

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
© Alan Malecki, 2020. All rights reserved.

Permission is herewith granted to University of Technology Sydney
to circulate and to have copied for non-commercial purposes,
at its discretion, the above title upon request of individuals and institutions.

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

1	Introduction	1
2	Hunter Valley Coal Train Dataset and Previous Analyses	3
2.1	Introduction	3
2.2	Hunter Valley Coal Train dataset	3
2.3	Previous analyses	6
2.3.1	Initial analysis by a private company	7
2.3.2	Re-analyses of the the initial study.	7
2.4	Temporal aggregation	8
2.5	Long memory dependence	11
2.6	Discussion	12
3	Linear Regression of a Long Memory Time Series with ARFIMA Errors	13
3.1	Introduction	13
3.1.1	Literature review	14
3.2	Long memory regression models	17
3.3	Simulations	18
3.3.1	Data generating model	19
3.3.2	Comparison of regression with iid errors and ARFIMA errors	20
3.3.3	Effect of temporal aggregation	25
3.3.4	Review of simulations	28
3.4	Application	29
3.4.1	Temporal aggregation of the Hunter Valley Coal Train dataset	30
3.4.2	Application: Linear regression with ARFIMA errors	31
3.4.3	Review of application	36
3.5	Discussion	37
4	Impact of Model Misspecification for Time Series Modelling	39
4.1	Introduction	39
4.1.1	Literature review	40
4.1.2	Linear regression with ARFIMA errors	41
4.2	Model misspecification	41
4.2.1	Simple model: No tail	42
4.2.2	Assumed model: Short tail	45
4.2.3	Assumed model: Long tail	47
4.3	Simulations	49
4.3.1	Data generating model	50
4.3.2	Evaluating our theoretical results: Unaggregated data	51
4.3.3	Simple model: No tail	51
4.3.4	Assumed model: Short tail	52
4.3.5	Assumed model: Long tail	53
4.3.6	Comparison of simple and assumed models against the true model: Independent error structure	53
4.3.7	Effect of temporal aggregation on simple, assumed, combined and true models: Long memory error structure	55

4.3.8	Selection of tail length for the assumed model	59
4.3.9	Review of simulations	61
4.4	Application: Hunter Valley Coal Train dataset	61
4.4.1	Comparison of the simple and assumed models	62
4.4.2	Selection of tail length for the assumed model	64
4.4.3	Fitting of the combined model	69
4.4.4	Review of application	70
4.5	Discussion	70
5	Bivariate Time Series Modelling	73
5.1	Introduction	73
5.2	Aggregation and mixed effect models for bivariate time series	74
5.2.1	Temporal aggregation for a bivariate time series	74
5.2.2	Mixed effect models for bivariate time series	75
5.2.3	Model 1	76
5.2.4	Model 2	76
5.2.5	Model 3	79
5.2.6	Model 4	82
5.2.7	Model 5	84
5.3	H-Likelihood for bivariate time series	86
5.4	Simulations: Fits of models 2 through 5	89
5.4.1	Data generating model	90
5.4.2	Fitting model 2	91
5.4.3	Fitting model 3	93
5.4.4	Fitting model 4	95
5.4.5	Fitting model 5	97
5.4.6	Review of model fits	99
5.5	Simulations: Effect of aggregation	100
5.5.1	Data generating model	101
5.5.2	Fitting for data from option 1	102
5.5.3	Fitting for data from option 2	105
5.5.4	Review of effect of aggregation	108
5.6	Application	109
5.6.1	Fit of model 1	109
5.6.2	Fit of model 2	110
5.6.3	Fit of model 3	112
5.6.4	Fit of model 4	112
5.6.5	Fit of model 5	113
5.6.6	Review of model fits	115
5.7	Discussion	115
6	Divide And Recombine in a Time Series Setting	117
6.1	Introduction	117
6.1.1	Literature review	118
6.2	Divide and Recombine process	119
6.2.1	Division step	120

