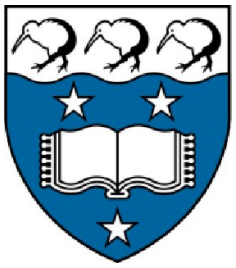
Time Series Visualisation for Auckland Air Quality



ZHAOMING SU

DEPARTMENT OF STATISTICS

THE UNIVERSITY OF AUCKLAND

supervised by

DR. EARO WANG

PROF. CHRIS WILD

A dissertation submitted in partial fulfilment of the requirements for the degree of Bachelor of Science (Honours) in Statistics, The University of Auckland, 2021.

Contents

Abstract

The abstract should outline the main approach and findings of the thesis and must not be more than 500 words.

Acknowledgements

I would like to thank my pet goldfish for ...

Copyright notice

(Standard thesis)

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

(Thesis including published works declaration)

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes (insert number) original papers published in peer reviewed journals and (insert number) submitted publications. The core theme of the thesis is (insert theme). The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the (insert name of academic unit) under the supervision of (insert name of supervisor).

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.) Remove this paragraph for theses with sole-authored work

In the case of (insert chapter numbers) my contribution to the work involved the following:

I have / have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: ZHAOMING SU

Student signature:

Date:

Thesis chapter	**Publication title**	**Status (published, in press, accepted or returned for revision)**	**Nature and % of student contribution**	**Co-author name(s), nature and % of co-author's contribution**	**Co-author(s), Monash student Y/N**
2	xx	xx	xx	xx	N
3	xx	xx	xx	xx	N
4	xx	xx	xx	xx	N
5	xx	xx	xx	xx	N

Preface

The material in Chapter 1 has been submitted to the journal *Journal of Impossible Results* for possible publication.

The contribution in Chapter 2 of this thesis was presented in the International Symposium on Nonsense held in Dublin, Ireland, in July 2015.

Chapter 1

Introduction (Template Demo)

This is where you introduce the main ideas of your thesis, and an overview of the context and background.

In a PhD, Chapter 2 would normally contain a literature review. Typically, Chapters 3–5 would contain your own contributions. Think of each of these as potential papers to be submitted to journals. Finally, Chapter 6 provides some concluding remarks, discussion, ideas for future research, and so on. Appendixes can contain additional material that don't fit into any chapters, but that you want to put on record. For example, additional tables, output, etc.

1.1 Rmarkdown

In this template, the rest of the chapter shows how to use Rmarkdown. The big advantage of using Rmarkdown is that it allows you to include your R code directly into your thesis, to ensure there are no errors in copying and pasting, and that everything is reproducible. It also helps you stay better organized.

For details on using *R Markdown* see `http://rmarkdown.rstudio.com`.

1.2 Data

Included in this template is a file called `sales.csv`. This contains quarterly data on Sales and Advertising budget for a small company over the period 1981–2005. It also contains the GDP (gross domestic product) over the same period. All series have been adjusted for inflation. We can load in this data set using the following command:

```
sales <- ts(read.csv("data/sales.csv")[, -1], start = 1981, frequency = 4)
```

Any data you use in your thesis can go into the data directory. The data should be in exactly the format you obtained it. Do no editing or manipulation of the data outside of R. Any data munging should be scripted in R and form part of your thesis files (possibly hidden in the output).

1.3 Figures

Figure 1.1 shows time plots of the data we just loaded. Notice how figure captions and references work. Chunk names can be used as figure labels with `fig:` prefixed. Never manually type figure numbers, as they can change when you add or delete figures. This way, the figure numbering is always correct.

