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: Dr Kasun Bandara

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

Team Introduction

THE UNIVERSITY OF MELBOURNE

Chi Zhang

Work Coordinating

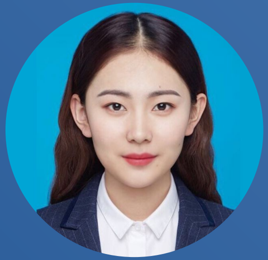
Data Analysis

Eric Luanzon

Data Preprocessing

Haitong Gao

Forecast Reconciliation



Meijun Yue

Data Visualisation

Yuexin Li

Data Visualisation

Meeting Agenda Arranging

Related Works Reading and Preliminary Models' Investigation

Project Background

THE UNIVERSITY OF MELBOURNE

Uses of forecasting

- Planning
- Marketing
- Distribution of funds

Problems with small-area forecasting:

1. Sparse data
2. Larger data requirements
3. Estimation of input data

->

Look into machine learning methods

Data Description

The Main Structure of the Australian Statistical Geography Standard

Australia — State and Territory (S/T) — Statistical Area Level 4 (SA4) — Statistical Area Level 3 (SA3) — Statistical Area Level 2 (SA2) — Statistical Area Level 1 (SA1) — Mesh Block (MB)

Data Description

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
…	…	…

1991
-
2011

Data Description

Male & Female

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)

		Males																	
SA3 Code	SA3 Name	m0-4	m5-9	m10-14	m15-19	m20-24	m25-29	m30-34	m35-39	m40-44	m45-49	m50-54	m55-59	m60-64	m65-69	m70-74	m75-79	m80-84	m85+
10101	Goulburn - Yass	2603	2565	2517	2472	2178	2392	2478	2312	2296	1999	1735	1502	1513	1170	765	506	260	159
10102	Queanbeyan	1593	1362	1223	1406	1743	1803	1670	1398	1346	1094	879	670	617	478	330	198	95	43
10103	Snowy Mountains	741	641	577	591	848	1037	1047	890	743	550	476	402	361	331	242	112	48	28
10104	South Coast	1985	1987	1697	1358	963	1258	1838	2141	1861	1451	1287	1569	2222	2353	1570	966	367	175
10201	Gosford	5373	5182	4631	4802	3928	4659	5303	5160	5048	3972	2993	2518	3055	3389	2852	1936	996	460
10202	Wyong	4332	4072	3481	3469	2956	3482	3957	3646	3392	2564	2108	2108	2809	3162	2477	1736	803	321

Partial Dataframe (Male Group)

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
10401	Clarence Valley
10402	Coffs Harbour
10501	Bourke - Cobar - Coonamble
10502	Broken Hill and Far West
10503	Dubbo
10601	Lower Hunter
10602	Maitland
10603	Port Stephens
10604	Upper Hunter

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)
49999	Special Purpose Codes SA3 (SA)
59797	Migratory - Offshore - Shipping (WA)
59999	Special Purpose Codes SA3 (WA)
69797	Migratory - Offshore - Shipping (Tas.)
69999	Special Purpose Codes SA3 (Tas.)
79797	Migratory - Offshore - Shipping (NT)
79999	Special Purpose Codes SA3 (NT)
80102	Cotter - Namadgi
80104	Gungahlin