6.2.2	Replicate division	120
6.2.3	Conditioning-variable division	120
6.2.4	Analysis step	121
6.2.5	Recombine step	121
6.2.6	Divide and Recombine for time series data	122
6.3	Simulations	122
6.3.1	Data generating model	123
6.3.2	Models in simulations	124
6.3.3	Linear regression with iid errors	124
6.3.4	Linear regression with long memory errors	125
6.3.5	Review of simulations	127
6.4	Application: Seperate train and tails model	128
6.4.1	Unaggregated Divide and Recombine model	128
6.4.2	Aggregated Divide and Recombine model	131
6.4.3	1 minute aggregation	131
6.4.4	5 minute aggregation	133
6.4.5	Review of application	135
6.5	Application: Combined train and tails model	138
6.6	Comparison of Divide and Recombine with models using full dataset (as in chapter 4)	141
6.6.1	Seperate train and tails model	141
6.6.2	Combined train and tails model	142
6.6.3	Time comparison of chapter 4 and chapter 6 analyses	142
6.7	Discussion	143
7	Discussion and Future Research	145

List of Figures

2.1	Recorded levels of air particulates PM1, PM2.5, PM10 and TSP for a ten minute period on 30 November 2012.	4
2.2	Time taken to pass the monitor for each train type (in minutes).	5
2.3	Log transformed TSP data during a 6-hour period on 9 December 2012. Plotting symbols are colour-coded to indicate the presence of various train types.	6
2.4	Comparison of unaggregated and aggregated log(TSP+1) data, with an aggregation period of 5 minutes.	9
2.5	Comparison of unaggregated and aggregated data for each passing train variable. The aggregation period is 5 minutes. We show the indicator variable for passing coal(empty and loaded), freight and passenger trains, as well as their respective proportions as a result of temporal aggregation. The green vertical lines indicate each 5 minute block upon which the aggregation is performed.	10
2.6	ACF and PACF plots for the unaggregated and aggregated (5 minutes) log transformed TSP measurements.	11
3.1	Inspection of residuals for one of the 20 replicates of the linear regression with gaussian iid errors and ARFIMA long memory errors, for both unaggregated and aggregated data. ACF and PACF plots for each model. ACFs show the autocorrelation structure of the errors as well as the presence of long memory in the data. The PACF's show the moving average aspect of the ARMA structure.	22
3.2	Comparison of intercept estimate for linear regression with gaussian iid errors and arfima long memory errors, for both unaggregated and aggregated data. Standard Error bars are included as well as the true $\beta_0 = 3$ as indicated by the solid green line. The solid black and blue lines are the aggregated models, for gaussian and ARFIMA models respectively. The dotted cyan and red lines are the unaggregated gaussian and ARFIMA models.	23
3.3	Comparison of passing train covariate for linear regression with gaussian iid errors and arfima long memory errors, for both unaggregated and aggregated data. Standard Error bars are included as well as the true $\beta_1 = 5$ as indicated by the solid green line. The solid black and blue lines are the aggregated models, for gaussian and ARFIMA models respectively. The dotted cyan and red lines are the unaggregated gaussian and ARFIMA models.	24
3.4	Comparison of the intercept estimates, $\hat{\beta}_0$, for aggregated gaussian iid and ARFIMA(p,d,q) regression as the aggregation interval increases. The true value is denoted by green line. Gaussian iid error implementation is shown by the blue shaded boxplots, while the ARFIMA errors are shown by the grey shaded boxplots.	26
3.5	Comparison of passing train covariate estimates, $\hat{\beta}_1$, for aggregated gaussian iid and ARFIMA(p,d,q) regressions as the aggregation interval increases. The true value is denoted by green line. Gaussian iid error implementation is shown by the blue shaded boxplots, while the ARFIMA errors are shown by the grey shaded boxplots.	27
3.6	Outcomes of linear regression with ARFIMA errors, as in model (20), for the aggregation periods of 5 minutes to 2 hours. The intercept is the black line, with loaded empty coal trains in red, freight trains in green, loaded coal in blue, passenger trains in light blue and unknown in purple.	33

