Developing Advanced Methods to Predict Air Traffic Network Growth

Judit Guimera Busquets

PhD Dissertation

CITY
UNIVERSITY OF LONDON
EST 1894

City, University of London

School of Mathematics, Computer Science & Engineering

October 2019

Contents

List of Figures vii

List of Tables xiii

Acknowledgements xvii

Declaration xix

Abstract xxi

Nomenclature xxiii

1 Background and motivation 1

1.1 Introduction 1
1.2 Importance of air travel forecasts 2
1.2.1 The role of air travel forecasts 3
1.2.2 Expected aviation growth 4
1.3 Motivation of the thesis 4
1.3.1 Airport connectivity 5
1.3.2 Air itinerary choice modelling 7
1.4 Outline of the thesis 9

2 Forecasting Models in the Literature Review 11

2.1 Classification of forecasting methodologies in aviation 11
2.1.1 Forecasting approaches used in aviation 13
2.1.2 Discussion 18
2.2 Research work in air travel demand forecasting 18
2.2.1 Machine learning within aviation 19
2.2.2 Machine learning in air travel demand forecasting 20
2.2.3 Application of network theory within aviation 22

2.2.4 Existing research in air itinerary choice models 27
2.2.5 Discussion 30

3 Modelling Framework 33
3.1 Scope 35
3.1.1 Geographic scope 35
3.1.2 Types of flights 35
3.1.3 Types of passenger 36
3.1.4 Aggregation level 36
3.2 Data sources 37
3.3 Methodology Overview 39
3.4 Module 1: O&D passenger demand model 41
3.4.1 O&D demand models estimated results 47
3.4.2 O&D demand models validation 50
3.5 Module 2: Airport connectivity model and itinerary choice model . . 58
3.6 Module 2: Airport connectivity model 58
3.6.1 Network Theory metrics to study airport connectivity within the US 59
3.6.2 Example application of network theory to characterise the US Air Transport System 63
3.6.3 Aviation-related variables to study airport connectivity within the US 68
3.6.4 Airport connectivity model specification 69
3.6.5 Airport connectivity model validation 78
3.7 Module 2: Itinerary choice model using multinomial logit 85
3.7.1 MNL model specification 85
3.7.2 MNL model results 94
3.7.3 MNL model validation 97
3.8 Module 2: Itinerary choice model using neural networks 99
3.8.1 NN model specification 101
3.8.2 NN model results and validation 105
3.9 Module 2: Comparison of the multinomial logit model and the neural network model 109
3.10 Module 3: Flight frequency model 111

3.10.1 Model specification: Model 1-2SLS 112
3.10.2 Model specification: Model 2-OLS 117
3.10.3 Model validation: Model 1-2SLS and Model 2-OLS 119
3.11 Summary 121

4 Example application of the modelling framework: US Air Transportation System 125

4.1 O&D passenger demand model 127
4.2 Airport connectivity model 138
4.3 Case 1: Evolution of the US network not considered 144
4.3.1 Case 1 - Itinerary choice model: multinomial logit model . . 144
4.3.2 Case 1 - Itinerary choice model: neural network 146
4.3.3 Case 1 - Itinerary choice model: comparison 147
4.3.4 Case 1 - Air traffic levels model: Model 2-OLS 150
4.3.5 Case 1 - Air traffic levels model: Model 1-2SLS 153
4.4 Case 2: Evolution of the US network considered 156
4.4.1 Case 2 - Itinerary choice model: multinomial logit model . . 156
4.4.2 Case 2 - Itinerary choice model: neural network model 158
4.4.3 Case 2 - Itinerary choice model: comparison 160
4.4.4 Case 2 - Air traffic levels: considering Airport-Pair connectivity 162
4.5 Comparison between Case 1 and Case 2 165

5 Conclusions and future work 169

5.1 Overall achievements compared to objectives 169
5.2 Modelling framework 171
5.2.1 O&D demand model 171
5.2.2 Airport connectivity model 171
5.2.3 Itinerary choice model 172
5.2.4 Air traffic model 173
5.3 Projections 174
5.4 Suggestions for further work 175
5.4.1 Airport connectivity 175
5.4.2 Itinerary choice modelling 176
5.4.3 Modelling framework 176

A Socio-economic data sources **179**

B Value of time and CPI calculation **185**

B.1 Value of Travel Time 185

B.2 Consumer Price Index 188

C Previous attempts - O&D demand model **191**

C.1 Previous modelling specifications 191

C.2 Error Investigation 192

D Previous attempts - Airport connectivity **199**

D.1 Airport connectivity models including O&D passenger demand . . . 200

D.2 Link additions and removals in 2011 and 2012 206

E Previous attempts - Itinerary choice model **209**

E.1 Multinomial logit model 209

E.2 Neural network model 215

List of Figures

2.1 Flowchart on Kotegawa's ATS capacity network model 26

3.1 Modelling framework. 34

3.2 Distance distribution amongst city-pairs considered in this research. 43

3.3 Observed against predicted passenger demand for short-haul city-pairs. 51

3.4 Observed against predicted passenger demand for medium-haul city-pairs. 52

3.5 Observed against predicted passenger demand for long-haul city-pairs. 52

3.6 Observed and predicted total number of passengers throughout validation years (i.e. 2008-2013) for short-haul city-pairs 53

3.7 Observed and predicted total number of passengers throughout validation years (i.e. 2008-2013) for medium-haul city-pairs 53

3.8 Observed and predicted total number of passengers throughout validation years (i.e. 2008-2013) for long-haul city-pairs 54

3.9 Weighted average fare by distance group at the market level. 54

3.10 Observed and predicted number of passengers throughout validation years (i.e. 2008-2013) for all city-pairs considered in this study. . . . 56

3.11 Observed and predicted system-wide total number of passengers throughout validation years (i.e. 2008-2013). 56

3.12 Average and weighted average fares for the US domestic network from 2007 to 2013. 57

3.13 Degree distribution. 64

3.14 Node weight (s) against node degree (k) 65

3.15 Eigenvector centrality (EVC) against node degree (k). 65

3.16 Clustering coefficient against node degree for all the nodes considered in this study. 66

3.17 Adjacent matrix of AMA airport with its neighbouring nodes. 67

3.18 AMA airport's network. Size of a node is representative of the node degree. 67
3.19 Distribution of number of links connected the previous year for those markets that experienced a link addition. 80
3.20 Distribution of number of links connected the previous year for those markets that experienced a link removal. 81
3.21 Actual and predicted network degree evolution between 2008 and 2013. 84
3.22 Actual and predicted network eigenvector centrality evolution between 2008 and 2013. 85
3.23 Journey fare distribution across the feasible itineraries based on the connecting airport. 93
3.24 Journey time distribution across the feasible itineraries based on the connecting airport. 94
3.25 Observed against predicted number of passengers throughout the validation years. 99
3.26 Example of a Neural Network. 100
3.27 An artificial neuron. 101
3.28 Comparison of average training time for all neural network topologies. 106
3.29 Comparison of average number of epochs during training for all neural network topologies. 107
3.30 Comparison of average Mean Square Error (MSE) of the validation dataset for all neural network topologies. 107
3.31 Observed against predicted number of passengers throughout the validation years using the estimated neural network model i.e. *NN30/10.* 108
3.32 RPK computed for the validation years (2008-2013) based on validation results obtained from the MNL and the NN model. Actual RPK values are also included. 110
3.33 Predicted and actual proportion of non-stop passengers throughout the validation years (2008-2013). 111
3.34 Comparison between observed and predicted air traffic levels across validation years. 120
3.35 Comparison between observed and predicted air traffic levels across validation years for model OLS-3. 121

4.1 Population projections. 129
4.2 Mean household income projections. 129
4.3 Oil fossil fuel price actuals and projections. 130
4.4 Comparison between airfare weighted mean and oil fossil fuel price between 2007 and 2013. 131
4.5 Mean airfare projections for low, central and high scenarios. Actual mean airfare up to 2015. 132
4.6 Projected O&D passenger demand for city-pairs that are between 186 and 400 miles apart for years between 2008 and 2025. Actual passenger demand levels up to 2018. 133
4.7 Projected O&D passenger demand for city-pairs that are between 400 and 2113 miles apart for years between 2008 and 2025. Actual passenger demand levels up to 2018. 134
4.8 Projected O&D passenger demand for city-pairs that are more than 2113 miles apart years between 2008 and 2025. Actual passenger demand levels up to 2018. 134
4.9 Projected O&D passenger demand for the US domestic network. Actual values up to 2018. 136
4.10 Observed against predicted O&D passenger demand for the US domestic network for years 2008, 2010, 2012, 2014, 2016 and 2018. . . . 137
4.11 Projections for the US Air Transportation average network degree between 2008 and 2025. Actuals for years between 2008 to 2018. . . 142
4.12 Projections for the US Air Transportation network average eigenvector centrality between 2008 and 2025. Actuals for years between 2008 to 2018. 143
4.13 Projections for the US Air Transportation network average path length between 2008 and 2025. Actuals for years between 2008 to 2018. . . 143
4.14 Itinerary fares actuals and projections. 145
4.15 RPK actuals and projection values. 146
4.16 RPK actuals and projection values. 147
4.17 RPK actuals and projection values for both models: multinomial logit and NN. 149

4.18 Comparison of percentage proportion of non-stop passengers obtained from the MNL and NN models up to 2025. And actual proportion of non-stop passengers up to 2018. 149
4.19 Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the MNL model. Model used: Model 2-OLS. 151
4.20 Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the NN model. Model used: Model 2-OLS. 151
4.21 Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN. Actual traffic levels up to 2018. Model used: Model 2-OLS. 152
4.22 Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the MNL model. Model used: Model 1-2SLS with auto-regressive term. 154
4.23 Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the NN model. Model used: Model 1-2SLS with auto-regressive term. . . 154
4.24 Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN; and using Model 1-2SLS with auto-regressive term. Actual traffic levels up to 2018. 155
4.25 Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN; and using Model 1-2SLS with auto-regressive term. Actual traffic levels up to 2018. 155
4.26 RPK actuals and projection values when applying the MNL model and considering evolution of the US ATS. Model used: MNL. 158
4.27 RPK actuals and projection values when applying the NN model and considering evolution of the US ATS. 159
4.28 RPK actuals and projection values comparison for both models (MNL and NN) when considering evolution of the US ATS. 161

4.29 Comparison of percentage proportion of non-stop passengers obtained from the MNL and NN models up to 2025. And actual proportion of non-stop passengers up to 2018. Case when considering the evolution of the US ATS. 161

4.30 Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the MNL model. Case when considering the evolution of the US ATS. 163

4.31 Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the NN model. Case when considering the evolution of the US ATS. . . 163

4.32 Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN. Actual traffic levels up to 2018. Case when considering the evolution of the US ATS. 164

4.33 Average LF over time obtained up to 2025 for the 2 sets of results: MNL and NN. Actual traffic levels up to 2018. Case when considering the evolution of the US ATS. 164

4.34 RPK projections for the Central scenario and considering no-evolution (Case 1) and evolution (Case 2) of the network. 166

4.35 Proportion of non-stop passengers projections for the Central scenario and considering no-evolution (Case 1) and evolution (Case 2) of the network. 168

4.36 Air traffic levels projections for the Central scenario and considering no-evolution (Case 1) and evolution (Case 2) of the network. 168

C.1 Error distribution throughout the validation years, closed up to error values between -100,000 and +100,000. 193

C.2 Total error against mean passenger demand per bin. Bin size 10,000. Range years 2008-2013. 196

C.3 Total error against mean passenger demand per bin for O&Ds with less than 20,000 passenger demand. Bin size 500. Range years 2008-2013. 197

D.1 Links that were added in 2011 (airport-pairs that were unconnected in 2010). 206
D.2 Links that were added in 2012 (airport-pairs that were unconnected in 2011). 206
D.3 Links that were removed in 2011 (airport-pairs that were connected in 2010). 207
D.4 Links that were removed in 2012 (airport-pairs that were connected in 2011). 207

E.1 Observed against predicted number of passengers using model 1 throughout the validation years. 212
E.2 Observed against predicted number of passengers using model 2 throughout the validation years. 213
E.3 Observed against predicted number of passengers using model 3 throughout the validation years. 213
E.4 VOT calculation from model 1. 215
E.5 Validation performance comparison (MSE) between the best performance results for several NN architectures with 3 different training algorithms. 216
E.6 Caption for 217
E.7 Time comparison between the best performance results for several NN architectures with 3 different training algorithms. 217
E.8 Validation performance comparison (MSE) between the best performance run considering four activation function combinations from table E.5. 218
E.9 Number of eqpochs comparison between the best performance run considering four activation function combinations from table E.5. . . 219
E.10 Time comparison between the best performance run considering four activation function combinations from table E.5. 219

List of Tables

3.1 Distance groups considered in O&D demand models. 42

3.2 Dummy variables considered for each O&D demand model. 45

3.3 Coefficient estimates obtained for the O&D demand models by distance group. 48

3.4 Percentage error between predicted and observed total passenger demand for the different distance groups and at the aggregate level throughout the validation years (i.e. 2008-2013). 55

3.5 Adjusted R^2 obtained across the validation years (2008-2013) when considering the aggregated results from the 3 O&D models developed by distance group. For comparison, adjusted R^2 from a previous attempt is also included. 57

3.6 Airport-pairs with AMA as origin airport. 67

3.7 Airport-pair connectivity between 2007 and 2013. An airport-pair is considered connected if there are at least 52 flights operating between them[1]. 69

3.8 Connectivity grid: removed links against added links. Number of airport-pairs removed against number of airport-pairs added to the network. Numbers are computed across the range years 2007 - 2013. 70

3.9 Distribution of connected years for those links that change their connectivity status twice between 2007 and 2013 -i.e. distribution of the 505 airport-pairs from Table 3.8-. 71

3.10 Airport-pair connectivity summary after excluding the 345 airport-pairs due to their unstable connectivity. Mean values across years 2007-2013. 72

3.11 Estimated model results for the link removal model. 75

3.12 Estimated model results for the link addition model. 77

3.13 Performance metric summary from the validation of the estimated link removal model. 82
3.14 Performance metric summary from the validation of the estimated link addition model. 83
3.15 Specification table of the utility function for stage two of the 2SCF model estimated using Berkson-Theil method. 91
3.16 Key characteristics of the input dataset. 92
3.17 Estimation results for the 1st stage of the control function model aim to predict journey fares. 94
3.18 Estimation results for the second stage of the control function model, when applying Berkson-Theil method. 96
3.19 Comparison of Value of Time obtained for the itinerary share model using 2SCF model and using WLS regression 98
3.20 Neural Network topologies considered during the estimation process of an air itinerary choice model using artificial neural networks. . . . 103
3.21 Comparison of adjusted R^2 obtained for each of the validation years when using the MNL model and the NN model. 110
3.22 1st stage model specification. 114
3.23 2nd stage model specification. 114
3.24 Coefficient estimates obtained for the first stage of the Model 1-2SLS. 116
3.25 Coefficient estimates obtained for the second stage of the Model 1-2SLS.116
3.26 Coefficient estimates obtained for the Model 2-OLS. 119
3.27 Comparison of adjusted R^2 obtained for each of the validation years when using the Model 1-2SLS and the Model 2-OLS. 120
3.28 MSE obtained from applying Model 1-2SLS and Model 2-OLS to those cases were a new link is added into the network. 121

4.1 Set of projections presented in this Chapter. 127
4.2 Rate of change year on year comparison between oil fossil fuel and average airfares. 132
4.3 Difference between projected and observed total passenger demand by distance group for years between 2008 and 2018. 133
4.4 Difference between projected and observed total passenger demand for the US domestic network for years 2008-2018. 136

4.5 Airport-pair connectivity between 2014 and 2018. An airport-pair is considered connected if there are at least 52 flights operating between them. 140
4.6 Observed number of airport-pair connected and disconnected against Predicted number of airport-pair connencted and disconnected for years 2008, 2010, 2012, 2014, 2016 and 2018. 141
4.7 RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using multinomial logit. . . 146
4.8 RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using neural networks. . . . 147
4.9 Adjusted R^2 computed from the projections generated from both models up to 2018. Values referred to the Central scenario. 148
4.10 Air traffic levels projections for 2008, 2018 and 2025 for the two sets of results and scenarios. Actuals for 2008 and 2018. 152
4.11 Air traffic levels projections for 2008, 2018 and 2025 for the two sets of results and scenarios; when using Model 1-2SLS with auto-regressive term. Actuals for 2008 and 2018. 156
4.12 RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using multinomial logit and considering evolution of the US ATS. 158
4.13 RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using neural networks and considering evolution of the US ATS. Model used: NN. 159
4.14 Adjusted R^2 computed from the projections generated from both models up to 2018 when considering network evolution (Case 2). Values referred to the Central scenario. 160
4.15 Air traffic levels projections for 2008, 2018 and 2025 for the two sets of results and scenarios. Actuals for 2008 and 2018. 165

B.1 Business and personal/leisure travel weights used to calculate value of time. 186
B.2 Value of business and personal travel time for years between 2003 and 2025. 188

B.3 $Factor_{CPI}$. calculation in order to convert monetary data to 2007's dollars. 190

C.1 1st attempts: estimated coefficients obtained when using OLS and WLS as estimation process. Error refers to the average difference of predicted total network passenger demand against observed during the validation years (2008-2013). 192

D.1 Model specifications used during early attempts to model airport connectivity. 203

D.2 Confusion matrix obtained during validation. Average values across all validation years: 2008-2013. 204

D.3 Performance metric for all 5 estimated models when validating the models over data from 2008 to 2013. Values shown are average across the 6 validation years. 205

E.1 Specification table of the utility function for models estimated using Berkson-Theil method. 211

E.2 Table results for Berkson-Theil. 212

E.3 Comparison of predictive powers between the 6 models estimated using Berkson-Theil method. 212

E.4 Comparison of Value of Time. Note the average VOT values for model 1 are presented in this table. 214

E.5 Combination of activation functions considered in this research. . . . 218

Acknowledgements

I would like to express my gratitude to all the people who supported and encouraged me during the journey to my PhD. First of all, I would like to thank my supervisors, Prof. Chris Atkin and Dr. Eduardo Alonso, for giving me the opportunity to do my PhD at City, University of London. Thanks Chris, for providing guidance (and no filter) during the project.