Current Findings

THE UNIVERSITY OF MELBOURNE

Data Analysis – Data Sparsity & Population Growth Trend

Related Research

Potential Methods – Forecast Reconciliation & Theta / ETS Model

Preliminary Model – LSTM

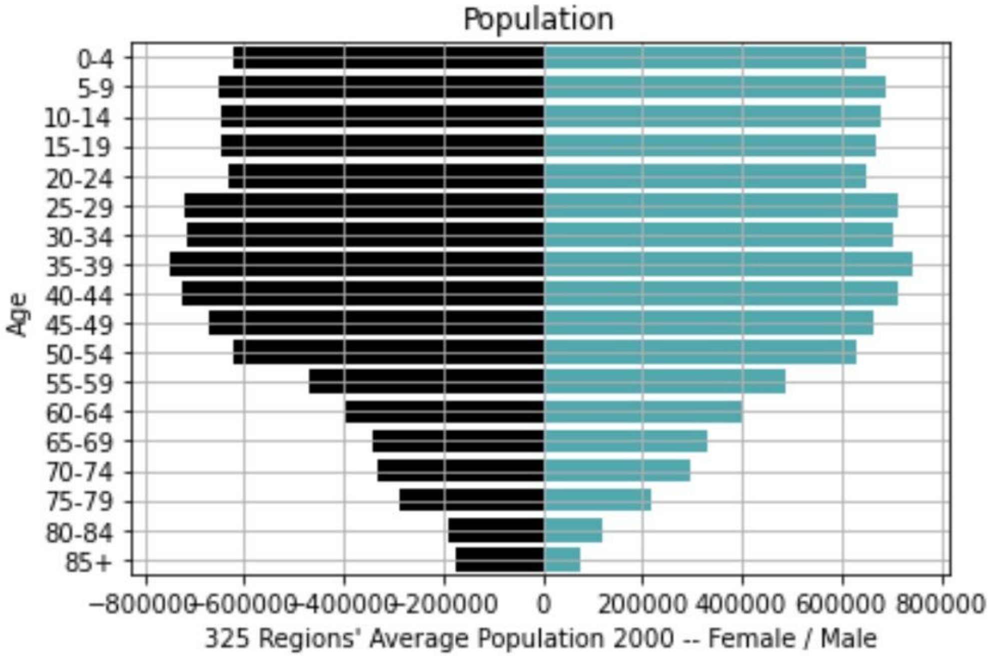
Data Analysis – Descriptive Statistic

Maximum and Minimum of each cohort's population is less informative

Anchoring Year 2000 total population value across all age-sex cohort in all regions

Data Analysis – Data Sparsity

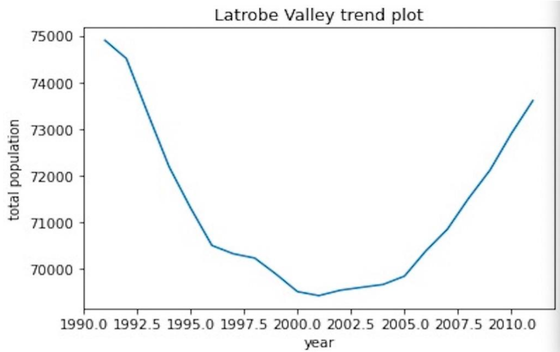
Elder Age-Sex Cohorts Population's Lacking

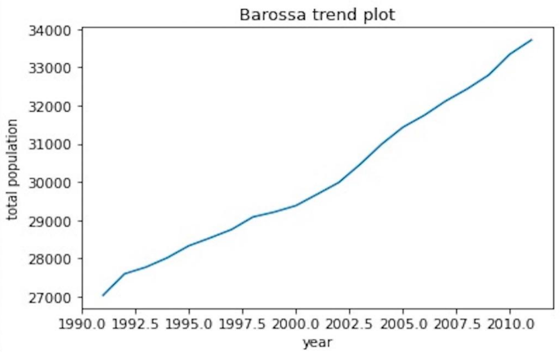


Data Analysis – Population Trends & Clustering

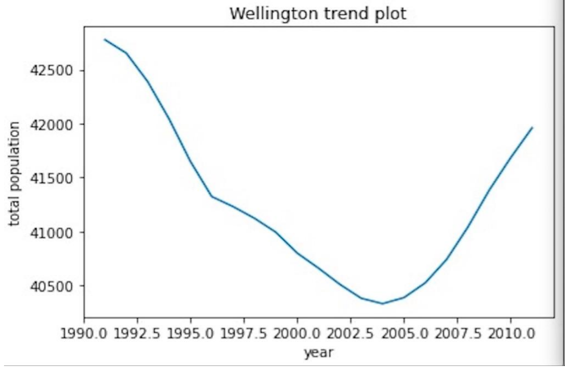
Data Trends – Region's Population Growth Trends' Difference / Similarity

