University of Technology Sydney (UTS)
School of Mathematical and Physical Sciences
Faculty of Science

Aggregation in Regression Analysis of Very Large Time Series Datasets

by

Alan Malecki

A thesis submitted
in partial fulfilment of the
requirements for the degree

Doctor of Philosophy

Sydney, Australia

Certificate of original ownership

I, Alan Malecki declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Science, at the University of Technology Sydney (UTS).

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training program.

Signature:

Production Note:
Signature removed prior to publication.

Date: 30/04/2020

Acknowledgement

This thesis and my life as a statistician would not be possible without my supervisor Louise Ryan. You have moulded me into the researcher and person I am now. You took me on as an honours student, showed me the exciting possibilities of statistics, and developed an interest that cannot be satisfied. You have made this journey an enjoyable and rich experience, and furthermore, you have made an immeasurable contribution to my life. When we began this journey, I never thought I would have such an incredible mentor and friend. Scott Sisson, my co-supervisor, I have truly appreciated the time you have given me. It has been a pleasure coming to your office and discussing my thesis, as well as many other topics. To my partner, Georgia, who supported me at all times, and gave me hope and joy throughout, we've done it together. Lastly, I thank my parents, who have given me everything. Nothing would have been possible without your love and support.

Contents

List of Figures

List of Tables

Abstract

The focus of this thesis is on the analysis of large and complex data. Computer memory constraints can prohibit the analysis of large datasets, and this issue is further complicated when faced with complex data. We are motivated by an environmental dataset concerning air particulate measurements and the impact of passing coal transport trains. This dataset has over 600,000 observations and is complicated by it's long memory dependence. Current methods for long memory time series are limited to small datasets. To overcome these issues, we consider two approaches for the analysis of large and complex data:

1. transforming data such that its volume and complexity is reduced, and,
2. extending current statistical methods for big data to allow for complex data structures.

The use of temporal aggregation transforms the dataset to a more manageable size. This permits the use of an AutoRegressive Fractionally Integrated Moving Average (ARFIMA) process on our motivating dataset. We also consider transforming the data to a bivariate series to reduce the loss of information due to this temporal aggregation.
Divide and Recombine is a modern approach to analysing big data. This approach for big data analysis has not yet been extended to the time series setting. We explore this situation and extend the D&R process for long memory time series.