I would like to express my special thanks to Dr. Antony Evans, who has guided me and supported me during all my PhD years, I really appreciate all your continuous support and advice. And I only have words of gratitude to Dr. Lynnette Dray for her kind and valuable help and for offering constant support, sharing her knowledge and providing guidance during this last year. I have learnt a lot from you both and I am really grateful.

I would like to thank Isabella, Weimiao and Evelien, fellow PhD students. Weimiao, I still remember when we met, two "freshly-newly" PhD students that did not really know what to expect. Together with Weimiao, Isabella and Evelien thanks for coming along in our adventure as part of Bolleboos (still think Pindakaas could have been a good team name too!). AFYI was an experience I would never forget! And "here WEGO" to a new adventure! My PhD journey would not have been the same without any of you. Having each other all these years made the tough moments easier to take along. I do believe the PhD has brought a friendship that would be impossible to break.

Also, thanks to all the PhD family in CG41 who made the PhD life easier and funnier. And a really special thanks to Alberto (sai che ti voglio bene). You have made a difference in my life at City and you know I am and will be here for anything you need. Thanks also to the entire aeronautical research group.

Thanks to all my London friends who have been supporting me during the tough moments of the PhD. I know it has not been easy, now there is no excuse to meet. For this, thanks to Corina, Fabrizio, Nick, Charlotte, Carmen, Enrique, Mariana, Gabriel and especial mention to Laura "sempre na torcida ainda seja des de lonje". Also, thanks to those that are far away but always supported me (Cesar, Laura, Jordi, MJ, tuda a tchurma brasileira and many more that cannot fit in just one page).

I would like to express my gratitude to my parents, who have been a strong support. Nothing would have been possible without them and their patience. Also, thanks to my brother that for once in a lifetime I let him finish something first and became a Dr ahead of me! Thanks to Josie (Dr. Golding) too, who is incredible knowledgeable and has helped me a lot. I have learned a lot from you, and I am really happy you are part of the family. And thanks to my family-in-law who always believed in me. Special mention to 'hermano' Renato for being close to me (and Marcelo) always, even if it is through a screen.

And finally, 'muito obrigada' to Marcelo. I would not be at this stage without you. You have been my support during the entire journey and encouraged me to keep going even during tough times. I am sorry your endless weekends watching football with a beer on hand are coming to an end. Thanks for all your support, your guidance, your most needed hugs and your unconditional love.

Author's Declaration

I declare that the work carried out in this thesis at City, University of London for a research degree between September 2013 and October 2019 is original. I also certify that I have written all text herein and have clearly indicated by suitable citation any part of the dissertation that has already appeared in publication (Guimera Busquets, 2015; Guimera Busquets et al., 2015; Busquets et al., 2018).

Judit Guimerà Busquets

October 2019

Abstract

This dissertation describes a forecasting methodology that takes into account changes in the connectivity of an air transportation system and assesses the impact at other levels of the network, such as route demand and air traffic levels. To achieve this, the modelling framework looks at city-pair demand generation, route demand assignment and air traffic estimation. While generating air traffic forecasts, the resulting model is also intended to highlight the most important factors driving air traffic network growth. This is achieved by considering a larger set of drivers than those considered in existing methodologies and research as well as exploring the use of alternative modelling techniques.

Network evolution is incorporated in the method through an airport connectivity model which identifies how and when airport-pairs across the network change their connectivity status. The problem is split into two models: one identifying those airport-pairs that are added to the network; and another one identifying those airport-pairs that are removed from the network. The modelling approach explores the use of network theory metrics along with other input variables, such as passenger demand, to see whether existing models employing only network theory metrics could be improved.

The impact of network evolution is assessed by the effect on air itinerary shares. Two itinerary choice models are developed using two different modelling approaches: multinomial logit and neural networks. While the multinomial logit formulation is the most common approach used to model itinerary shares, only few studies have looked at modelling itinerary shares at the network level. Neural networks have yet to be explored in this field. In this research, air itinerary choice models have been developed at the most aggregate level, using open-source booking data, for a large group of city-pairs within the US Air Transportation System. The output of the itinerary choice models, influenced by the consideration of network evolution, is then used to project air traffic levels and assess the impact of network structure changes in the number of operations in the US ATS.

The results reflect the complexity behind network evolution, especially for cases when a mature system is considered (e.g. US ATS): comparisons between the case of a static network and the case when network evolution is considered indicate that the impact of network changes on overall system metrics is relatively minor in the US. However, they indicate that changes in fossil fuel prices may influence changes in the overall network characteristics, and consequently network evolution. The results also prove the feasibility of estimating a single itinerary choice model at the network level for an entire air transportation system. Although the multinomial logit model results have better accuracy, the potential of neural networks for this purpose is also demonstrated, the latter being more representative of the hub-and-spoke network strategy.

Nomenclature

Acronyms

2-SCF two-stage control function

2-SLS two-stage least squares

AI Artificial Intelligence

AIM Aviation Integrated Model

ANN Artificial Neural Network

APD Air Passenger Duty

ASC Alternative Specific Constant

ASPM Aviation System Performance Metrics

ATAG Air Transport Action Group

ATM Air Transport Movement

ATS Air Transportation System

BEA Bureau of Economic Analysis

BLS Bureau of Labour Statistics

BNN Biological Neural Network

BTS Bureau of Transport Statistics

CAA UK Civil Aviation Authority

CC Clustering Coefficient

CPI Consumer Price Index

IATA airport codes have been excluded from the nomenclature, however they can be found in https://www.iata.org/publications/pages/code-search.aspx

DB1B Airline Origin and Destination Survey

DECC Department of Energy & Climate Change

DfT UK's Department for Transport

DoT US's Department of Transportation

EVC Eigenvector Centrality

FAA Federal Aviation Administration

FLEET Fleet Level Environmental Evaluation Tool

FNR False Negative Rate

FPR False Positive Rate

GDP Gross Domestic Product

IATA International Air Transport Association

ICAO International Civil Aviation Organization

IV Instrumental Variable

LCC Low Cost Carrier

LF Load Factor

LMA Levenberg-Marquart algorithm

MLE Maximum Likelihood Estimation

MMNL Mixed Multinomial logit

MNL Multinomial logit

MSA Metropolitan Statistical Areas

mSA micropolitan Statistical Areas

MSE Mean Squared Error

NAPAM National Air Passenger Allocation Model

NAPDM National Air Passenger Demand Model

NextGEN Next Generation Air Transportation System

NHTS National Household Travel Survey

NN Neural Network

O&D	Origin & Destination
OLS	Ordinary Least Squares
QSI	Quality of Service Index
RF	Random Forest
RP	Revealed preference data
RPK	Revenue Passenger Kilometers
RPM	Revenue Passenger Miles
SARPs	Standards and Recommendations Practices
SP	Stated preference data
SVM	Support Vector Machine
TAF	Terminal Area Forecasts
TNR	True Negative Rate
TPR	True Positive Rate
VIF	Variance Inflation Factor
VOT	Value of Time
WLS	Weighted Least Squares
YoY	Year-on-Year

Chapter 1

Background and motivation

1.1 Introduction

Just over a century has passed since the first scheduled flight in history took place between St. Petersburg and Tampa in the United States. Although the event was an important milestone in aviation history, commercial aviation was not of interest to the public. It was not until the World War I (WWI) when the military value and strategic advantage of aircraft were identified, leading to technological advances in the aviation sector. Thanks to these technological innovations, the first freight services were born. However, people remained sceptical of the idea of travelling by air, based on a combination of safety concerns and trust towards the service provided. Larger, faster and safer aircraft were needed to change people's aviation vision. It was not until the 1930s when the aviation sector started to grow. Important aircraft innovations led to the creation of the first modern airlines focused on the transport of passengers, such as United Air Lines or Transcontinental and Western Air (TWA) among others.

Advances in aviation technology during war time years along with governments' post-war support helped to make the aircraft a leading mode of transport during the 1950s. In the late 1970s, deregulation opened up the aviation market in the Unites States generating room for the expansion of the industry. Unconstrained competition, freedom to expand the route system and flexibility to develop innovative pricing structures became the new norm with deregulation. As a result, new airline business models came to the market, prompting a rapid aviation growth since the liberalisation of the skies in the late 1970s. In the last year of government regulation (1977) in the US, the airline industry carried approximately 240 million passengers in scheduled service, while in 2016 they carried 823 million passengers system wide (BTS, 2017) against an economic growth (measured in real GDP[1]) of 184% during the same period (BEA, 2019).

With the aim to accommodate this increase in demand capacity expansion was accompanied by investing in new infrastructures (i.e. such as expansion of airports), and new technological improvements for aircraft, air management systems and air traffic systems.

The fast growth and expansion experienced by the aviation industry over the years show how important this industry has become to society. Its influence on the worldwide economy and connectivity has led aviation to be one of the key driving factors of today's society and economy. Globally, aviation has an economic impact of $2.7 trillion equivalent to 3.6% of the global Gross Domestic Product (GDP) and supports 62.7 million jobs worldwide [2]

1.2 Importance of air travel forecasts

The air transportation industry is a large, complex sector involving numerous stakeholders with different agendas all playing an important role. In order to maintain the welfare and prosperity of the industry, it is essential to understand the behaviour of what drives the industry's demand, which involves both passengers and freight. Consequently, not only the trends of passengers and freight of past years need to be studied, but also the intrinsic reasons that explain such behaviour need to be understood. This knowledge is used by stakeholders to produce air travel demand forecasts, permitting them to plan ahead in an effort to maintain the well-being of the industry. As Schäfer (2007) states, anticipating changes in air travel demand on aggregated levels is critical for all the major players of the industry. Failing to do so may end up not only in high economic losses but also in slowing down the society and economic development of a region.

[1] Real GDP is a measurement of economic output that accounts for the effect of inflation or deflation giving a more realistic assessment of growth than nominal GDP.

[2] These numbers are referred to 2016 according to ATAG statistic and include direct, indirect, induced and tourism catalytic. Source: https://aviationbenefits.org/media/166344/abbb18_full-report_web.pdf

Unexpected disruptions are another major concern for aviation if supply is not ready for it, involving also major economic losses. For example, in 2010 the eruption of the Icelandic volcano Eyjafjallajökull took Europe by surprise, leaving the air traffic over Europe in a state of total collapse. This natural adversity caused the closure of large portions of European air space for a week due to safety concerns related to the effect that ashes blown over the European air space could have on aircraft engines. According to the Air Transport Action Group (ATAG) (2014) not being able to supply the air travel demand for about a week meant that around 10 million passengers were affected costing the global GDP about £3bn ($5bn).

Closer to home, the controversy that the UK air transport sector generates, especially in the South East of England, is often broadcast in the news. Experts in the aviation sector believe that if nothing is done to improve London airports' capacity soon, the current level will not be able to meet the UK's air travel demand forecast for the medium- and long-term periods (Airports Commission, 2013). The consequences of not being able to meet the requirements of air travel demand in the future are believed to impact directly on the high economic and social competitiveness of London, one of the main gateway cities of the UK and Europe. For instance, the Airports Commission (2013) suggests that failing to address the situation in the South East of England would cost, over a sixty-year time period, between £21 to £23 billion to users and providers of airport infrastructure and between £30 to £45 billion to the wider economy (Airports Commission, 2013).

1.2.1 The role of air travel forecasts

The aim of forecasting is to determine how patterns of demand will change over time, reflecting external factors such as growth of income, demographic changes and changes in transport prices. The use and scope of aviation forecasts can vary. For example, airport developers will use air passenger demand forecasts to make decisions regarding infrastructure expansion plans; whereas airlines will use them when planning their network and fleet management taking into account demand seasonality.

Air travel forecasts are also essential for governments and regulatory bodies, who generate them to help take informed decisions when establishing policies. Those policies are not only focused on maintaining the well-being of all industry stakeholders and to achieve a balance between industry and society, such as policies created to reduce the environmental impact of aviation. Forecasting is therefore a key tool for decision-making and is used across all industry players and Government in both business planning and policy decision making.

1.2.2 Expected aviation growth

Although forecasts differ in scope and purpose by which they are produced, they all agree in predicting significant aviation growth. For example, the two main manufacturers, Airbus and Boeing, have similar projections, forecasting a global annual air traffic growth of just under 5% over the next 20 years. Airbus (Airbus, 2017) predicts that average annual traffic growth (measured in Revenue Passenger Kilometre, RPK[3]) over the next 20 years will be 4.4% with a more rapid increase during the first decade (4.9% growth p.a.) than the second (4.1% growth p.a.) (Airbus, 2017). Similar to Airbus, Boeing predicts an annual world passenger growth of 4.7% over the next 20 years (Boeing, 2017).

According to statistics from the International Civil Aviation Organitzation (ICAO), the aviation industry has seen a dramatic growth over the past 20 years, with passengers numbers rising from 1.46 billion in 1998 to 3.98 billion in 2017[4]. Aviation growth prospect is high, with forecasts of number of passengers is expected to reach 8.2 billion in 2037 according to the International Air Transport Association (IATA)[5].

1.3 Motivation of the thesis

The literature review reveals the importance of forecasting within the air travel industry. It is a key tool for decision-making used across all industry stakeholders. Nowadays, issues to be addressed in air travel are much broader and complex than in the past, affecting other areas outside air travel; today's decision makers need to consider not only the impact of aviation growth towards economy, but also social and environmental effects, a wider range of options alternatively to building new

[3] RPK is the basic measure of passenger traffic and reflects how many available seats were actually sold. RPK is defined as the product between number of passengers and kilometres they have flown.

[4] Source: https://data.worldbank.org/indicator/IS.AIR.PSGR

[5] This number is based on a 3.5% compound annual growth rate (CAGR). Source: https://www.airlines.iata.org/news/passenger-numbers-to-hit-82bn-by-2037-iata-report

infrastructures and resource limitations.