3.7	Outcomes of linear regression with ARFIMA errors, as in model (20), for the aggregation periods of 5 minutes to 2 hours. Empty coal trains are in red, freight trains in green and loaded coal in blue. We have included standard errors for each train type.	34
3.8	Residual analysis showing ACF and PACF plots for unaggregated iid model, aggregated iid model, and aggregated ARFIMA(1,0.43,2) model at $J = 100$ (10 minute aggregation).	36
4.1	Illustration of passing train and tail indicators.	42
4.2	Comparison of alternate tail lengths under the assumed model with ARFIMA errors on simulated data with long memory and $\beta_1 = \beta_2 = 5$. We analyse the unaggregated data and then we consider the aggregated data for the aggregation periods from $J=5$ to $J=40$. The solid lines show the $\hat{\beta}_1$ estimates which correspond to the train effect, and the dotted lines show the $\hat{\beta}_2$ estimates which correspond to the assumed train tails. The true set value is shown by the cyan line.	60
4.3	Comparison of empty coal train and tail covariates from the simple and assumed models (35) and (36) with both iid and ARFIMA error structures over the aggregation periods of 5 minutes to 2 hours.	64
4.4	Comparison of Empty Coal train and tail covariates ($\hat{\beta}_1$ and $\hat{\beta}_2$) for 1, 2, 3, 4 and 5 minute tails. We analyse the assumed model for 5 to 10 minute aggregations.	65
4.5	Comparison of Freight Coal Train and Tail Covariates ($\hat{\beta}_3$ and $\hat{\beta}_4$) for Alternate Tail Choices	66
4.6	Comparison of Loaded Coal Train and Tail Covariates ($\hat{\beta}_5$ and $\hat{\beta}_6$) for Alternate Tail Choices	67
5.1	Model 2 Simulations, 50 Replicates; plot (i) shows the beta coefficient estimates, and plot (ii) shows the variance components	92
5.2	Method of Moments Model 2: (i) & (iv) have $n = 1,000$; (ii) & (v) have $n = 10,000$; and (iii) & (vi) have $n = 100,000$.	93
5.3	Model 3 Simulations, 50 Replicates; plot (i) shows the beta coefficient estimates, and plot (ii) shows the variance components	94
5.4	Method of Moments Model 3: (i) & (iv) have $n = 1,000$; (ii) & (v) have $n = 10,000$; and (iii) & (vi) have $n = 100,000$.	95
5.5	Model 4 Simulations, 50 Replicates; plot (i) shows the beta coefficient estimates, and plot (ii) shows the variance components	96
5.6	Method of Moments Model 4: (i) & (iv) have $n = 1,000$; (ii) & (v) have $n = 10,000$; and (iii) & (vi) have $n = 100,000$.	97
5.7	Model 5 Simulations, 50 Replicates; plot (i) shows the beta coefficient estimates, and plot (ii) shows the variance components	98
5.8	Method of Moments Model 5: (i) & (iv) have $n = 1,000$; (ii) & (v) have $n = 10,000$; and (iii) & (vi) have $n = 100,000$.	99
5.9	Method of Moments Model 5b: (i) & (iv) have $n = 1,000$; (ii) & (v) have $n = 10,000$; and (iii) & (vi) have $n = 100,000$.	100
5.10	First 100 observations of Option 1 simulated dataset including aggregated data	103
5.11	Aggregation Effect for Model 1:i. $J=10$, ii. $J=20$, iii. $J=50$, iv. $J=100$; $n=10,000$, with red crosses indicating the initial true parameter as set in the unaggregated data simulation; 5 replicates; Beta Coefficients	104