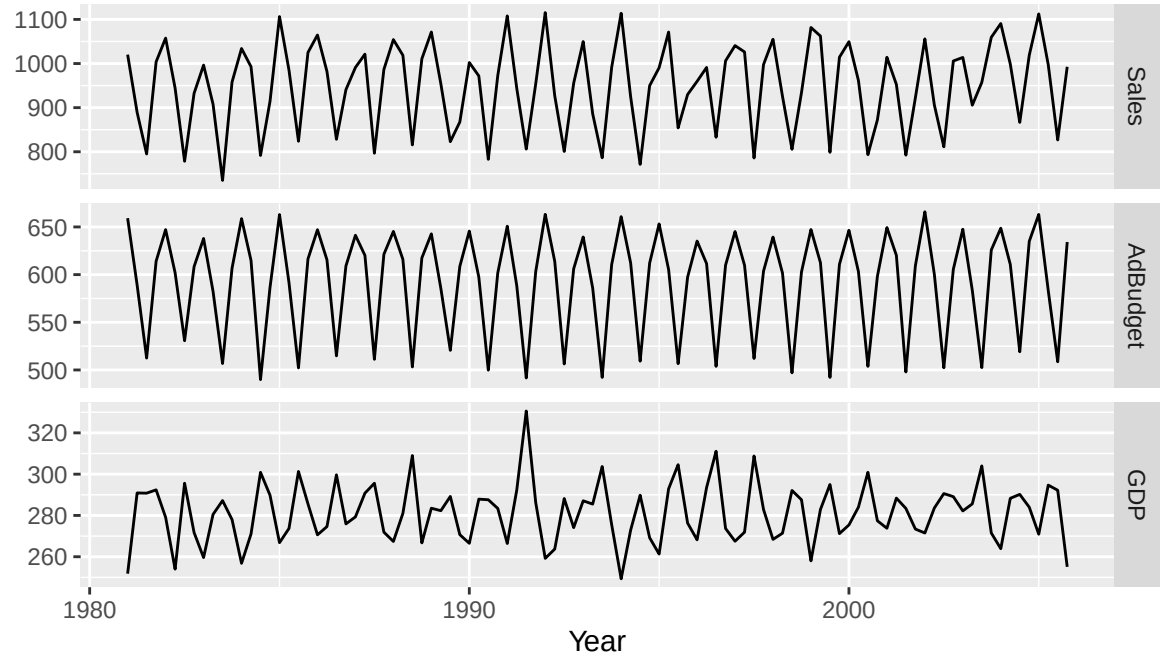


Figure 1.1: *Quarterly sales, advertising and GDP data.*

1.4 Results from analyses

We can fit a dynamic regression model to the sales data.

If y_t denotes the sales in quarter t, x_t denotes the corresponding advertising budget and z_t denotes the GDP, then the resulting model is:

$$y_t - y_{t-4} = \beta(x_t - x_{t-4}) + \gamma(z_t - z_{t-4}) + \theta_1\varepsilon_{t-1} + \Theta_1\varepsilon_{t-4} + \varepsilon_t \tag{1.1}$$

where $\beta = 2.28$, $\gamma = 0.97$, $\theta_1 = NA$, and $\Theta_1 = -0.90$.

1.5 Tables

Let's assume future advertising spend and GDP are at the current levels. Then forecasts for the next year are given in Table 1.1.

Again, notice the use of labels and references to automatically generate Table numbers. In this case, we need to generate the label ourselves.

Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1000.2	947.7	1052.7	919.9	1080.5
1013.1	959.3	1066.8	930.9	1095.3
1076.7	1022.9	1130.6	994.4	1159.0
1003.5	949.7	1057.4	921.2	1085.8

Table 1.1: *Forecasts for the next year assuming Advertising budget and GDP are unchanged.*

The `knitLatex` package is useful for generating tables from R output. Other packages can do similar things including the `kable` function in `knitr` which is somewhat simpler but you have less control over the result. If you use `knitLatex` to generate tables, don't forget to include `results="asis"` in the chunk settings.

Chapter 2

Literature Review (Draft)

2.1 Exponential smoothing (Template Demo)

Exponential smoothing was originally developed in the late 1950s (Brown, 1959, 1963; Holt, 1957; Winters, 1960). Because of their computational simplicity and interpretability, they became widely used in practice.

Empirical studies by Makridakis and Hibon (1979) and Makridakis et al. (1982) found little difference in forecast accuracy between exponential smoothing and ARIMA models. This made the family of exponential smoothing procedures an attractive proposition (see Chatfield et al., 2001).

The methods were less popular in academic circles until Ord, Koehler, and Snyder (1997) introduced a state space formulation of some of the methods, which was extended in Hyndman et al. (2002) to cover the full range of exponential smoothing methods.

Chapter 3

The Auckland Environmental Data

3.1 Introduction

The data, provided by Auckland Council (2003-2021), includes data on 14 parameters from 10 monitoring stations in Auckland from as North as Takapuna to as South as Patumahoe. Parameters consist of air quality index (AQI), 10 pollutant levels and four other meteorological variables, reported in hourly resolution, with various starting dates (since as early as 2003 in Takapuna) until April 2021. All time-stamps are in New Zealand Standard Time (UTC+12) to avoid duplicated index upon boundaries of daylight saving.