The constant evolution of the industry and increased complexity require continuous work on forecasting methodologies for air transportation. While significant contributions have been made to date, existing research are typically based on the assumption that the air transport system is static and therefore routes are neither added or removed to/from the network. While network evolution is only considered when airline information is available, this information is limited to the short-term since airlines do not plan their network structure far ahead, and often network changes information is not available to everyone. Considering the fact that air transport network evolution occurs - e.g. in 2016, 20% of the schedule seats were on routes that did not exist in 2000 (SABRE database, 2017) - this research focuses specifically on modelling network evolution and the impact that this evolution might have at different levels of the air transport system such as number of passengers per kilometre, proportion of non-stop passengers or air traffic levels. By considering network evolution, the impact of network changes to airport congestion, environment and local and global emissions could be better understood. In order to do that, this research explores two areas within aviation forecasting that are briefly introduced below: airport connectivity; and air itinerary choice modelling approaches. Both areas of research are presented within a single modelling framework that looks at city-pair demand generation, itinerary demand assignment and air traffic estimation, so that the impact of network evolution can be analysed at different levels across the network (i.e. how it affects passengers' choice of route and air traffic levels).

1.3.1 Airport connectivity

To study the evolution of the network, one requirement is to predict whether airport connectivity will develop, so if two airports will have a direct flight between them in the future. Airport connectivity is investigated through the use of complex network theory. Network theory is a way of representing networks through studying the components that form them and the interactions and connectivity amongst these components, since those are what shape and define the network as a whole[6]. Therefore, network theory is used to form a better understanding of the structure and dynamics of many real-world systems, such as the air transportation system,

[6]Further information regarding network theory is explained in Chapter 3.

by characterising the system's components and their interactions through a set of mathematical metrics.

Initial studies using network theory looked at characterising the air transport network through a set of mathematical metrics and understanding its characteristics (Guimera et al., 2005; Cheung and Gunes, 2012). Research done by Cheung and Gunes (2012) showed that in general the US Air Transportation System (ATS) exhibits similar network characteristics to the World-wide Airport Network (WAN), which is characterised by showing a power-law distribution, meaning that is a network formed of large amount of small airports (i.e. with a low number of connections) and a small amount of large airports (i.e. with a large number of connections). Results from Cheung and Gunes' research highlighted that, similarly to the WAN characteristics, the US ATS showed just a partial power law degree distribution, by which it was suggested that within the US ATS highly densely populated areas grow at a slower rate than those located in less populated areas. Their work also showed that compared to earlier years, the US ATS is more vulnerable to airport closures today than it was in the past due to the number of airports increasing faster than the number of flight routes, prompting the network to become less dense, and therefore more susceptible to failures. More recent studies (Lacasa et al., 2009; Fleurquin et al., 2013) adopted network theory to explore and identify the dynamics taking place in the aviation system and how these affect the efficient functioning of the network, such as simulating congestion in the airspace (Lacasa et al., 2009), and the spread of epidemics (2006). The work done by Lacasa et al. (Lacasa et al., 2009) used network theory to simulate the dynamics of the network when the diffusion of a given number of aircraft through that network occurs so that congestion effects could be simulated. Results obtained from the simulation were compared to those obtained when applying the model to real network data from the European air transportation system, leading to similar results to those previously obtained: above a certain threshold, the amount of airport queues and operational efficiency sharply decreases due to congestion effects. The work done by Colizza et al. (2006) studied how the characteristics of the world-wide air transport network influences the global spread of diseases by combining network theory and the susceptible-infectious-recovered (SIR) model, which describes mathematically an influenza epidemic. Results showed that air transportation network properties are partly responsible for the global pattern of emerging diseases, with models having an 80% accuracy at predicting pathways of epidemic diffusion during the early stage of the epidemics. Being this stage the most relevant phase during epidemic surveillance, such models could be used as a tool to test the effect of plans of actions to avoid the spread of diseases such as travelling restrictions and vaccination policies. These examples outline the potential of applying network theory to understand the dynamics of an air transport system, and therefore, understand and characterise its evolution.

Although the application of network theory as a method to model the response of an existing air transport system to events has become a trend in the last decade, few have used this technique to predict the future structure of an ATS. The main research in this field is the work done by Kotegawa (2012), who applied network theory to model the US ATS to analyse the evolution of airport connectivity. Kotegawa's research looked at the likelihood of airport-pairs changing their connectivity - i.e. identification of which new airport-pairs would appear in the future and which of the existing ones would disappear - by applying different modelling techniques in which all input variables were only network theory metrics. The accuracy of these early models ranged between 20% and 40%, leaving room for improvement and further exploration of the subject. Inspired from Kotegawa's results (2012), one of the main motivations of this thesis is to explore whether these accuracies can be improved, in particular by using a broader set of input variables beyond network theory metrics, such as considering passenger demand; as well as considering the effect of the US network evolution would have to other system-wide characteristics such as itinerary shares and air traffic levels.

1.3.2 Air itinerary choice modelling

Air itinerary choice modelling is a representation of the proportion of passengers choosing one itinerary out of several options - i.e. also called air itinerary share models. Air itinerary share models aim to predict customer behaviour, helping to understand what drives air passenger choices when it comes to travel. Itinerary choice models can become crucial to support airlines in their network planning and scheduling since important decisions on resources allocation and pricing are made based on itinerary demand. One of the most common air itinerary share model used across the industry are the Quality of Service Indices (QSI), which base their predictions on the relationship between itinerary share and frequency share as well as itinerary level of service (i.e. whether itinerary is a direct route or involves a connection). However, since QSI do not take into account other factors that might influence itinerary shares, such as itinerary fares and time or characteristics of the passenger (e.g. business/leisure passenger), there is a growing interest in developing better itinerary choice models; consequently, for the last 15 years efforts have been focused on shifting away from Quality of Service indices (QSI).

Most of the current research focused on predicting air itinerary shares centres around discrete choice models, which use logit formulation to calculate the probability of a passenger choosing a specific itinerary based on its utility value[7]. Discrete choice models are widely used in urban transportation, however, they are usually built using disaggregate data and include information about the individual making the decision (i.e. the passenger); whereas in air transport, data at the disaggregated level as well as data accessibility are limiting factors.

Most of the early studies on demand assignment for air travel focus on studying the distribution of demand across one single dimension. For example, analysis of air travellers' choice within multi-airport cities or regions (Hansen, 1995; Windle and Dresner, 1995) or across airlines (Proussaloglou and Koppelman, 1995). Although these research gave a deeper understanding of the relationship between airport attributes and airport market shares, a more aggregated assignment of air travel volumes is also needed. In recent years, efforts focus on modelling itinerary market share across multiple dimensions (Adler, 2001; Coldren et al., 2003; Grosche and Rothlauf, 2007; Atasoy and Bierlaire, 2012; Coldren and Koppelman, 2005; Hsiao and Hansen, 2011). And only one research has attempted to model air itinerary shares using a machine learning technique (Grosche and Rothlauf, 2007); the work done by Grosche and Rothlauf was a comparative study between multinomial logit, neural network and a custom model developed by the authors, with the latter being the most accurate.

There is a growing trend in using discrete choice methodology in the aviation industry and existing air itinerary share models are mostly focused on support-

[7]The utility value of an itinerary is calculated based on the characteristics of that itinerary. It is expected that a passenger would choose the itinerary with the largest utility.

ing carrier decision-making. Consequently, those studies define itineraries at a more disaggregate level, using variables describing airline and travel time preferences (Coldren et al., 2003; Coldren and Koppelman, 2005; Grosche and Rothlauf, 2007; Atasoy and Bierlaire, 2012). Also, data used in most of these studies is either proprietary or from surveys which tend to only represent a small subset of passengers through surveys and are time consuming and costly to complete, making its availability scarce. Lastly, the computing power needed to handle discrete choice modelling estimation is, in cases when the size of the dataset is large, a known limitation (Coldren et al., 2003; Li et al., 2017).

With the aim of modelling demand at the network level -i.e. annual share of passengers per each itinerary available in the entire network -, and therefore focusing on developing a tool that models the network dynamics as a whole rather than carriers share as current research, another main motivation of the present work is to develop an effective air itinerary share model which uses aggregated data, such as open-source booking data, for a large group of city-pairs within the US Air Transportation System. This is done through the use of alternative techniques: a discrete choice model and the use of an artificial neural network model, with the latter being a line of research not seen in the literature at this level of aggregation and without considering passenger preference information. Since the modelling framework presented in this dissertation considers the evolution of the network, the development of a model at the network level will allow to assess the overall impact that evolution has to the hub-and-spoke routing structure that characterises the US ATS and how the change of it may affect passengers. An itinerary share model at the system level can also capture the impact of airport capacity on passenger travel behaviour and airport congestion, and it can be used to evaluate the benefit of airport expansion projects. Lastly, by capturing the changes on travel behaviour that network evolution may influence, the impact of aviation has towards the environment can also be captured through the relationship between aircraft movements and emissions.

1.4 Outline of the thesis

This dissertation is structured as follows: Chapter 2 presents a literature review on existing forecasting methodologies that are of relevance regarding the work done in this dissertation; in particular the use of network theory used as modelling approach for evolution of the air transportation system and an overview on existing methodologies used to determine air itinerary shares.

The literature reveals the potential of applying network theory to understand the evolution of any air transportation system. From existing research, work has been done on analysing how connectivity between airports changes over time based on airport topology metrics. But this work has been limited to the use of network theory variables. This has given a motivation to the current investigation, since consideration of a broader set of variables to study airports' connectivity may improve existing modelling performance.

The research carried out in this dissertation is presented through a single modelling framework that looks at city-pair demand generation, route demand assignment and air traffic estimation. Modelling approaches used for each of these stages of the framework are described in Chapter 3. For each of the sub-models within the framework the estimation model results are presented and validated.

A predictive example application of the modelling framework is presented in Chapter 4. In this section mid-term projections for domestic US air traffic levels are produced for years from 2008 to 2025. Results are compared to actual data when possible (i.e up to 2018).

Conclusions drawn from the research carried out are presented in Chapter 5 with a section outlining suggestion for future work and possible improvements steps.

Chapter 2

Forecasting Models in the Literature Review

Considering the important role that forecasting plays within aviation and that is highlighted in Chapter 1, an overview of the existing methodologies used across the industry and published research is presented in the following sections. This Chapter is structured as follows. First, a classification of forecasting models most commonly used to forecast air travel demand is presented. This is followed by a summary of methodologies currently used by aviation stakeholders, such as aircraft manufacturers or international agencies. Finally, a more detailed review of published research that are of relevance regarding the work done in this dissertation is presented. The latter section includes an overview on the use of machine learning techniques within the topic of forecasting, a review on the use of network theory as a modelling approach for airport connectivity and an overview on methodologies used to determine air itinerary shares.

2.1 Classification of forecasting methodologies in aviation

Econometric modelling, which is the application of statistical methods to establish quantitative relationships between a particular phenomenon and the economic variables affecting it, is one of the most common approaches used in aviation forecasting. Econometric models establish the relationship between travel demand and income, which is one of the key variables explaining aviation growth. For example, Boeing states that economic and income growth are key drivers of travel demand, especially for emerging markets such as China and India (Boeing, 2017). Historical data is also used to identify air travel demand trends which then are used to generate future projections. The majority of these approaches are characterised by their simplicity and by the use of similar explanatory variables, such as socio-economic information or airfares, with these often chosen based on the judgement of domain experts.

Several classifications of forecasting methods have been proposed over the years (1971; 1995), with the most recent done by Swan (2008) who identified three common methods of forecasting air travel demand: trends, gravity models and simulation. Trends are the most common forecast technique used for air travel demand, which involves the use of econometric models in which passenger and freight demand are regressed against economic activity over time periods, such as the change in Gross Domestic Product (GDP). In turn, factors that induce economic growth are sometimes also taken into account, for example, demographic variables such as population or middle-class growth. Time-series is another type of econometric model widely used in aviation, where the formulation of the model is based on the past behaviour of a variable -i.e. air travel demand- in order to account for patterns in the past movements of that variable (Pindyck and Rubinfeld, 1998). These methodologies clearly outline the close relationship between economic and aviation growth highlighting the importance of considering such variable when modelling future air travel demand.

Gravity models consider that the level of traffic between origin and destination city is based on the attraction and the spatial separation between both. The attraction of the origin and destination city can be defined in many ways, all representing the importance of the two edges (i.e. cities) of the connection. Examples of city-level attractiveness variables include population or GDP; whereas examples of spatial separation are travel time or distance between the two cities.

Finally, simulation models are those which estimate the rise in traffic from changes in fares and service levels. Additionally, it is worth briefly mentioning qualitative techniques, which are based on the intuition and the subjective evaluation of expert's opinions (Teyssier, 2012). Surveys and questionnaires are two common ways to collect data for this type of models. Often, qualitative techniques are used in conjunction with techniques mentioned above and help bring some clarity to explain passenger and/or freight movement patterns. For example, the passenger interview surveys conducted by the Civil Aviation Authority (CAA) feed the model's base year demand produced by the UK's Department for Transport (DfT) (DfT, 2013). Among several other purposes, these surveys are used to provide some information about journey purpose, to supply time series for international-to-international interlining[1] passengers and to account for the share of domestic interliners on domestic routes. Surveys can help understand passenger behaviour, however they tend to represent a small subset of passengers, are time consuming and costly to complete; making surveys not an adequate resource when modelling the air travel demand at the network level (i.e. where a large number of passengers and itinerary options are considered) as this research does.

The above forecasting approaches are some of the most common techniques currently used across the industry. Forecasts are derived in a range of ways depending on both the time and data available as well as the questions the forecasts are trying to address. And although the choice of methodology depends on the use and scope of aviation forecasts, approaches used amongst different aviation stakeholders are quite similar. A summary of the forecasting methodologies used by several industry stakeholders is introduced in the following sub-section.

2.1.1 Forecasting approaches used in aviation

Regression models

The most common methodology used across aviation stakeholders is regression models, mainly relating economic growth with air travel demand growth. For example, Boeing, whose forecasts are focused on developing aircraft demand forecast for their customers (i.e. the airlines), bases its methodology on a regression equation stating the relationship between passenger traffic, measured by Revenue Passenger kilometres (RPK), and a set of input variables that can be split in three groups: economic activity, ease of travel and local market factors (Boeing, 2017). Economic activity includes GDP development, per capita income, labour force composition and international trade and investment links. Within the ease of travel category examples

[1]Interlining, also known as interline ticketing or booking, is a commercial agreement between airlines to handle passengers travelling on itineraries that require multiple flights on multiple airlines.

include market liberalisation effects, such as the arrival of low-cost carriers and the consideration of new routes and/or greater frequencies of existing ones due to the lifting of constraints that existed prior to liberalisation. And local market factors relate to any factor that is not directly connected to macroeconomic aspects or to ease of travel, such as not having enough capacity growth to accommodate the demand.

Similarly to Boeing, The International Civil Aviation Organization (ICAO) and Eurocontrol, which are two examples of international agencies that generate air travel demand forecasts in order to set up the principles that will help maintain a safe and orderly growth of the industry, use regression models based on economic and non-economic factors, such as cost of travel and a set of dummy variables to take into account random events, such as 9/11 (Teyssier, 2010; ICAO, 2016; Eurocontrol, 2017). Both organisations use a bottom-up approach starting with specific sub-models and then combining the results of these sub-models to produce the final traffic forecast for each country of the European Union (EU) for Eurocontrol (Eurocontrol, 2013, 2017), and globally for ICAO (Teyssier, 2010).

All forecasting approaches mentioned above have gone through improvement phases over the years derived from the need of more accurate and compact econometric models and aggregated air traffic forecasts from which more detailed forecasts could be derived for various purposes and to better capture the dynamism of the industry (Boeing, 2017; ICAO, 2016; Eurocontrol, 2017). Although improvements have been made throughout the years, all models are characterised by their simplicity, and in some cases by a perceived lack of impartiality (e.g. aircraft manufacturers' forecasts tend to be less conservative since their purpose is to encourage airlines' spending on new aircraft), as well as similarities in methodology and input variables. Moreover, none of these methodologies consider disaggregate evolution of the air transport system -i.e. connectivity changes-, and therefore, are considered to be a better fit for short-term forecasts.

