (ii) Non-Negative — Reasonable Application (No Negative Population)

  (iii) Random-Splitting — Allows Best Year(s) to be Included for Model Fitting

  (iv) Learning Rate / Auto-Stop — No Significant Improvement

  (v) Extra Features — Birth Rate & State & Average Total Population

LSTM Model Extra (Type Extra)

THE UNIVERSITY OF MELBOURNE

- Sliding Window with Step = 1
- Gap = 5
- Inherit Implementation from the Standard Model (Type 2)
- Only Predict the Population in 2006 & 2011 (Special Offer)
- Reduce Computational Consumption (Measured in 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 (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) 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)

Result Table (Basic vs. Benchmark)

THE UNIVERSITY OF MELBOURNE

LSTM Basic Model (Type 1)

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

Benchmark Model

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)

THE UNIVERSITY OF MELBOURNE

LSTM Basic Model (Type 1)

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 Implemented Model (Type 2)

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)

THE UNIVERSITY OF MELBOURNE

LSTM Basic Model (Type 1)

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 Extra Model (Type Extra)

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

THE UNIVERSITY OF MELBOURNE

- Recommend

  (i) External Variable Improvement

  (ii) Easy Application without Complicate Coding

- Not Recommend

  (i) Performance — Does not Outperform than the Benchmark

  (ii) Data Characteristic — Too Short Time-Series

  (iii) Interpretation — Lack of Interpretability of the Black-Box Model

  (iv) Computational Consumption — Too Long for Computing the Result Comparing with the Benchmark

- Overall Recommendation

Conclusion & Report

THE UNIVERSITY OF MELBOURNE

- Conclusion of Work

  (i) Data Preprocessing + Reconstruct Benchmark Model in R

  (ii) Mainly Three Steps of LSTM Model Implementation (Introduce Based on Type 1)

  (iii) Result and Recommendation

- Brief Introduce of Report

  Deeper Illustration of Related Work, Model Interpretation, Result & Discussion

THE UNIVERSITY OF MELBOURNE

Faculty of Engineering and Information Technology

Thank you!

Reference

https://blog.paperspace.com/time-series-forecasting-regression-and-lstm/