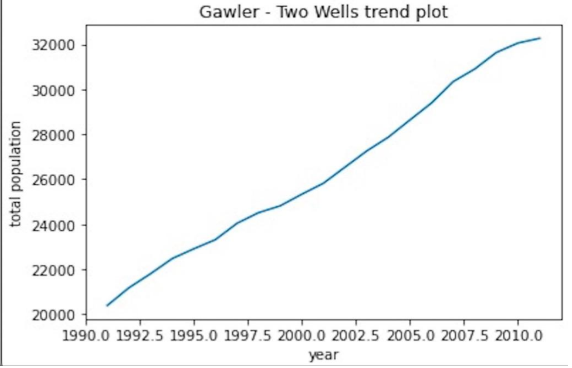
Latrobe Valley vs. Barossa





Wellington vs. Gawler Two Wells





Related Research

1. Hamilton-Perry model:

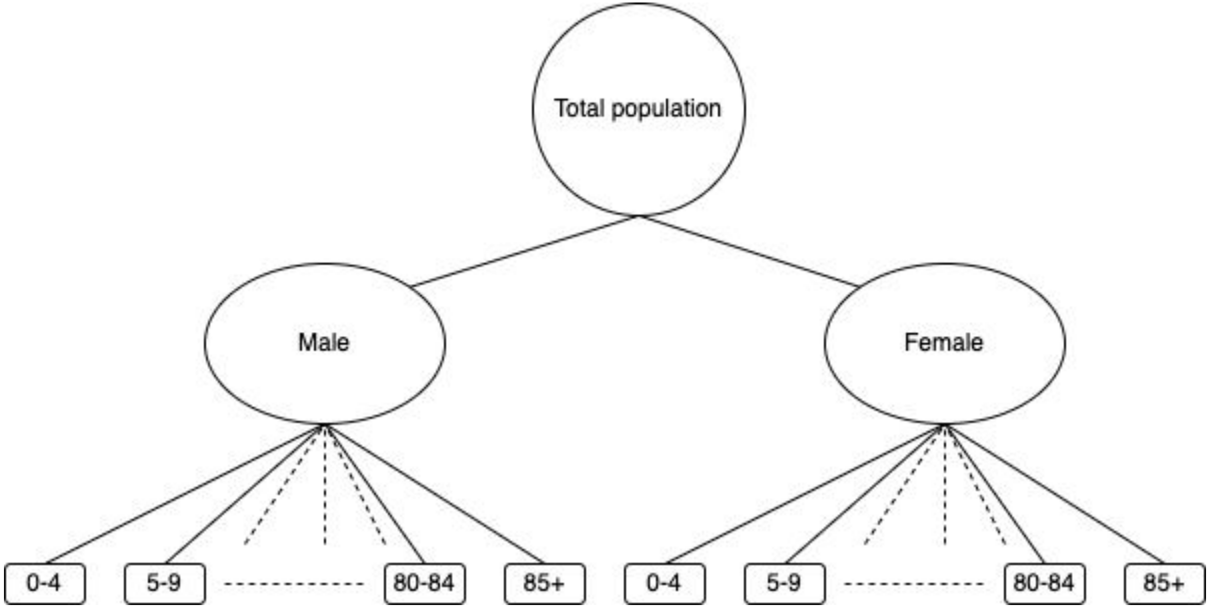
- Could be implemented without migration data, easy to implement
- But less details output

2. Synthetic migration cohort-component model:

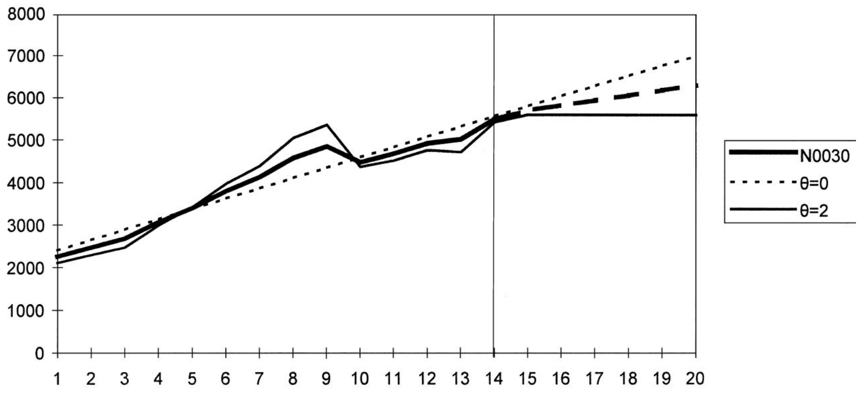
- Contained the impact of the total population of the area
- But changing the inward migration flows to maintain consistency