Modular structure models

Other stakeholders, such as the Department for Transport (DfT) -i.e. which produces extensive aviation forecasts in the United Kingdom (UK)- or the Federal Aviation Administration (FAA) -i.e. which is the national agency responsible for the air traffic forecast in the US- use a more elaborated methodology based on a combination of regression models, splitting the forecasting process in several phases. Both the FAA's and UK DfT's forecast estimates are produced in order to inform policy making.

On the one hand, UK Aviation Forecasts are based on a top-down econometric approach, split into three steps. First, the national air passenger demand at an aggregated level is produced. This first stage is built on the basis of unconstrained capacity. Secondly, introducing the effects of capacity constraints, the previous demand predicted at the national level is broken down to compute air traffic demand at airport level, which is measured by Air Transport Movements (ATMs) counted as landing or take-off of an aircraft. Finally, the forecast air traffic levels are then used to produce information regarding passenger movement, costs or to estimate greenhouse emission levels (DfT, 2017).

UK DfT's aviation forecasts are based on a set of regression equations, which vary, in terms of model formulation and set of input variables, depending on the stage of the modelling approach they are used at. For example, during the first stage, which focuses on forecasting air travel demand at the national level, the set of influencing factors are grouped into two separated contributions: the continuous decrease in airfares, which is derived from the fuel price fluctuation, the decrease on non-fuel related costs, the Air Passenger Duty (APD) - accounting for the tax revenues paid by passengers- and the carbon costs; and a steady economic activity growth, which is measured as a combination of the UK and foreign GDP, the UK consumption and the imports and exports of the country (DfT, 2013).

Differently to stage one, UK DfT's second stage, which focuses on breaking down the aggregated demand from stage one to several sub-networks that form the under layers of the system, uses a different formulation for the regression equation: a multinomial logit. This model is used to forecast the volume of passengers across UK airports and the factor driving that demand is the cost of travelling through a specific airport. For this stage, input data used to obtain the parameter estimates comes from the Civil Aviation Authority[2] (CAA) airport choice data -e.g. using the CAA Passenger Survey.

The sum of estimated forecasts for a given origin airport will give the total unconstrained demand of that given airport. For the final stage, UK DfT's method-

[2]The CAA is the specialist aviation regulator.

ology applies the current airport capacity levels in order to derive the constrained air traffic forecast for each airport through an iterative process. In this iterative demand re-allocation process, shadow costs[3] - i.e. taking form of runway slot cost or terminal cost - are added to the costs of using each over-capacity airport before repeating the passenger allocation stage. After each re-calculation of the ATM numbers a check is performed to see if those new predicted passenger and ATM numbers fit terminal and runway constraints across all modelled airports. If they do the model is said to have converged for that year, if not the iterative process continues until a solution is found in which both types of capacity (i.e. terminal and runway) are not exceeded at any airport. UK DfT's methodology also does not consider connectivity changes, unless previously announced by airlines, and therefore, the evolution of the network is not considered.

On the other hand, in the US the FAA produces the Terminal Area Forecast (TAF), which is the official FAA aviation demand forecast for the US air transportation system. The TAF is based upon historical local and national measures that influence aviation activity as well as drivers within the industry itself, such as fossil fuel prices. A particular airport demand is derived independently of the ability of that airport and its air traffic control system to furnish the capacity required for meeting that specific demand. However, if the airport has been historically capacity constrained, this would be reflected in the forecasts as the factors considered to influence airport demand are embedded in historical data (FAA, 2016). The FAA's methodology is also based on regression models.

Similarly to the UK DfT methodology, the FAA's forecasting approach follows a top-down approach, split into three stages. The forecasting process starts by producing the origin and destination (O&D) market demand forecasts; these forecasts are then combined with the US Department Of Transport (DOT) T-100 segment data[4] in order to estimate the passenger demand by airport- and segment-pair; and finally the segment-pair air traffic level forecasts are assigned to aircraft equipment

[3]Shadow costs refers to costs of using an airport that exceeds its capacity until its demand falls within its maximum capacity and it is used to adjust the passenger demand considering a scenario when capacity constraints exists. It is mainly represented in two ways and it is added to the passenger cost of using a specific airport: as runway slot shadow cost; and as terminal shadow costs. But it can also represent the value a marginal passenger would place on flying to/from that airport if extra capacity were available.

[4]T-100 data (Form 41) is a database where commercial airlines report all passengers that flew routed segments. T-100 data, which contains the most recent airline schedule, is used to identify those available itineraries connecting a given airport-pair.

so that segment-pair operation forecasts are produced.

The factors considered to influence the air travel demand include airfares, income, distance and number of routes available between a given market. These variables are similar to those used by the UK DfT, reflecting mainly on the cost of travelling and capacity available.

In contrast to the UK DfT methodology, the allocation of air traffic across airports (i.e. trip distribution model) does not use multinomial logit formulation but instead a growth factor method is used. Growth factor methods are based on the assumptions that the trip making pattern will remain the same in the future as it was in the base year; while the volume will increase along with the growth in the generating zones as well as with the growth in area attractiveness.

The itinerary assignment is done by the Fratar algorighm, a type of growth factor method, which allocates the traffic previously forecast at the market level across all possible routes of the network constructed from the airline schedule. This allows to evaluate the connectivity between two trip ends is evaluated (Viken et al., 2006) and as a result, the future daily airline schedule among O&D airport-pairs of the current schedule is generated.

The air traffic allocation methodology used by the FAA, the Fratar algorithm, bases its success on the fact that it is simple to use and understand as well as that it conserves the observations as long as is consistent with the information on growth rates available. However, those advantages become its drawbacks, making its approach more reliable for short-term predictions (Ortuzar and Willumsen, 2001). One of the main limitations is the fact that it does not consider changes in transport costs and assumes that resistance to travel will remain the same, and therefore, the addition of new facilities and/or routes is neglected. Consequently, FAA's methodology does not consider the evolution of the network, unless information of airlines' network changes is made previously available, reducing its reliability for medium- and long- term forecasts. Moreover, growth factors methods are highly dependent on the accuracy of the base year trip matrix, which is usually not that high, meaning that the final matrix is not highly reliable.

The resulting passenger demand at segment-level is then compared and adjusted with T-100 segment passenger data (BTS-RITA, Bureau of Transportation

Statistics. Research and Innovative Technology Administration, 2014). T-100 also gives information regarding types of aircraft flown on those segments, and therefore, can be used along with the predicted passenger demand to project aircraft departures by segment.

2.1.2 Discussion

As discussed in section 1.2, forecasting plays an important role across aviation, being a key tool for decision-making processes in which stakeholders are involved. It is clear that the constant evolution of the industry and increased complexity require continuous work on forecasting methodologies for air transportation, and that often current industry forecasting approaches are under review for further improvements, as presented in the previous section (2.1.1).

It is also clear that econometrics is the most common approach used amongst industry's stakeholders, with model specifications that are somehow similar mainly amongst them, showing the strong link between aviation and economic growth. With aviation being a large and complex industry in continuous evolution, the need for a modelling framework that considers the dynamics of the network (i.e. the consideration of routes being added and removed from the air transportation system) and the effect that these dynamics have at different levels of the network (i.e. passenger choice, airport congestion, etc) is necessary. Focused on developing a sophisticated and yet user-friendly model, -i.e. simple enough that there is a will to be used across the industry stakeholders-, the work carried out in this research looks at providing with a single modelling framework that models city-pair passenger demand, airport connectivity changes, itinerary demand assignment and air traffic estimation. By considering the possibility or airport connectivity changes (i.e. airport-pairs being added and removed to/from the network) the structure of the network and its characteristics will change over time and the impact that these changes have into itinerary choice and consequently to air traffic levels can be evaluated.

2.2 Research work in air travel demand forecasting

The tendency to use econometric models to estimate the air traffic demand forecasts has been emphasised in the previous section. Moreover, it can be concluded that the majority of aviation stakeholders that produce air traffic forecasts use this type of methodology based on analysing past data and trends.

Considering the important role that forecasting air travel demand plays for the future of the aviation industry, as mentioned in section 1.2, a large amount of research in this topic has been carried out in the past years. A large part of the research has focused on improving existing methodologies (i.e. by using alternative econometric models), such as the research done by Evans (Evans, 2010) which combined a gravity model with an agent-based model in order to model airlines' operational responses to environmental constraints. However, there is a growing trend towards the use of machine learning techniques with the aim of predicting future air traffic levels. Technically speaking, machine learning is the field of study that gives computers the ability to learn without being explicitly programmed, i.e. a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E (Ng, 2013). In contrast to statistical models, which are designed for inference about the relationships between variables, the purpose of supervised machine learning is obtaining a model that can make repeatable predictions; and that can provide various degrees of interpretability depending on the model, from highly interpretable such as lasso regression to non-interpretable, such as neural networks (i.e. machine learning models generally sacrifice interpretability for predictive power).

2.2.1 Machine learning within aviation

Machine learning has experienced an enormous development since its beginnings in the 1930s with Ronald A. Fisher and in 1950s with Frank Rosenblatt's linear perceptron (Alexander, 2013). In the last decade, its range of applications has grown extensively, especially for complex learning problems, partly because of increase in data processing power and faster compilers. Therefore, what started as a simple attempt to separate points in a plane evolved into, among other developments, solving structured learning problems, such as speech recognition and medical diagnosis, and learning with massive amounts of data. Machine learning models are designed to make the most accurate predictions possible.

The range of fields in which machine learning is applied is vast, from biology to cognitive or sociological science. Aviation is a large, complex system involving the generation of large scale and unstructured data in various data formats. This makes air transportation an interesting system to apply machine learning techniques aiming to transform those datasets into applied knowledge. From the application of machine learning techniques to air transport data, several benefits can be obtained across different areas, such as better optimisation of airlines' fleet management, advances in safety within the air transport system, as well as improving existing air travel demand forecasts.

Focusing on airline's performance, Lawson and Castillo (2012) employed large amount of data available on flight punctuality to predict whether or not a flight will be delayed.

In the area of safety, the extension in which machine learning techniques have been applied covers mostly algorithms for anomaly detection, which consists of the analysis of air safety data in order to identify general patterns that define the majority of flights. As a result, approaches to identify possible anomalies that could lead to an incident or accident during the riskiest phases of a flight have been developed, producing a shift towards a more proactive safety management. Cluster analysis on continuous flight parameters (Iverson, 2004), multiple kernel anomaly detection which considers both discrete and continuous data streams (Das et al., 2010) and text mining to predict and discover precursors to safety incidents (Smalley, 2012; Srivastava, 2011) have all been applied.

2.2.2 Machine learning in air travel demand forecasting

Within the area of air travel demand forecasting, only a small number of studies have used machine learning, such as the work done by Nam and Schaefer (1995) and the work done by Cheung and Gunes (2012). The work done by Nam and Schaefer (1995), one of the first studies in this field, used neural networks for forecasting international airline passenger traffic between the US and South Korea. Results showed how the neural network model used to predict air travel demand outperformed in terms of accuracy the more conventional statistical analysis used at that time. The choice of neural network was driven by the need of an alternative technique free of any distributional assumption about the model errors - as opposed to regression modelling-; or reduced complexity of the model-building process that existed in some pattern recognition procedures.

Also, considering neural network models, Cheng et al. (2003), applied a hybrid model consisting of neural network and statistical analysis to forecast traffic flows in China's air network. Their research was driven by the need to create a model to predict air traffic flows at a network level using air traffic information from each Chinese regional control centre as opposed to the then current system where a distinct set of predictions was generated separately for each regional control center. The aim of this research was to construct an air traffic model that would improve the efficiency of the current air traffic flow management system, which directly depends on the accurate predictions of the air traffic flow of each regional control center independently. And with a slightly different line of research Kotegawa et al. (2012) explored the use of machine learning techniques, such as neural networks and support vector machines[5], alongside network theory to study the evolution of the US air transportation system (i.e. identification of airport-pairs that will be added and removed to/from the network).

As the examples above show, there is great potential in applying machine learning techniques to those most commonly used within aviation to explore whether better predictive accuracy can be obtained. As the air transportation system evolves, the need for new techniques when trying to model different areas of the system is proven - i.e. also shown through the continuous improvements that forecasting methodologies used across the industry have undergone in order to adapt to the increase complexity and factors influencing the air transport system. Within the work presented in this dissertation, the two main contributors are: the use of network theory to model the evolution of the air transportation system, which as highlighted in the previous section (2.1), is a key aspect to understand the dynamics of the network and the effect that this has on airport-, route- and system-level demand and it is not currently considered in the existing forecasting methodologies; and a comparison analysis of the predicting power of a discrete choice model and a neural network model (i.e. machine learning model) to determine air itinerary shares and assess how the consideration of network evolution affects the predicting power of each of these methodologies. A summary of existing research related to these two

[5]Support vector machine (SVM) algorithm is a supervised learning model most commonly used for classification problems and its objective is to find a hyperplane in an N-dimensional space (N - number of features) that distinctly classifies the data points.

areas is presented below.

2.2.3 Application of network theory within aviation

Understanding the underlying reasons of airport-pair connectivity and how these can change throughout the years can benefit communities of all sizes and be used to understand and predict network growth. For example, Wittman and Swelbar (2013b) showed that small- and mid-sized airports[6] have been largely affected by cuts in commercial air service in the US by looking at the reduction of number of scheduled domestic flights in US's largest airports compared to the reduction at smaller airports during the recession years (i.e. 2007 - 2012). Data showed that while US's largest airports lost 8.8% of their domestic scheduled services, smaller airports suffered a reduction of 21.3%, which combined with the analysis of airline behaviour at the time resulted on the conclusion that there was a trend towards consolidating service at the nation's largest airports.

In order to study airport-pair connectivity, Wittman and Swelbar (2013a) identify three categories of research approaches: network theory models; temporal sensitivity models; and more simple models referred to as *intuitive metrics* by Arvis and Sheperd (2011). Network theory models are those that provide a holistic perspective modelling the air transport system as a natural network that consists of well-defined nodes (airports) and links (flights that connect those nodes or airports) and by characterising these nodes through a series of mathematical quantities that define their importance and the influence they have towards the network.

The temporal sensitivity models can be classified as the most robust method within airport connectivity modelling. This type of model examines air transport connectivity of only those connections that are reasonable or feasible for a passenger to take. Consequently, itineraries involving lengthy layovers or unreasonably small connection times should be excluded when computing any connectivity metric if possible (Wittman and Swelbar, 2013a). To construct the set of feasible itineraries, much more detailed schedule data is necessary, making these models harder to generalised to large networks (i.e. such as the US ATS) when hundreds of airports need to be considered.

[6] The definition of small- and mid- size airports used in Wittman and Swelbar's work refers to medium-hubs, small-hubs or non-hubs as defined by the FAA.

Finally, the intuitive metrics are the simplest and easiest to understand of the three categories and considers airport connectivity based on a score computed by the product of those intuitive metrics. The most attractive intuitive metrics generated are those that measure the quantity (i.e. such as available seats per annum) and quality of available service (i.e. larger airport would tend to to be considered as more valuable) and destination (Reynolds-Feighan and McLay, 2006; Pearce, 2007). While the simplicity of the intuitive metrics makes this approach appealing, it has an important limitation: it only consider non-stop itineraries; limiting its application to those airports that are only served by airlines operating point-to-point (i.e. low-cost carriers). Consequently, this methodology cannot be used within the US ATS, where most of the major US carriers follow a hub-and-spoke network.