Parameter	Unit	Note
AQI		Air Quality Index
BC(370)	ngm^{-3}	Black carbon at 370nm wavelength
BC(880)	ngm^{-3}	Black carbon at 880nm wavelength
NO	μgm^{-3}	Nitrogen monoxide concentration
NO_2	μgm^{-3}	Nitrogen dioxide concentration
NOx	μgm^{-3}	Nitrogen oxides concentration
PM2.5	μgm^{-3}	Particulate matter with diameter <2.5μm
PM10	μgm^{-3}	Particulate matter with diameter <10μm

Table 3.1: *Parameters available in the data.*

Parameter	Unit	Note
SO_2	μgm^{-3}	Sulphur monoxide concentration
O_3	μgm^{-3}	Ozone concentration
CO	mgm^{-3}	Carbon monoxide concentration
Relative Humidity	%	
Temperature	°C	
Wind Speed	ms^{-1}	
Wind Direction	°	
AQI Level of Concern		Derived from AQI

Table 3.1: *Parameters available in the data.*

3.2 Data Cleaning

The final data set is composed of two separate data sets, each with a different data structure. Cleaning and manipulation are needed for ensuring the consistency of format for combination. The desired structure of the final data consists of observations, each uniquely identified by the date-time and location serving as the index and key, with records of all parameters.

The environmental parameter data set is in long format, with each observation consisting of a single record of one parameter. In total, there are 7,461,261 records from 10 locations with 14 parameters in the data set. Preliminary inspection finds that 4,654 records (0.06% of total) have a missing time-stamp and 60,312 records (0.81%) have missing value. Besides, 104,332 records are found to have a negative value. Nevertheless, all pollutants are reported in units in the form of mass per unit volume, and other parameters except for temperature are only sensible if positive (Boamponsem, 2021). Thus, 104,257 records of insensible negative values are removed. Of the 7,292,038 time-and-numerically sensible

records, 239,374 (3.28%) are duplicate with 120,207 redundant records. Further checking reveals that 230,822 of the duplicates have inconsistent values. However, as the scale of the inconsistency of most duplicate records is small, the first-appearing records of duplicates are kept. The cleaned data set consists of 7,171,831 (96.12%) valid records with 719,640 records (10.03%) of implicit time gaps.

The parameter wind direction is from a differently structured data set, with each observation consisting of wind direction records of all locations uniquely identified by a date-time stamp. Of the seven locations of interest (locations also in the environmental data set), the data set includes 940,632 records with 207,436 (22.05%) missing values but no implicit time gaps. It is worth noting that 39,193 records (4.17%) have an inconsistent format for their time-stamp.

The clean data set consists of 1,138,535 observations and 18 variables (date-time as the index, location as the key, 15 original parameters and one derived parameter) from Jun 1, 2003, to Apr 30, 2021. The time range is different for each location, with Customs St as short as since Jan 1, 2020. After accounting for all implicit time gaps, 9,173,003 records have missing values, implying a real missing rate of 53.71%. It is noteworthy that extreme values from input error are more frequent in the earlier years. The final data set is subsetted from the year 2016.

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.

Bibliography

Auckland Council (2003-2021). *The Auckland Environmental Data*. Accessed: 2021-05-25. Auckland, New Zealand. https://environmentauckland.org.nz/Data/DataSet/.

Boamponsem, L (2021). *Auckland's Air Quality*. Auckland Council. Auckland, New Zealand. https://sites.google.com/aucklanduni.ac.nz/louis/auckland/.

Brown, RG (1959). *Statistical forecasting for inventory control*. McGraw-Hill, New York.

Brown, RG (1963). *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, New Jersey: Prentice Hall.

Chatfield, C, AB Koehler, JK Ord, and RD Snyder (2001). A new look at models for exponential smoothing. *The Statistician* **50**, 147–159.

Holt, CE (1957). *Forecasting trends and seasonals by exponentially weighted averages*. O.N.R. Memorandum 52/1957. Carnegie Institute of Technology.

Hyndman, RJ, AB Koehler, RD Snyder, and S Grose (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **18**(3), 439–454.

Makridakis, S, A Anderson, R Carbone, R Fildes, M Hibon, RLJ Newton, E Parzen, and R Winkler (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting* **1**, 111–153.

Makridakis, S and M Hibon (1979). Accuracy of forecasting: an empirical investigation (with discussion). *Journal of Royal Statistical Society (A)* **142**, 97–145.

Ord, JK, AB Koehler, and RD Snyder (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of American Statistical Association* **92**, 1621–1629.

Winters, PR (1960). Forecasting sales by exponentially weighted moving averages. *Management Science* **6**, 324–342.