5.12	Aggregation Effect for Model 1:i. J=10, ii. J=20, iii. J=50, iv. J=100 ; n=10,000, with red crosses indicating the initial true parameter as set in the unaggregated data simulation; 5 replicates; Variance Components	105
5.13	First 100 observations of Option 2 simulated dataset including aggregated data . . .	106
5.14	Aggregation Effect for Model 2:i. J=10, ii. J=20, iii. J=50, iv. J=100 ; n=10,000, with red crosses indicating the initial true parameter as set in the unaggregated data simulation; 5 replicates; Beta Coefficients	107
5.15	Aggregation Effect for Model 2:i. J=10, ii. J=20, iii. J=50, iv. J=100 ; n=10,000, with red crosses indicating the initial true parameter as set in the unaggregated data simulation; 5 replicates; Variance Components	108
5.16	Application: Fit of Model 1	110
5.17	Application: Fit of Model 2	111
5.18	Application: Fit of Model 3	112
5.19	Application: Fit of Model 4	113
5.20	Application: Fit of Model 5	114
6.1	Comparison of estimated ARFIMA(p,d,q) parameters for simulations with K=2. 20 replicates for each model. The third and fourth rows of this figure have 40 estimates as the Divide and Recombine has split each replicate into two periods.	127
6.2	Lengths of L_k for each subset under conditioning-variable division.	128
6.3	Comparison of ARFIMA(p,d,q) order for the residuals of each Divide and Recombine subset for the division of 500 and 5000 observations. Unaggregated data.	130
6.4	Comparison of ARFIMA(p,d,q) order for the residuals of each Divide and Recombine subset for the division of 100 and 2000 observations. 1 minute aggregation.	133
6.5	Comparison of ARFIMA(p,d,q) order for the residuals of each Divide and Recombine subset for the division of 100 and 2000 observations. 5 minute aggregation.	135
6.6	Divide and Recombine coefficient estimates for the Assumed model. Data divided into replicates and by conditioning-variables for each day.	136
6.7	Divide and Recombine coefficient estimates for each train type by aggregation period in each subset. Data divided by conditioning-variables for each day. (In the unaggregated case, we have a further subdivision within each day.)	137
6.8	Divide and Recombine coefficient estimates for each train TAIL type by aggregation period in each subset. Data divided by conditioning-variables for each day. (In the unaggregated case, we have a further subdivision within each day.)	137
6.9	Divide and Recombine coefficient estimates for the Combined model. Data divided by conditioning-variables for each day, and in the unaggregated case, it is further divided within each day. 4 minute tails.	139
6.10	Divide and Recombine coefficient estimates for each train type by aggregation period in each subset. Data divided by conditioning-variables for each day. (In the unaggregated case, we have a further subdivision within each day.) Combined Model.	140

List of Tables

2.1	Information about trains passing during the study period. Columns show median value, along with lower and upper quartiles (LQ, UQ).	4
2.2	Results of the generalized additive model from equation (1) for logged TSP values. .	8
3.1	Simulation Results for ARFIMA(p,d,q) error structure for unaggregated and aggregated models ($J = 10$ as the aggregation block size).	28
3.2	Temporal aggregation effect on ARFIMA(p,d,q) errors from model (20) for aggregation periods from 5 minute to 2 hours.	35
4.1	Outcomes for the simple model in (23). The set value is the simulated value from the data generating model. The expected value is what we expect the estimate to be based off of the algebra above and the estimated value is the simulated estimate from our analysis. All simulations have an iid error structure. We include the means of all the simulation outcomes in the final row.	51
4.2	Assumed model with assumed tail length half of the true tail for the $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ estimates. The set value is the simulated value from the data generating model. The expected value is what we expect the estimate to be based off of the algebra above and the estimated value is the simulated estimate from our analysis. All simulations have an iid error structure. We include the means of all the simulation outcomes in the final row.	52
4.3	Assumed model with assumed tail length greater than the true tail for the $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ estimates. The set value is the simulated value from the data generating model. The expected value is what we expect the estimate to be based off of the algebra above and the estimated value is the simulated estimate from our analysis. All simulations have an iid error structure. We include the means of all the simulation outcomes in the final row.	53
4.4	Estimated ($\hat{\beta}_0$) from the simple, true and assumed mode from data generated from equation (32) with iid gaussian error structure. For the assumed model, we consider 3 possible tail covariates. The first column has estimates for the assumed tail of length 5 observations, the 2nd column has an assumed tail of 10 observations, which is equal to the true tail length. The final column shows the estimates for the assumed model with a tail length set to 15 observations. We compare all three assumed tail lengths with the simple model (no tail covariate), and the true model where the tail length is known.	54
4.5	Estimated $\hat{\beta}_1$ and $\hat{\beta}_2$ from models 1-3 from data generated from equation (32) with iid gaussian error structure. For the assumed model, we consider 3 possible tail covariates. The first column has estimates for the assumed tail of length 5 observations, the 2nd column has assumed tail of 10 observations, which is equal to the true tail length. The final column shows the estimates for the assumed model with a tail length set to 15 observations. We compare all three assumed tail lengths with the simple model (no tail covariate), and the true model where the tail length is known.	55