Considering the drawbacks of using temporal sensitivity models and intuitive metrics -i.e. difficulty generalising to large networks and the limitation of only considering non-stop itineraries-, this dissertation focuses on the potential of network theory to analyse the dynamics of the US air transport system. The application of network theory to model the air transport system has attracted attention recently, becoming a trend in the last decade. The great variety of tools developed for the analysis of different topologies has helped form a better understanding of the structure and dynamics of many real-world systems (Zanin and Lillo, 2013). The study conducted by Guimerà et al. (2005) was one of the first of many to have taken into account this methodology applied into the air transportation system. It was also one of the first studies that described measures of the mathematical qualities of this network such as eigenvalue centrality, node weight or node degree among others. Guimerà et al. (2005) concluded in defining the worldwide air transport system as *a small-world network in which (i) the number of non-stop connections from a given city and (ii) the number of shortest paths going through a given city has distributions that are scale-free.* Surprisingly, the resulting analysis showed that those nodes with more connections were not the most central ones -i.e. the nodes through which most shortest paths go- which is characteristic of scale-free networks. They suggested that this behaviour was due to the multi-community structure of the network, which are influenced not only by geographical constraints but also by geopolitical considerations.

Only focusing on the United States Air Transportation System (US ATS), Cheung and Gunes (2012) used network theory to analyse the network and its evolution from 1991 to 2011. In line with Guimera et al. (2005) work, results of this study show how the US ATS exhibits small world characteristics - i.e. such as small average shortest path[7] and a large clustering coefficient[8]; however, only a partial power law degree distribution was observed, implying that airports in high densely populated areas grow at a slower rate than those located in less populated areas. The study also showed that although most of the characteristics of the US Air Transport network remained the same through those 2 decades, the network had become more vulnerable to airport closures throughout the years.

Other research uses network theory to model the air transport system and better understand its fundamental characteristics. Bonnefoy and Hansman (2007) used network theory to analyse the Origin&Destination (O&D) routes flown by existing light jets to understand the principles underlying this air transport system and be able to assess how the introduction of very light jet traffic would affect the network's evolution. DeLaurentis et al. (2008) applied network theory to simulate the contraction of the air transport network.

Nevertheless, network theory has not just been used for the assessment of the topology (i.e. characteristics) and structure of the aviation system, but it has also been adopted to explore and identify the dynamics taking place in it. For example, Lacasa et al. (2009) proposed a network-based model for the ATS to simulate the effect of traffic dynamics and showed the appearance of air traffic jams. In the same line of research, Fleurquin et al. (2013) used network theory metrics to characterise the delay propagation of the US airport network. Others -e.g. Colizza et al. (2006)- used this research field to study the air transport system's impact on spreading epidemics.

Finally, the resilience and the vulnerability of the air transport system have also been evaluated through the use of network theory. Resilience is defined as the intrinsic ability of a system to adjust and quickly recover its functioning before, during or after changes and disturbances, so that it can sustain required operations under both expected and unexpected conditions (Hollnagel, 2013). Three studies

[7]The shortest path between airports i and j is the path with the fewest number of flights (i.e. links)

[8]The clustering coefficient captures the degree to which the neighbours of a given node link to each other. The higher the clustering coefficient is, the more robust the network is since in case of a link failure alternative paths would exist.

are worth mentioning. Chi et al. (2004) analyse how the main topological characteristics of the US air transport network respond to random failures and attacks. A similar analysis, performed by Wilkinson et al. (2012) focused on the European air transport system, concluding that severe disruptions due to natural disasters such as the eruption of the Icelandic volcano Eyjafjallajökull in 2010 are the result of the geographical correlation of the disturbances and the geographical correlation of hubs concentrated in central Europe. Finally, Kotegawa et al. (2010) studied the impact that targeted and random attacks have on multiple stakeholders that form the aviation system.

Even though many have used network theory as a method to study the topology and structure of aviation and better understand the fundamental principles that govern it, little research has been done to use network theory to predict the future evolution of the Air Transportation System. The main work within this area of research is the work done by Kotegawa (2012), who developed a network restructuring model for improving air traffic forecasts using several machine learning techniques.

Kotegawa's innovative approach used network theory metrics (i.e. quantitative parameters that characterise the airports that formed the network) as explanatory variables to constitute the input dataset that was used to train several algorithms in order to predict future air travel demand. The main objective of his research was to develop a model that captured the mechanisms of the US Air Transportation System network evolution. In other words, an algorithm that predicted the likelihood of unconnected city-pairs being connected by service in the future and the likelihood of connected city-pairs being unconnected in the future.

The application of network theory defined the structure of the US ATS as a network composed of nodes (i.e. represented by airport) and links connecting those nodes between each other (i.e. represented by flights covering route segments). The nodal network properties were used to characterise all existing connections between the nodes and as input variables to the forecast algorithms that were considered.

The problem was separated into two different paths in order to focus on the link addition and the link removal process separately[9]. Figure 2.1 shows the flowchart of the ATS capacity network model used by Kotegawa (2012). In Figure 2.1 the separated paths mentioned can be seen as the link addition (i.e. phase A) and the

[9]Note that link refers to flight.

link removal (i.e. phase B) module. After the link addition module (i.e. phase A) in which newly established links are identified, historical distribution based on the distance between airports are used to assign number of operations (i.e. weights) to those new links (i.e. phase C). Weights (i.e. number of operations) for all links projected to exist in the network (i.e the sum of those added and those not removed) are adjusted (i.e. phase D) so that the aggregated value matches the total air traffic levels projected by the Terminal Area Forecast (TAF). This adjustment process is done by using the Fratar algorithm. Two set of models were generated based on the two tools used to assess the impact of the network structure changes would have to system-wide metrics, such as the impact on flight delay. The two tools used are: NASPAC, which is the FAA system using TAF and Fratar algorithm to allocate the growth in air traffic; the NASA-Purdue Fleet-level Environmental Evaluation Tool (FLEET) simulation, which is a tool set that investigates the airline fleet-level environmental impact from new aircraft technologies and concepts (Zhao et al., 2009). Different methodologies were used to generate the two sets of models, which considered a different set of airports.

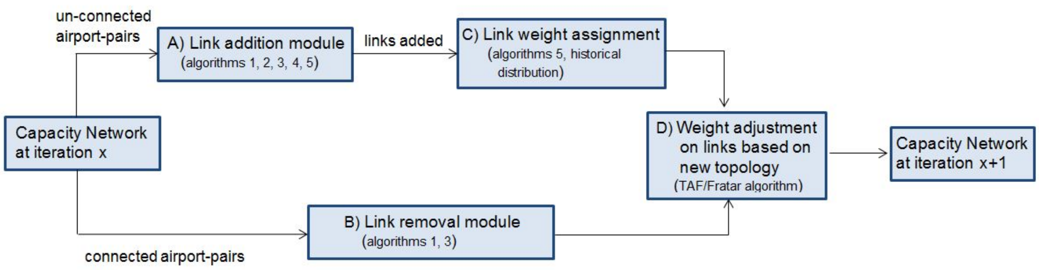

Figure 2.1: Flowchart on Kotegawa's ATS capacity network model

The link addition module was developed using Support Vector Machine (SVM) in combination with Logistic Regression for the model integrated with NASPAC; whereas Random Forest[10] (RF) was used for the model integrated with NASA-Purdue FLEET simulation. The link removal module, where only existing connected airport-pairs constitutes the input dataset, was developed using logistic regression. Amongst all the network theory metrics -i.e. also defined as nodal properties or topological parameters - that Kotegawa considered, four were identified as the ones giving the best modelling performance results: node degree, node weight, eigenvector centrality and clustering coefficient. Further explanation regarding those four nodal

[10]Random Forests is an ensemble method for classification (can also be used for regression) which consists of a collection of three-structured classifiers.

properties can be found in Chapter 3. These four nodal properties were used to define several model specifications for the link addition and link removal modules. The overall forecasting accuracy for the link addition and link removal modules were around 20% and 40% respectively, which is considered quite low.

The potential of using network theory in order to simulate air transportation system evolution is clearly shown through the literature presented in this section (Guimera et al., 2005; DeLaurentis et al., 2008; Lacasa et al., 2009; Kotegawa, 2012). With most of the existing research looking at the application of network theory as a method to understand the system dynamics and to model the response of an existing air transport systems to events (Colizza et al., 2006; Hollnagel, 2013; Kotegawa et al., 2010; Fleurquin et al., 2013), only Kotegawa (2012) used network theory combined with machine learning techniques to expand the air traffic demand forecasts produced by the FAA in an attempt to capture the mechanisms of the US ATS network evolution (Kotegawa, 2012). This research used network theory metrics as input variables, and only one demographic characteristic (i.e. distance between airports) was taken into account, which omitted many parameters that are equally important in underlying the ATS system behaviour, such as passenger demand. Moreover, the accuracy obtained was quite low, leaving room for improvement and further exploration of the subject. In addition, research focused on improving the allocation of air traffic demand (i.e. itinerary choice model) was not carried out. Instead, the FAA TAF's methodology was implemented. This uses the Fratar algorithm which is a type of trip distribution method using growth factors and has some important drawbacks such as not considering changes in transport costs and the assumption that resistance to travel will remain the same as the base year used as well as to be more suitable for short-term forecasts. Focusing on the impact that the evolution of the network would have for the air transport system, this dissertation also explores different methodologies to assess the impact that changes in network structure would have passenger behaviour at the aggregate level and that would help to determine itinerary shares, and consequently air traffic levels at the network level.

2.2.4 Existing research in air itinerary choice models

There is a growing interest in developing better itinerary share models than those already existing. Itinerary share models can become crucial to support airlines in their network planning and scheduling since important decisions on resources allocation and pricing are made based on itinerary demand.

Most of the current research centres on developing innovative approaches using discrete choice modelling, which aims to model competition and customer behaviour to determine air travel itinerary shares – also known as demand assignment models. While most of the discrete choice models applied in urban transportation are built using disaggregate data and include information about the individual making the decision (i.e. the passenger); in air transport, data disaggregation as well as data accessibility are limiting factors - i.e. disaggregated data consists of information at the individual level only possible to gather through surveys which tend to be costly and difficult to carry out for large aviation systems. The need to quickly adapt to changes in demand makes flexibility crucial for carriers and other stakeholders in the industry. For this reason, most of the models built, and found in the literature, to support decision-making rely on booking data which is generally proprietary. Furthermore, airlines do not typically record much of the passenger data that is relevant to passenger decision making, such as age, gender and income. This data is not typically available, except for a small subset of passengers which limits the applicability of the model to small networks or subset of airports, and therefore it cannot be used for the purpose of modelling itinerary shares at the network level -i.e. at the most aggregated level.

Most of the early studies in demand assignment for air travel focus on studying the distribution of demand across one single dimension, the choice to be made affects only one aspect of an itinerary such as choosing the origin airport, or the airline to fly with separately. These early models were mostly applied to analyse air travellers' choice within multi-airport cities or regions – i.e. airport choice models (Hansen, 1995; Windle and Dresner, 1995) – or across airlines – airline choice models (Proussaloglou and Koppelman, 1995). Although the former is the most widely studied topic in discrete choice modelling within air transport, and has given a deeper understanding to the relationship between airport attributes and airport market share, a more aggregated assignment of air travel volumes is also needed.

There are only few studies that present work for itinerary market share estimation across multiple dimensions using discrete choice modelling. Some of these studies used multinomial logit (MNL) formulation (Adler, 2001; Coldren et al., 2003;

Grosche and Rothlauf, 2007; Atasoy and Bierlaire, 2012); others apply nested logit (NL) formulation (Coldren and Koppelman, 2005; Hsiao and Hansen, 2011), mixed multinomial logit (MMNL) formulation(Warburg et al., 2006) and other alternatives methodologies (Grammig et al., 2005; Carrier, 2008). Also in this group the only existing research using machine learning to model air itinerary shares can be mentioned. The work done by Grosche and Rothlauf (2007) is a comparative study of three methods used for itinerary market share estimation: multinomial logit, neural networks and a custom model developed by the authors; using booking data (i.e. proprietary) and with the application example including markets between Germany and European countries. Results obtained showed that the custom model developed by the authors was the best performing model and disregarded the NN due to lack of model interpretability.

The studies in choice behaviour modelling mentioned above can be classified with respect to the type of data they are based on: revealed preference data (RP) or booking data (Coldren and Koppelman, 2005; Hansen, 1995; Windle and Dresner, 1995); stated preferences (SP) data or survey data (Hess and Polak, 2005; Pathomsiri and Haghani, 2005); and a combination of both (Atasoy and Bierlaire, 2012). Studies using RP data do not usually provide full insight into passenger choice behaviour since models are estimated based on real booking data, and no information regarding other alternatives at the moment of booking is generally available. This limitation might lead to RP models performing poorly due to the insufficient variability in the observations. In contrast, SP data collected from surveys allows for modelling of new alternatives - i.e. hypothetical situations that do not exist yet, as well as more accurate estimation of the sensitivity of travellers to characteristics of their trips. However, studies using SP data may be subject to bias due to the nature of the experiment as tailored by the survey writer. These studies are also often limited to a small range of markets, limiting their application to a small network set.

Most of the studies that focus on air itinerary choice models mentioned in this section have been developed using disaggregate data, with information obtained directly from airlines or surveys (i.e. SP data) (Coldren et al., 2003; Coldren and Koppelman, 2005; Atasoy and Bierlaire, 2012; Grosche and Rothlauf, 2007) containing preference information (e.g. airline preference and departure time preference), which as mentioned is most often not available. Most of the existing work also focuses on studying air travel demand mainly at the segment and market level. Little research has been done in modelling air itinerary shares at the network level -i.e. at the most aggregated level- on its own, rather than treating it as a summation of the demand of individual markets involved. Some advantages of modelling aggregate demand at the network level are the ability to capture the impact of airport capacity on passengers' travel behaviours as well as the impact of the hub-and-spoke routing structure that characterises the US air transportation system.

External factors, such as software computational limitations during the estimation process have also been a common issue encountered by some work in choice modelling, forcing some studies to limit the number of city-pairs being analysed (Weidner, 1996) or to split the problem into a set of sub-models estimated with a smaller dataset (Li et al., 2017; Busquets et al., 2018) -i.e. other studies such as Coldren (2003) Atasoy and Bierlaire (2012), or Ghobrial and Soliman (1992) also analysed a reduced number of city-pairs although is not explicitly stated whether it was a choice made due to software limitations.

At last, none of the existing research focused on itinerary choice modelling considers the evolution of the network, but only those itineraries that are available in the base year are considered as an option for future available choices.

2.2.5 Discussion

From the literature review presented above, it can be concluded that there is an effort to explore the potential of alternative data mining and machine learning techniques applied to improve air travel demand forecasts (Nam and Schaefer, 1995; Cheung and Gunes, 2012; Kotegawa, 2012). The constant evolution of the industry and increased complexity also require continuous work focused on improving forecasting methodologies for air transportation. And while significant contributions have been made to date, there is still room for improvement. This research focuses on exploring two areas within aviation forecasting: airport connectivity (i.e. network evolution) which is not considered in existing forecasting methodologies used across the industry; and air itinerary choice modelling at the network level.

When analysing the air traffic dynamics in the past changes in the air transport network structure experienced over time are easy to identify. However, forecasts generated by aviation stakeholders are based on the assumption that the future route network structure will remain the same as the current network structure (FAA, 2016). Therefore, when trying to project the ATS evolution mechanism, the ability to reproduce the addition and/or removal of routes are not considered, unless previously communicated by airlines that operate within an the air transportation system. Since airlines do not plan their network structure far ahead, those network changes would be only applicable to short-term forecasts and in addition this type of information is usually not available to everyone.

Only one research focused on modelling network evolution (Kotegawa, 2012), however accuracy obtained was measured at approximately 20% and 40% for the link (i.e. airport-pair) addition and removal forecasts respectively, leaving room for improvements. Kotegawa's work (2012) only considered network theory metrics and did not focused on improving the allocation of air traffic demand, and therefore without considering the effect that network structure changes would have to the overall system, such as itinerary shares or impact on aircraft operations. Inspired from Kotegawa's efforts and looking at providing a further understanding of the air transportation system evolution, the work presented in this thesis further explores the application of network theory to predict airport-pair connectivity within the US domestic network. This is done by considering the use of input variables beyond network theory metrics as opposed to existing research, which include passenger demand, distance between airports and dummy variables characterising whether an airport is a hub or not. By including O&D passenger demand as input variable, the predicting capabilities of the connectivity models are extended to those links formed by airports without any connection elsewhere.