3. Decided to implement hierarchy models

Forecast Reconciliation Method



Potential Methods



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

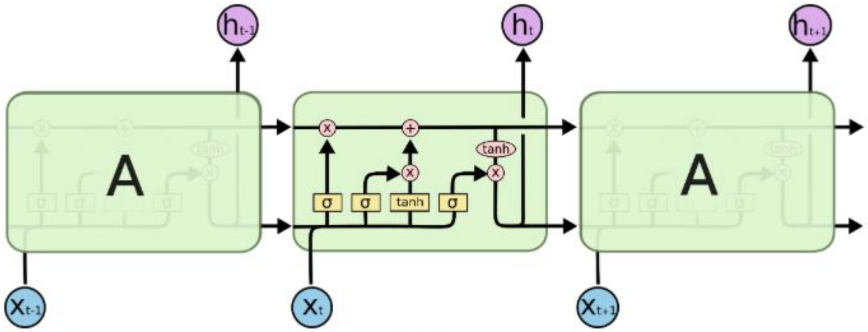
Fioruci, J. A., Pellegrini, T. R., Louzada, F. & Petropoulos, F., 2015. The Optimised Theta Method.

Trend	Seasonal	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

Hyndman, R.J., & Athanasopoulos, G. (2021) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 30 April 2022.

Long-Short Term Memory (LSTM)

THE UNIVERSITY OF MELBOURNE



The repeating module in an LSTM contains four interacting layers.

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

★ **Age & Sex Prediction**

Three Steps for modelling:

- Normalise input data
- Decide time step
- Unidirectional & Bidirectional

Motivations for this project:

- SA3s Data
- Time Range: 21 years
- Multivariate Input: age & sex
- New attributes if possible

Evaluation

THE UNIVERSITY OF MELBOURNE

$$APE_{age-sex} = \frac{\sum_s \sum_a |F_{s,a} - A_{s,a}|}{A} 100.$$

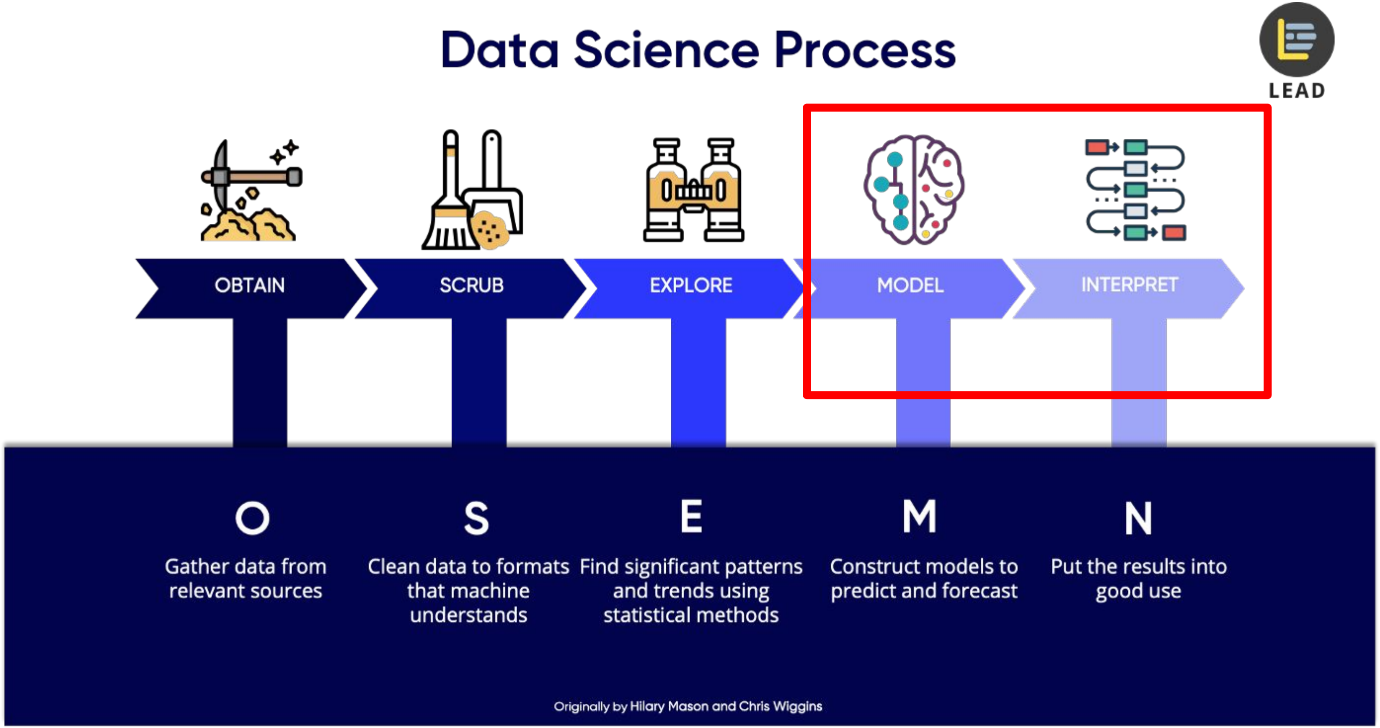
Fit on each:

1991 - 1996 →
1996 - 2001 →
2001 - 2006 →
2006 - 2011 →

Test:

2016
OR
2021

Next Semester Plans



Data Science Process (the O.S.E.M.N. framework)
https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Timeline

THE UNIVERSITY OF MELBOURNE

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, Eric Luanzon, Meijun Yue	
Model Interpretation	Week 6	Week 8	all	Result discussion, finalise code and tidyup.
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	

THE UNIVERSITY OF MELBOURNE

Faculty of Engineering and Information Technology

Thank you!

Exponential Smoothing (ETS)

THE UNIVERSITY OF MELBOURNE

Classification of ETS:

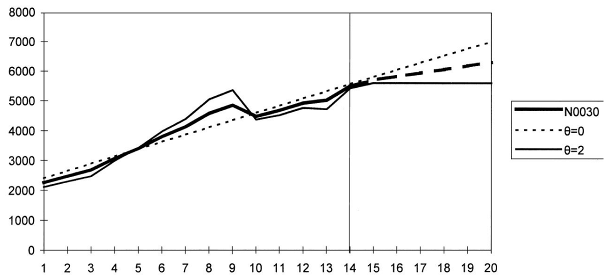
Trend Component	Seasonal Component	Method
N	N	Simple exponential smoothing
A	N	Holt's linear method
A	N	Additive Holt-Winters' method
A	M	Multiplicative Holt-Winters' method
Ad	A	Additive damped trend method
Ad	M	Holt-Winters' damped method

Model Selection:

R code:

```
model <- ETS( data )
```

Theta



	Year	SA3.Name	variable	value	fit	residuals	ape
5201	2007	Goulburn - Yass	m0.4	2153	2055.2944	-97.705642	4.5381162
5526	2008	Goulburn - Yass	m0.4	2214	2027.1605	-186.839466	8.4390003
5851	2009	Goulburn - Yass	m0.4	2286	1999.0267	-286.973289	12.5535122
6176	2010	Goulburn - Yass	m0.4	2358	1970.8929	-387.107113	16.4167563
6501	2011	Goulburn - Yass	m0.4	2350	1942.7591	-407.240936	17.3294015
5202	2007	Queanbeyan	m0.4	1841	1824.6099	-16.390102	0.8902825
5527	2008	Queanbeyan	m0.4	1869	1834.2834	-34.716572	1.8574945
5852	2009	Queanbeyan	m0.4	1897	1843.9570	-53.043043	2.7961541
6177	2010	Queanbeyan	m0.4	1917	1853.6305	-63.369513	3.3056606
6502	2011	Queanbeyan	m0.4	1937	1863.3040	-73.695984	3.8046455
5203	2007	Snowy Mountains	m0.4	613	572.1580	-40.841997	6.6626422
5528	2008	Snowy Mountains	m0.4	627	560.6533	-66.346703	10.5816112
5853	2009	Snowy Mountains	m0.4	649	549.1486	-99.851408	15.3854250
6178	2010	Snowy Mountains	m0.4	668	537.6439	-130.356114	19.5143884

Showing 1 to 14 of 58,496 entries, 7 total columns

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

Fioruci, J. A., Pellegrini, T. R., Louzada, F. & Petropoulos, F., 2015. The Optimised Theta Method.