4.6	Coefficient estimates for $(\hat{\beta}_0)$ from the true, simple, assumed and combined models, with data generated from equation (32) with long memory error structure. We consider 4 possible tail covariates. The first two columns have estimates for the assumed tail of length 5 observations, the 3rd and 4th columns have assumed tail of 10 observations, which is equal to the true tail length. The 5th and 6th columns show the estimates for the assumed model with a tail length set to 15 observations. The final two columns show the estimates for an assumed tail of length 20 observations. We compare all four assumed tail lengths for the simple model (no tail covariate), the assumed model, the combined model and the true model where the tail length is known. UA is short for the unaggregated data and AGG is the aggregated data, at aggregation period J	56
4.7	Coefficient estimates for $\hat{\beta}_1$ and $\hat{\beta}_2$ from models 1-4 from data generated from equation (32) with long memory error structure. We consider 4 possible tail covariates. The first two columns have estimates for the assumed tail of length 5 observations, the 3rd and 4th columns have assumed tail of 10 observations, which is equal to the true tail length. The 5th and 6th columns show the estimates for the assumed model with a tail length set to 15 observations. The final two columns show the estimates for an assumed tail of length 20 observations. We compare all four assumed tail lengths for the simple model (no tail covariate), the assumed model, the combined model and the true model where the tail length is known. UA is short for the unaggregated data and AGG is the aggregated data, at aggregation period J	58
4.8	Coefficient estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ and $\hat{\beta}_6)$ for the assumed model with 4 minute tails for the coal train data. We consider outcomes for the aggregation periods from 5 to 10 minutes.	68
4.9	Coefficient estimates $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ and $\hat{\beta}_6)$ and standard errors for the assumed model with 4 minute tails for the coal train data. We consider outcomes for the aggregation periods from 5 to 10 minutes.	68
4.10	Comparison of train coefficient estimates under the combined model with 4 minute tails.	69
4.11	Comparison of standard errors for the train coefficient estimates, from table 4.10 above, under the combined model with 4 minute tails.	70
5.1	Fixed effects parameter selection for data generation all models	90
5.2	Parameter selection for data generation Model 2	91
5.3	Parameter selection for data generation Model 3	94
5.4	Parameter selection for data generation Model 4	96
5.5	Parameter selection for data generation Model 5	98
5.6	Parameter selection for data generation Options 1 and 2	102
5.7	Parameter selection for data generation Option 1	102
5.8	Parameter selection for data generation Option 2	106
6.1	Simulation results for iid error structure. Values shown are the means of 20 replicates. The simulated data has length $N = 5000$, and for aggregated data, we aggregated every $J = 10$ observations. The number of subsets, K , for the D&R models is $K = 2$ and $K = 5$	125
6.2	Simulation results for long memory error structure. Values shown are the means of 20 replicates. The simulated data has length $N = 5000$, and for aggregated data, we aggregated every $J = 10$ observations. The number of subsets, K , for the D&R models is $K = 2$ and $K = 5$	126

6.3	D&R results. Unaggregated Data.	129
6.4	D&R results. Aggregation period of $J = 1$ minute.	132
6.5	D&R results. Aggregation period of $J = 5$ minutes.	134
6.6	Combined Model results for unaggregated to aggregated data of 1 to 5 minutes. Tail length is 4 minutes.	140
6.7	Comparison of full data (Chapter 4) and D&R (Chapter 6) results: 5 minute aggregation	141
6.8	Comparison of full data (Chapter 4) and D&R (Chapter 6) results: 5 minute aggregation. Combined model.	142
6.9	Comparison of full data (Chapter 4) and D&R (Chapter 6) analyses timings. Calculation time is in minutes. (*) here we have divided each day into two subsets due to the memory constraints.	143

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 its 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.