Regarding itinerary choice modelling, there has been a lot of effort in the study of air itinerary shares as presented in the literature review above. However, some limitations can be identified: little research has been done in modelling air itinerary shares at the network level -i.e. at the most aggregated level- on its own since most of the existing work focuses on studying air travel demand at the segment and market level; the applicability of some of the models is also limited -i.e. to specific markets- and further model refinement and verification is still required to better capture passenger choice effects, influenced by the fact that data used during the estimation process is a representation of a small subset of passengers and might not be reflective of the entire population; computational limitations during estimation process have also been a common issue limiting the predicting capabilities of the models to a subset of cities; and none of the existing research considers the dynamics of the network and therefore only itineraries available in the base year are considered as an option for future choices.

Considering these limitations, this dissertation also centres on developing an air itinerary share model at the network level (US ATS) considering non-stop and one-stop services. Since the work done in this dissertation considers network evolution, two different modelling methodologies are explored to assess how itinerary choice is affected by network changes in the long term. The two modelling techniques explored and compared are multinomial logit and neural network. In the case of multinomial logit, Berkson-Theil approximation method - i.e. which transforms the estimation process to a least square formulation - is used to overcome the computational limitations that the maximum likelihood estimation process has associated when considering large amount of data.

The interaction between both models (i.e. airport connectivity and itinerary choice) is integrated within a single modelling framework where the response from the system to network structure changes could be assessed, such as passenger choice, air traffic levels, airport congestion and environmental impact.

Chapter 3

Modelling Framework

Airline route planning decisions are a response mainly to changes in cost and passenger demand, which in turn affects competition and airlines' profitability. An airline's strategic decision to start operating a new route or cease an existing one is generally down to a simple question of whether that route is profitable or not. Because passenger demand is not the only factor to be considered in the equation when predicting air traffic levels - i.e. presence of competition or aircraft type operating a given route can also influence air traffic levels -, the approach employed in this research focuses on understanding the underlying principles driving demand. The modelling framework presented in this dissertation has a modular structure looking at city-demand generation, itinerary demand assignment and air traffic estimation. Broadly, a demand module projects passenger demand between cities -i.e. O&D passenger demand. Then, the itinerary demand assignment module is divided into two parts: first, a connectivity model identifies which airport-pairs will be added and removed from the network allowing the identification of available itineraries serving each city-pair; second, considering the available routes previously identified, an air itinerary choice model projects which airports and route the projected O&D passenger demand will choose. Finally, the air traffic module projects the number of operations between airport-pairs (i.e. air traffic levels between airport-pairs). A simplified diagram of the structure of the modelling framework presented in this dissertation is shown in Figure 3.1.

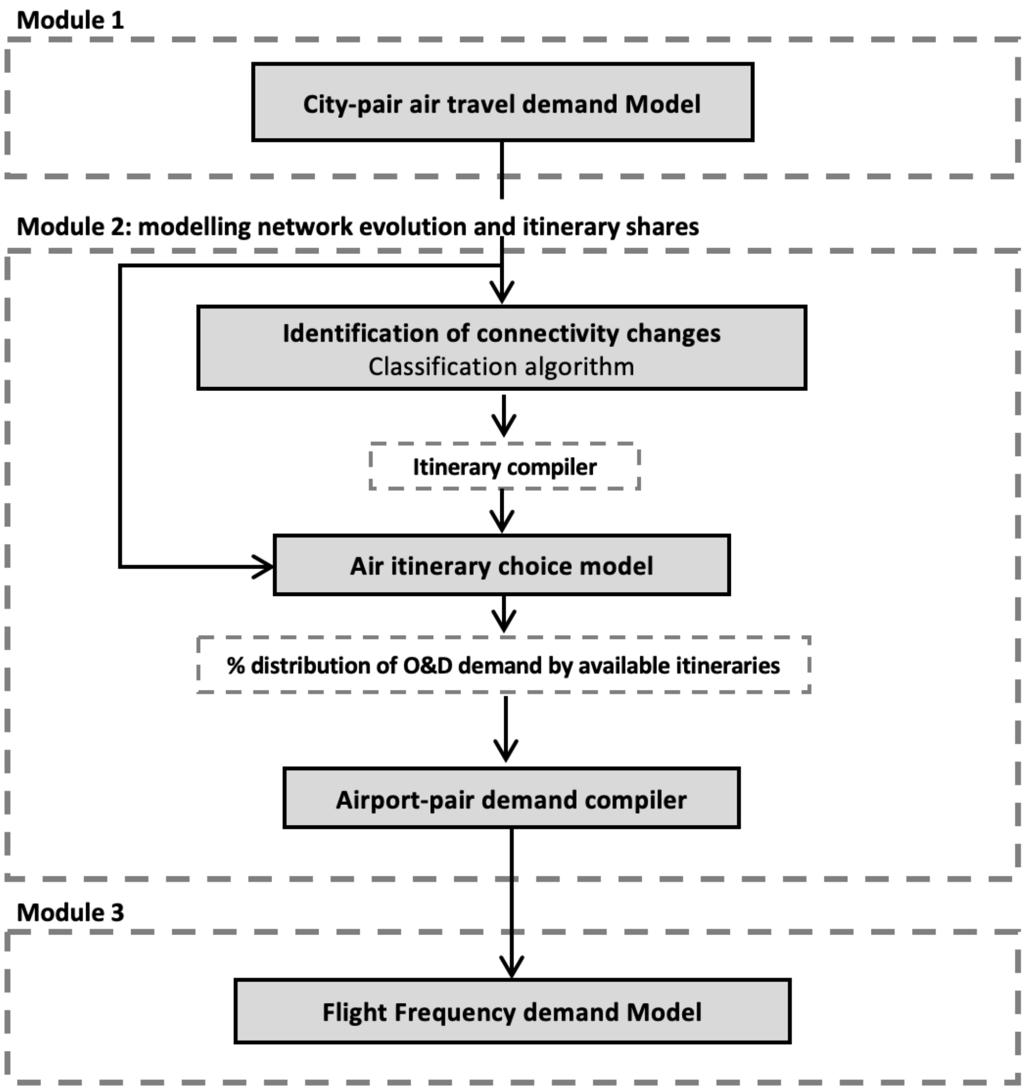

Figure 3.1: Modelling framework.

The structure of the rest of this chapter is as follows. Section 3.1 discusses the scope of the model in terms of geographic scope, types of flights, airports and passengers covered in this study, as well as aggregation level at which the modelling framework in Figure 3.1 refers to. Section 3.2 describes the different sources used to compile all the necessary data to estimate the different models and generate projections. Section 3.3 gives an overview of the methodology and the detailed information regarding each of the modules is presented in the following sections: O&D demand model in Section 3.4; airport connectivity model in Section 3.6, which includes a detailed description of network theory and an application example of this to characterised the US air transportation system; Sections 3.7 and 3.8 describe the two methodologies used to develop the itinerary choice models with Section 3.9 presenting a comparison between those; and Section 3.10 for the air traffic model. And Section 3.11 presents a brief summary of the findings associated with the modelling framework used in this dissertation. At last, projections for future years using the entire modelling framework will be presented in Chapter 4.

3.1 Scope

3.1.1 Geographic scope

The geographic scope of the model presented in this dissertation is the US Air Transportation System (ATS) and only information regarding US domestic flights has been used. The size of the network accounted in number of airports used in this research considers 337 airports within the US ATS. This subset of airports is aligned with the subset of US airports used by the AIM Project (Schäfer and Dray, 2015) which would enable to integrate both models if further research is done. The set of 337 airports represents the US airport set included in the global 1,277 airport set that contributes to the 95% of global RPK (Revenue Passenger Km). Airports have been classified as hubs an non-hub airports.

Information regarding US cities in which these airports are located are also aligned with the AIM project (Schäfer and Dray, 2015). In total, a set of 178 US cities account for the corresponding 337 airports. City characteristics include information regarding city attractiveness (i.e. indicator to whether the city is a major tourism or business destinations) and accessibility by other types of transport (i.e. road and/or rail).

3.1.2 Types of flights

Flights can carry passengers and/or cargo, flights can be scheduled or unscheduled, and they can be classified as commercial, general or military aviation. While a large amount of data is available for commercial scheduled flights, information for other types of flights can be scarce. Therefore, for the purpose of this research, only scheduled flights from commercial aviation have been considered. This includes flights in which aircraft carry only passengers and aircraft with mixed configuration which carry passenger and cargo at the same time. No limit on aircraft size or type has been applied when processing the input data.

For the purpose of simplicity, only non-stop and one-stop itineraries, accounting for more than 97% of the total passenger demand, have been considered in this study. This high percentage is representative of the fact that the geographic scope of this research is limited to the US domestic network and it is observed that a small amount of passengers will be travelling through itineraries with more than one-stop. From the literature, this is also a common practice for studies looking at US hub-and-spoke network (e.g. the work done by Hsiao and Hansen (2011)). For one-stop flights, the connecting airports are limited to a subset of 25 US hub airports[1], which still captures more than 96% of the total passenger demand of the original dataset.

3.1.3 Types of passenger

For the purpose of this study, passengers have not been divided according to the purpose of their trip (e.g. business or leisure) nor by demographic characteristics such as age or gender. Only one passenger category is used and all information regarding passengers (e.g. airfares) is computed as a weighted average using the number of passengers as weight across all the available data. This is due to data linked to the US datasets on trip purpose or demographics is not available. As a consequence, different responses of business and leisure passenger to changes in fare -i.e. price elasticity- are not captured. Instead, a single response to fare changes is tracked.

3.1.4 Aggregation level

The time-step considered in this study is annual. Consequently, both passenger demand and air traffic levels projections generated by the estimated models will refer to aggregated annual volumes. Similarly, input data used in the several models is annually aggregated, either as a sum or as annual weighted average as in the case of airfares. Models are estimated with 2007 data, and those will be used to model the years after 2007. Each sub-model is being validated in isolation using actual data from 2008 to 2013. Projections are generated from 2008 to 2025, with the first

[1]IATA codes for the 25 hub airports as considered and used in this research are as follows: 1-ORD, 2-ATL, 3-DFW, 4-LAX, 5-IAH, 6-DEN, 7-DTW, 8-PHL, 9-CVG, 10-MSP, 11-PHX, 12-EWR, 13-CLT, 14-IAD, 15-JFK, 16-LAS, 17-MIA, 18-SFO, 19-SLC, 20-SEA, 21-BWI, 22-STL, 23-CLE, 24-MEM, 25-PIT. Source: https://www.iata.org/publications/pages/code-search.aspx

10 years projections being used for validation of the modelling framework since they can be compared to actual data. The decision to use 2007 data to estimate the sub-models is due to data availability at the start of this research as well as 2007 being a relatively stable year not affected by any major external factor such as the economic crisis, making the models more representative of the general trends. Due to models being estimated with 2007 year data, all monetary-related variables (e.g. airfares and mean household income) need to be converted to 2007 value of money. This monetary conversion is done based on the Consumer Price Index (CPI) to 2007 US dollar values (Bureau of Labour Statistics, US Department of Labour, 2014a). Please refer to Appendix B for further information regarding CPI.

In all stages of the modelling framework at least 52 flights per year must operate between an airport-pair for it to be considered operational. Based on the research done by Dennis (2002) airlines would not normally start a service with less than a double-daily frequency short-haul and close to a daily long-haul. The present decision of 52 flights as the minimum number of flights for any given airport-pair to be considered connected, rather than a daily service as suggested by Dennis (2002), is due to several reasons: the US ATS is defined by a range of short-, medium- and long-haul markets and a mix of different sized airports as opposed to Dennis' (2002) research which focused on medium-sized European airports; the level of aggregation in this research is annual; and to be able to capture seasonal services that might be only operating during holiday periods, or regional services that might operate less frequently than other commercial services.

Further information regarding input variables used for each sub-model is explained in the following sections below.

3.2 Data sources

As mentioned in Section 3.1, the geographic scope of the modelling framework presented in this dissertation is the US Air Transportation System (ATS) and only information regarding US domestic flights has been used. The list of the 337 airports and 178 cities considered in this study are obtained from the AIM Project (Schäfer and Dray, 2015). Information regarding whether other transport links exist and whether a city is a major tourism and/or business attraction is also extracted from there. Such information is considered to be static through the base (2007) and projection years (2008-2025).

In terms of demographic data, the US Census Bureau (2014a) website is the source where population data was obtained for the years up to 2016. The US population statistics are generated from decennial censuses. For the in-between years, the US Census Bureau estimates population levels based on times series applied to the most recent decennial census and considering population changes due to births, deaths and migration numbers. The in-between years population estimates are updated annually. For demographic data related to period years between 2017 and 2025, please refer to Appendix A.

For the economic data (mean household income), the Bureau of Economic Analysis (2014) website has been consulted. This measure is computed by taking the personal income of the residents of a given area divided by the resident population of the same area, being the latter taken from the US Census Bureau (2014a).

Historical flight frequency data is extracted from the US Department of Transport T-100 data (2014), while historical information on passenger demand data and airfares is extracted from the Airline Origin and Destination Survey (2014) -i.e. known as DB1B, which contains a 10% sample of airline tickets from reporting carriers. Variables used to develop the itinerary choice model are also extracted from the DB1B. Flight delay information is obtained from the FAA Aviation System Performance Metrics (ASPM) database (2014) and is considered constant throughout the projection years (2008-2025). While this might be a poor assumption, especially in the short-term, it is assumed that airport capacity is added to reduce delays or maintain delays at existing levels.

Fuel prices have been extracted from the the US Bureau of Transportation Statistics (BTS) (2019), which provides with information regarding monthly fuel cost per gallon for domestic and international operations up to 2018. The monthly fuel cost for domestic operations have been used and an annual average has been calculated. Fossil fuel price projections have been obtained from the on-line database provided by the Department of Energy & Climate Change (DECC). Actuals values have been used up to 2015, while fossil fuel price projections up to 2025 have been used to generate simulations.

As mentioned in Section 3.1, the time-step considered in this study is annual. Data from 2007 is used to estimate each sub-model, while validation of each of these sub-models in isolation is done with data from 2008 to 2013. Projections using the entire modelling framework (considering interactions between the different sub-models) are generated for years from 2008 to 2025. Since models have been estimated with 2007 data, any monetary data used for different years is converted to 2007 US dollar values. This is done using the Consumer Price Index (CPI) (Bureau of Labour Statistics, US Department of Labour, 2014a). Please refer to Appendix B for further information regarding CPI and how the monetary conversion to 2007 US dollars has been computed.

3.3 Methodology Overview

As shown in Figure 3.1, the model presented in this dissertation looks at city-pair demand generation, itinerary demand assignment and air traffic estimation within a single framework. The approach presented in this thesis is the result of a comprehensive analysis of a variety of frameworks and approaches which evolved to the present work. Further information regarding those early modelling frameworks can be found in Busquets et al. (2015; 2015).

As discussed in Section 2.2.5 the work presented in this thesis centres in: further exploring the application of network theory to predict airport-pair connectivity within the US domestic network, by considering the use of input variables beyond network theory metrics as opposed to existing research (Kotegawa, 2012); and a comparison study of two modelling approaches to estimate itinerary shares, and that are used to assess the impact that network evolution would have for the air transport system. Below the classification of the set of input variables considered in this research, followed by a detailed description of each of the sub-models the modelling framework consists of.

Classification of input variables used in this research

The decision to explore a larger set of explanatory variables than typically considered is based on research showing that other inputs, apart from those typically considered such as socio-economic factors, can be important to understand the growth of the air traffic system. Examples include the research done by Evans and Schäfer (2014), which focused on simulating the operational responses of airlines to environmental constraints and showed that airline response to any type of capacity constraints and competitiveness is important when trying to understand the underlying principles behind the evolution of the air transport system; or the work done by Guimera et al. (2005) and Kotegawa (2012) which used network theory to characterise the air transport system demonstrating the potential of this area to understand network dynamics. This extended set of explanatory variables is expected to better capture the underlying behaviour that drives the aviation industry, including the underlying drivers of demand for air travel, and the underlying drivers of airline decision making to supply flights to serve this demand.

The set of input variables used in the overall modelling framework can be classified into the following three groups:

- Network theory quantitative measures. Their inclusion is based on a representation of the air transportation system using topology and mathematical graph theory, as by Kotegawa (2012) and Guimera et al. (2005) and discussed in the literature (Section 2.2.3). Given this, the ATS is represented as a natural network that consists in well-defined nodes (airports) and links (flights that connect these nodes or airports);

- Socio-economic variables, which are widely used in the literature across industry stakeholders and existing research as discussed in the literature (Section 2.1);

- Aviation-related variables. These variables characterise the trip, and hence have an influence on passenger decisions. Examples of widely used aviation-related variables include journey fare and time or presence of other modes of transports (i.e. competition) amongst others; as well as information regarding whether an airport is a hub, and whether they are located in the US mainland.

The following sub-sections describe each of the sub-models presenting the methodology used for each of them and further information regarding the corresponding set of input variables used. Each sub-section also includes the validation of the models in isolation (i.e. without considering interaction amongst the other sub-models).

3.4 Module 1: O&D passenger demand model

The first stage of the modelling framework looks at projecting passenger demand by city-pairs. This is done by using an O&D demand model that projects true origin-ultimate destination demand among a set of cities within the US Air Transportation System (ATS). For this stage a gravity-type model, similar to that used by Dray et al.(2019), is used. Passenger demand between cities o and d is calculated using a simple one-equation linear regression model of the form of

$$\begin{aligned} \ln N_{od} = \beta_0 + \beta_1 \ln(\sqrt{P_o P_d}) + \beta_2 \ln(\sqrt{I_o I_d}) + \\ + \beta_3 \ln(f_{od} + vot * t_{od} + vot_{delay} * t_{delay,od}) + \sum_i \beta_i D^i_{od} \end{aligned} \tag{3.1}$$

where N_{od} is the total passenger demand between cities o and d through any available route, P_o and P_d are the population related to cities o and d respectively, I_o and I_d are the mean household income per capita related to cities o and d respectively, f_{od} is the average fare per passenger travelling from city o to city d, vot is the value of time, t_{od} is the average travelling time between cities o and d, vot_{delay} is the value of delay time, $t_{delay,od}$ is the average delayed time that passengers encounter when travelling between the two cities, D_{od} are a set of dummy variables capturing other elements of the city-pair connection and the set of β are the parameters to be estimated. Note that the coefficients associated to generalised cost and the income variable are known as elasticities and measure the change in passenger demand as a result of changes to these economic variables, providing a key insight into the proportional impact of different economic actions and policy decisions.

Logarithmic transformation is used in both the dependent and independent variables. This resulted in more accurate forecasts, while maintaining the simplicity of the linear model (Benoit, 2011) (i.e. since OLS or WLS can be used to estimate the log-transform model.

Weighted least squares (WLS), with the number of passengers used as weights, is used as estimating process. Three sets of parameters are estimated, each one for one distance group -i.e. short-, medium- and long-haul. The distance groups have been derived by examining the distribution of distances between US cities (Figure 3.2[2]) and experimentally exploring different distance thresholds, choosing

[2]Distance distribution has been plotted considering all city-pairs with passenger demand during

the thresholds that resulted into three models with the highest predictive power. The decision of estimating three different models based on distance groups comes after investigating one single model for the entire US ATS, which resulted in low model accuracy. Consequently, one set of parameters is estimated for each of the three distance groups as per Table 3.1 below. Model specification for each of the three models is the same with the exception of the set of dummy variables included for each distance group -i.e. set of variables D in Equation 3.1. Further information regarding previous attempts to estimate a single O&D demand model is presented in Appendix C. The literature review shows that considering several distance groups when estimating O&D demand models is a common practice. Examples include the work done by Dray et al. (Dray et al., 2014), which presents an O&D demand model used within the Aviation Integrated Model (AIM) framework in which separate parameters are estimated based on different world region-pairs and distance-groups.

Distance type	**Distance Group in Miles (km)**	**Number of city-pairs in training set (undirected)**
short-haul	186 - 400 miles (300-644 km)	279
medium-haul	400 - 2113 miles (644-3400 km)	2423
long-haul	>2113 miles (>3400 km)	378

Table 3.1: Distance groups considered in O&D demand models.

2007 base-year, which adds to a total of 11670 undirected city-pairs.

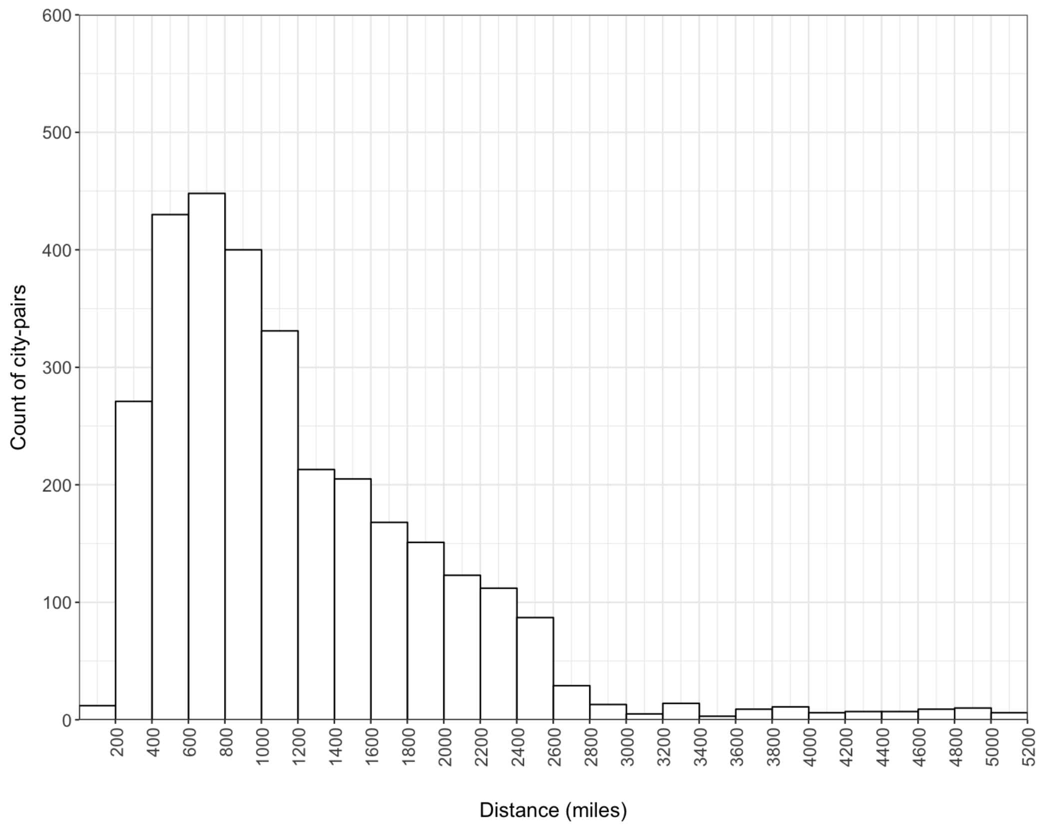

Figure 3.2: Distance distribution amongst city-pairs considered in this research.

Demand models for city-pairs that are less than 300km apart have been also explored. These O&Ds are expected to have a strong competition from other type of transport (e.g. public and/or private transport, such as rail services), and therefore are most likely to have lower demand than modelled[3]. Because the models presented here are developed at the aggregated level and applied to a large number of city-pairs, estimated models had a large error for such close O&Ds which tend to behave differently than the rest of the network, due to over-prediction. Consequently, these O&Ds have been excluded from this study. Moreover, passenger demand linked to O&Ds less than 300 km apart correspond to less than 2% of the total network passenger demand.

Based on Equation 3.1, the set of input variables for the O&D demand model can be split between socio-economic factors and aviation-related variables. In terms of socio-economic factors, in line with industry practice, population and income information have been included as input variables. Population is defined as the

[3]There is also some exceptions above 300km, such as the LAX-SFO corridor which is an O&D about 540km apart and is considered one of the busiest US domestic airline routes (Source: https://www.forbes.com/sites/ericrosen/2019/04/02/the-2019-list-of-busiest-airline-routes-in-the-world/)

number of inhabitants that the Metropolitan Statistical Areas (MSA) of the US around each airport has. The area defined in each case is sometimes the result of the combination of few MSA and/or the addition of few micropolitan Statistical Areas (mSA). In terms of economic information, the mean household income per capita for the MSA considered has been used. The inclusion of these variables amongst the input set is due to the proven relationship between economic/social growth and aviation growth. Although other socio-economic indicators also exist, this study limits their inclusion to the two variables mentioned above due to data availability.

Given that each city-pair has associated two population and income values -i.e. one associated to the origin city o and another one to the destination city d -, data transformation for these two variables is performed by applying equation 3.2 below where P_{od} refers to the population variable associated to the city-pair between cities o and d, and P_o and P_d refer to population of cities o and d respectively. Similarly, for income the same data transformation is used.

$$P_{od} = \sqrt{P_o * P_d} \tag{3.2}$$

In terms of aviation-related factors, generalised cost and a set of dummy variables have been considered. Generalised cost (GC_{od})) is referred as the cost a passenger has to incur to travel from city o to city d. It considers airfare, the value of travelling time and the value of delay time. Generalised cost is defined by Equation 3.3[4].

$$GC_{od} = f_{od} + (vot * t_{od} + vot_{delay} * t_{delay,od}) \tag{3.3}$$

Passengers value travel time and delay time differently, with delayed travel time being valuing three times more than the expected travel time as used by Evans and Schäfer (2014). Appendix B presents further information regarding the calculation of value of time. By including delay information, the generalised cost variable also captures capacity constraints of specific airports. Delay is calculated through the time-lag between the expected gate-to-gate time and the real time that a specific airport-pair route service has related. Consequently, for those air routes involving

[4]Explanation of each of the terms for the generalised cost formulation is done when Equation 3.1 is presented. Data sources for the different variables used in this equation are presented in Section 3.2.

airports or airspaces that are already operating nearly at or completely at capacity a larger delay will be associated and therefore, a larger generalised cost is expected. It is expected that, the higher generalised cost, the less likely people will be willing to travel.

As discussed in Section 3.1.4 all variables (city-pair delay, airfare and travel time) are computed as weighted average using number of passengers as weight. A summary of the available sources used to obtain information regarding the input variables is presented in Section 3.2.

As mentioned above, the three O&D demand models for the three distance groups differ in the set of dummy variables used for each distance group, as per Table 3.2. The inclusion of these dummy variables in the models is to be able to capture other elements of the city-city connection that might not be a generalised characteristic of the city-pairs within each of the models, such as whether a road or high-speed rail link exists between cities or whether one of the cities is a major tourist or business attraction.

Distance Group	Dummy variable	Symbol
short-haul	East Coast transport links	R_{EC}
	South transport links	R_{South}
	LAX-SFO / LAX-LAS	LAX
medium-haul	City attractiveness	S_1 and S_2
	Transport links	R
	Offshore territories	I_1 and I_2
	MIA-NYC holiday destination	MIA
	Presence of hub airports	h_1 and h_2
long-haul	City attractiveness	S_1 and S_2
	Offshore territories	I_1
	Presence of hub airports	h_1 and h_2

Table 3.2: Dummy variables considered for each O&D demand model.

Dummy variables included in the short-haul O&D demand model are as follows:

- East Coast transport links - indicates whether high-speed rail links exist between cities located in the East Coast, such as NYC-Boston and NYC-Philadelphia. It also includes San Antonio-Houston which has a large bus network connecting both cities at low prices (e.g. less than $10), having a negative impact on air passenger demand, resulting in lower air demand for this O&D due to ground transport competition.

- South transport links - indicates whether rail network exists amongst cities located in Florida[5].
- LAX-SFO and LAX-LAS - these city-pairs show a specially high demand although they are less than 400 miles apart. The market between Los Angeles and the Bay area is one of the largest O&D market in the US[6].

Dummy variables included in the medium-haul O&D demand model are described as follows:

- City attractiveness - indicates whether one or both cities in the pair are major tourism or business destinations.
- Transport links - indicates whether high-speed rail links exist between cities that fall within this distance group.
- Offshore territories - indicates whether one or both cities in the pair are located in one of the US offshore territories or Alaska.
- MIA-NYC - similarly to LAX-SFO for short-haul, MIA-NYC corridor shows a much larger demand compared to the majority of city-pairs within this group. This is believed to be due the leisure attractiveness that Miami has as a city[7].
- Presence of hub airports - two dummy variables indicating whether zero, one or two hub airports are present on a O&D market.

Thirdly, dummy variables included in the long-haul O&D demand model are described as follows:

- City attractiveness - indicates whether none or one of the cities in the pair is a major tourism or business destination.
- Offshore territories - indicates whether one city in the pair is located in one of the US offshore territories or Alaska.

[5] Amtrak runs the South Train Routes (https://www.amtrak.com/regions/south.html) with Brightline working with Virgin Train USA having launched services across Florida last year (https://www.virgin.com/richard-branson/introducing-virgin-trains-usa)

[6] Source: https://www.forbes.com/sites/ericrosen/2019/04/02/the-2019-list-of-busiest-airline-routes-in-the-world/

[7] This route is also known as 'snowbird' route from the many retirees who spend winter in Miami.

- Presence of hub airports - the same as for the medium-haul demand model.

3.4.1 O&D demand models estimated results

To train the models, 2007 data has been used. During the estimation process city-pair's direction of travel is not considered - i.e. city-pair between cities o and d is considered the same as city-pair between cities d and o. Consequently, input variables have also been defined by city-pair without considering direction - e.g. city-pair airfare and flight time (included in the definition of generalised cost) are computed as the weighted average of airfare and flight time of both directions using passenger numbers as weight. The observed passenger demand for a specific city-pair, which is used to compare the predicted passenger demand with, has also been computed as a weighted average of passenger demand from both directions. The consideration of directed city-pairs have been also explored, however, better results emerge when model estimation process is done considering undirected city-pairs, which is expected since the model does not include factors that account for directionality. The total number of undirected pairs of cities used during training is 3,080.

Validation of the O&D demand model, estimated using 2007 data, is done using actual data from the individual years between 2008 to 2013. Estimated model results for the three O&D demand models are presented in Table 3.3. All estimated coefficients are statistically significant at the 95% confidence level and are of the expected sign. Adjusted R^2 during training across the 3 models falls between 0.84 and 0.91. This is a significant improvement compared to the results obtained when a single set of coefficients covering all distances were estimated, in which the adjusted R^2 obtained was 0.61[8]. Estimated coefficients are also of the expected sign, with all coefficients related to other transport modes competition negative (i.e. R_{EC}, R_{South}, R). It is interesting to see how the set of dummy variables for the medium- and long-haul models indicating whether zero, one or two airports are present on an O&D-pair have opposite sign -i.e. h_1 is 1 when there is at least a hub airport in each of the cities forming the city-pair, 0 otherwise; h_2 is 1 when none of the cities have a hub airport, 0 otherwise. This might be caused by the fact that long-haul markets tend to be one-stop itineraries between two small cities, which usually do

[8] A briefly summary of early modelling attempts can be found in Appendix C.

not have a hub airport[9].

	short-haul	medium-haul	long-haul
(Intercept)	11.43 ***	13.29 ***	11.34 ***
Population	0.94 ***	1.18 ***	1.47 ***
Income	2.35 ***	1.04 ***	1.28 ***
Generalised Cost	−2.45 ***	−2.56 ***	−2.91 ***
S_1	0.67 ***	0.55 ***	
S_2	−0.93 ***	−0.74 ***	−0.79 ***
R_{EC}	−0.63 ***		
R_{South}	−1.27 ***		
LAX	0.63 ***		
R		−0.52 ***	
I_1		0.84 ***	1.28 ***
I_2		−2.07 ***	
MIA		1.01 ***	
h_1		0.22 ***	−0.50 ***
h_2		−0.14 ***	0.25 *
Adj. R^2	0.91	0.84	0.88
RMSE	151.2	142.8	130.3

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.3: Coefficient estimates obtained for the O&D demand models by distance group.

From results presented in Table 3.3, income and generalised cost elasticities can be compared to those in the literature. In 2007, Intervistas Consulting Inc. (2007) prepared a report for IATA, in which they studied air travel demand elasticities. From this report the following is concluded:

- Generalised cost elasticity - Generalised cost elasticities obtained in this study range from -2.45 to -2.91, which are comparable to those obtained by Dray et al. (2014) for the Intra-North America O&D demand models (i.e. ranging between -2.04 to -2.61 for the different distance groups). Also, results show passengers in long-haul trips to be more fare-sensitive than on short-haul ones, which is

[9]Note that for this study the consideration of only 25 large hub airports has been used, and therefore the inclusion of mid- and small- hub airports could potentially have an impact on the results.

expected as there is typically a higher proportion of business passengers on short-haul trips, which are less fare-sensitive (2017, 2007). A more popular measure used to evaluate the impact in passenger demand due to a change in fares is fare-only elasticities which are typically slightly less negative as stated by Dray et al. (2014). At the route/market level, airfare elasticities for the US domestic network should be in the range of -1.2 to -1.5.

- Income elasticity - Income elasticities obtained in this study range from +1.04 to +2.35. The analysis performed by IATA (2007) found a wide range of income elasticities in the literature, generally between +1.0 to +2.0, meaning that air travel is generally found to be income elastic. Their analysis also produced mixed results and stated that results of the models using US data generated income elasticities between +1.6 and +1.8. These reference values are above to what it is obtained in this research for the medium and long haul - i.e. +1.04 and +1.28 respectively -. However, IATA (2007) report also states that results are not robust and that any small change in model specification could result in sizeable changes in these values. Nevertheless, income elasticities for the three distance groups obtained in this research are positive as expected from the literature review.

All models are also tested for multicollinearity. Multicollinearity, also known as collinearity or ill-conditioned data, occurs when two or more variables are highly, but not perfectly related, correlated with each other. This effect will make impossible to define or interpret the regression coefficients as every time a given change to one of the correlated variable occurs, the corresponding observation on other highly correlated variables is likely to change in a similar manner. Consequently, it will be impossible to measure any change on y (the independent variable) solely due to the change of one of the correlated variables, while keeping the rest of variables equal.

The existence of multicollinearity can be discovered through several indicators, such a relative high R^2 value along with a few low t-statistics, or a high F-statistic while none of the coefficients are significant at the 95% confidence level (i.e. low t-statistics of the estimated parameters).

A formal test for multicollinearity is the calculation of a condition number associated with the input variables, such as the Variance Inflation Factor (VIF). A

VIF value of more than 20 indicates the presence of multicollinearity. VIF for a given explanatory variable i (VIF_i) is computed through Equation 3.4, where R^2 is the multiple correlation coefficient of X_i regressed on the remaining explanatory variables.

$$VIF_i = \frac{1}{1 - R^2} \tag{3.4}$$

VIF values obtained across the three O&D models are lower than 5, hence confirming absence of multicollinearity.

3.4.2 O&D demand models validation

Validation of the O&D demand models estimated with 2007 data is performed by using annual data from years 2008 to 2013. By applying Equation 3.1 for each of the corresponding distance groups, estimated number of passengers for the subset of city-pairs in each distance group are generated. For each validation year, predicted number of passengers are compared to observed ones in Figures 3.3, 3.4 and 3.5, showing results for short, medium and long haul O&D demand models respectively. Also, the total estimated passenger demand is compared to the total number of passenger demand for each of the distance group O&D demand models as shown in Figures 3.6, 3.7 and 3.8.

Overall, the models seem to have a good fit and although a similar trend to those observed can be spotted some differences arise across the three distance group models. For example, while the medium- and long- haul models seem to over-predict passenger volumes in 2009 with an increase with respect to 2008; the short-haul model captures the observed smooth declining trend for that year. This can be explained by the fact that the economic recession led to a decrease in airfares, and therefore generalised cost, prompting the model to generate over-predictions due to the negative generalised cost elasticity. Secondly, estimated model results from the three distance groups (Table 3.3) show that short-haul markets are slightly less inelastic to airfare changes, and hence the decline on airfares would have less influence in short-haul passengers, which is expected as there is typically a higher proportion of business passengers in this distance group, than those in long-haul trips. At last, a further check is performed to verify whether airfares have changed differently across the three distance groups, which is presented in Figure 3.9 confirming the more steep decline in airfares for the mid- and long-haul distance groups. Figure 3.9 verifies that the combination of a larger decrease in airfares for the long and mid-haul markets along with a more negative generalised cost elasticity these distance group has, prompted the models to perform differently for this particular year (i.e. 2009).

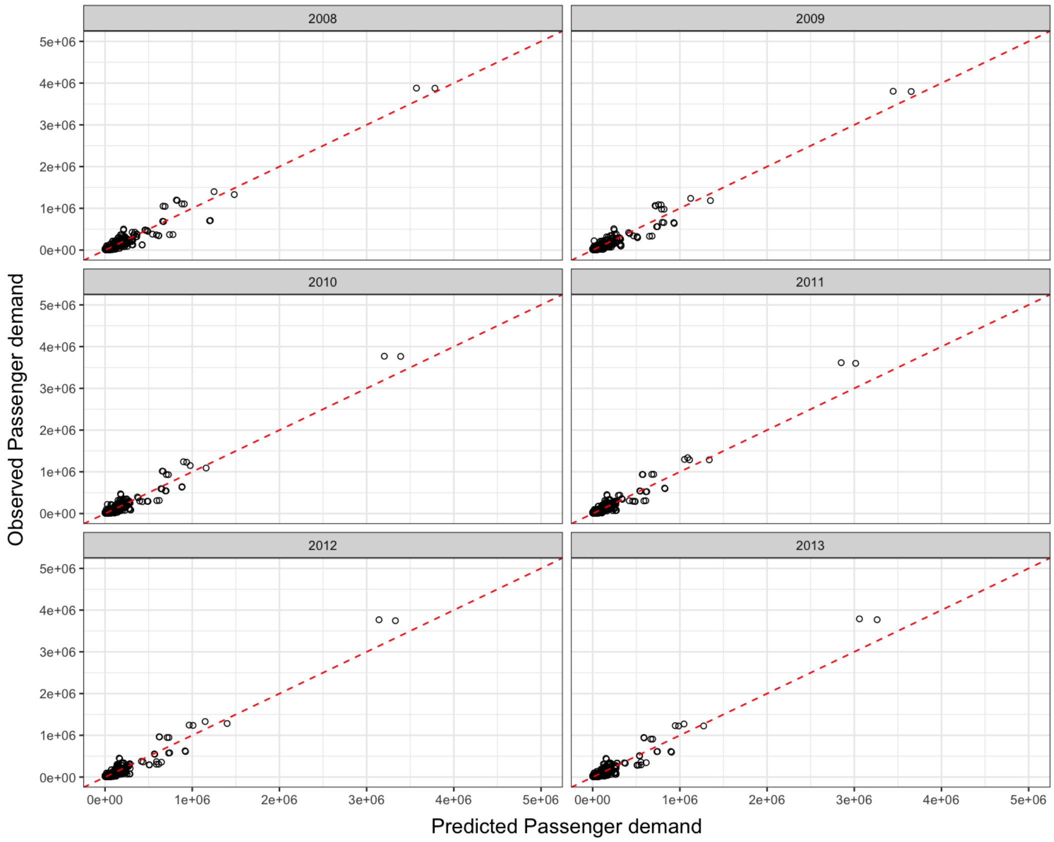

Figure 3.3: Observed against predicted passenger demand for short-haul city-pairs.

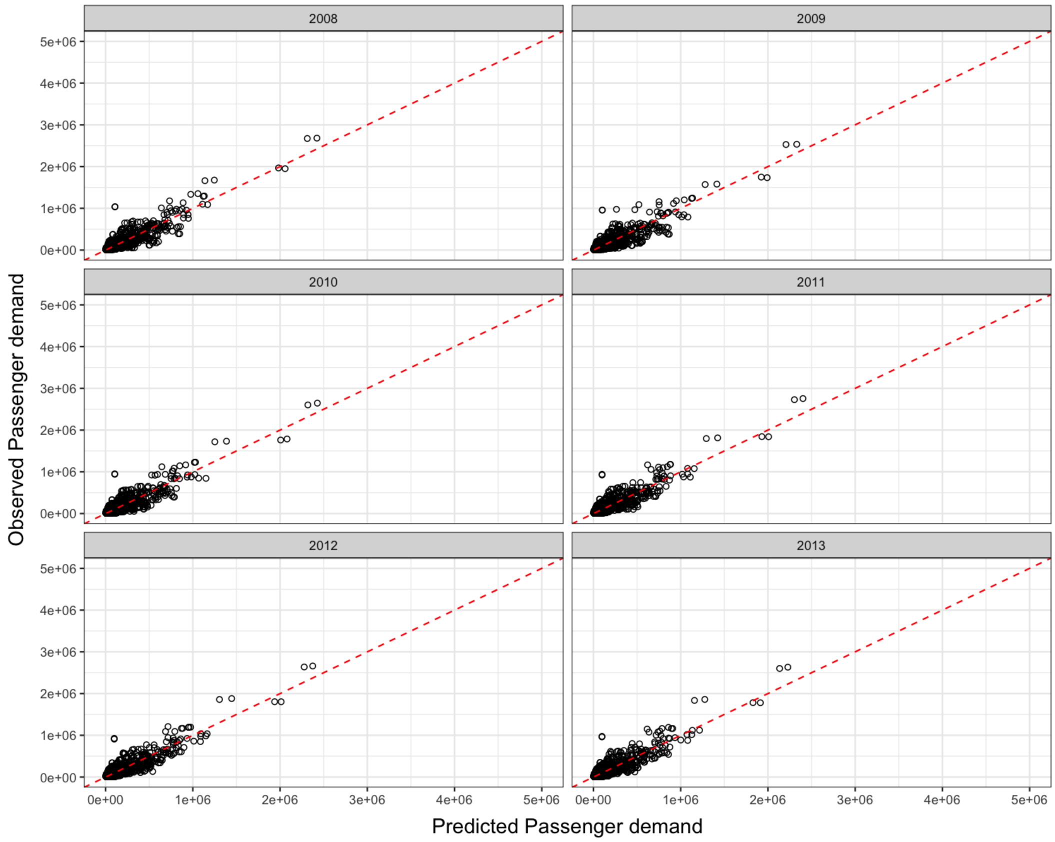

Figure 3.4: Observed against predicted passenger demand for medium-haul city-pairs.

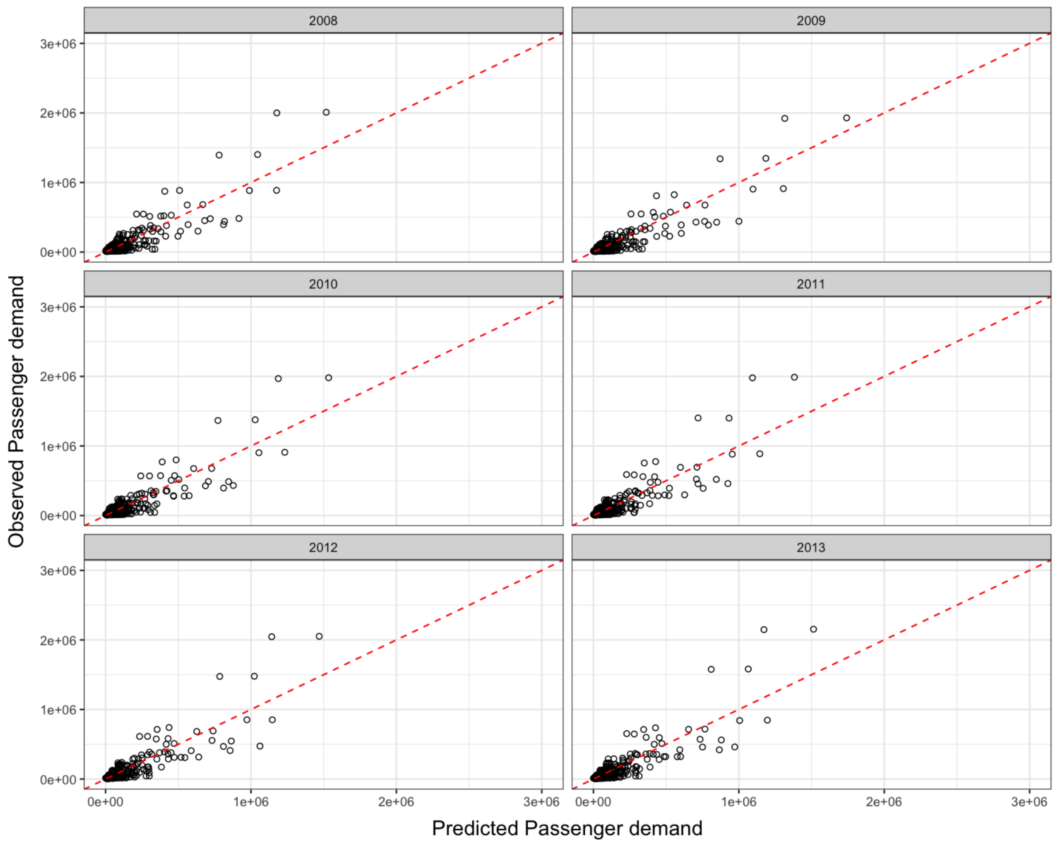

Figure 3.5: Observed against predicted passenger demand for long-haul city-pairs.

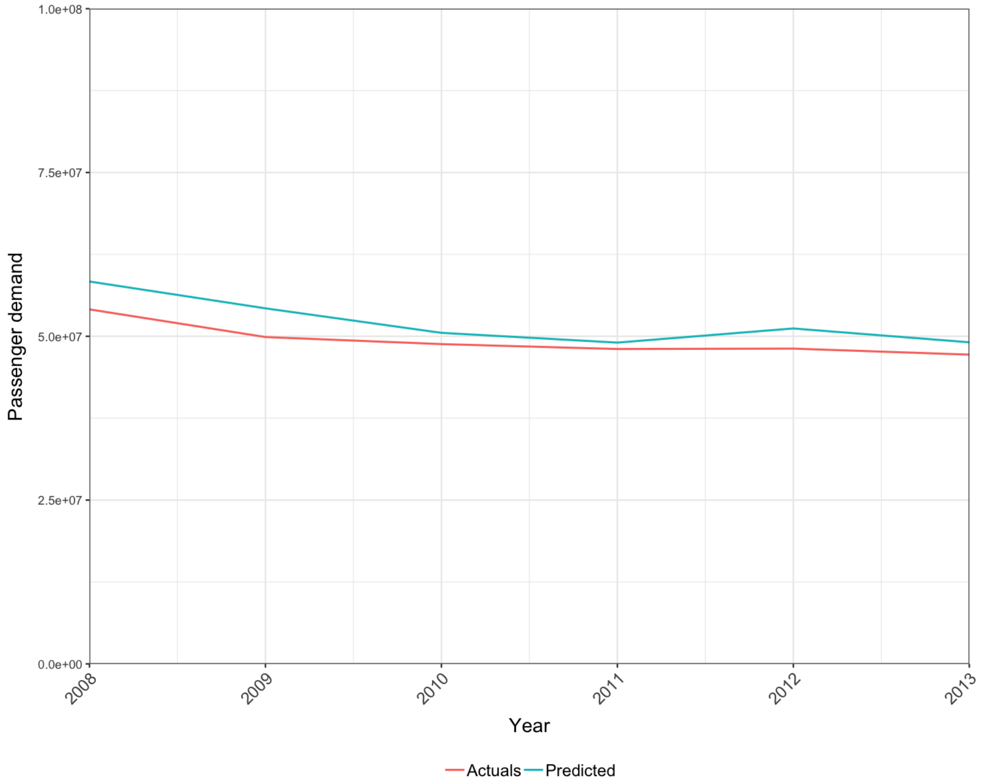

Figure 3.6: Observed and predicted total number of passengers throughout validation years (i.e. 2008-2013) for short-haul city-pairs

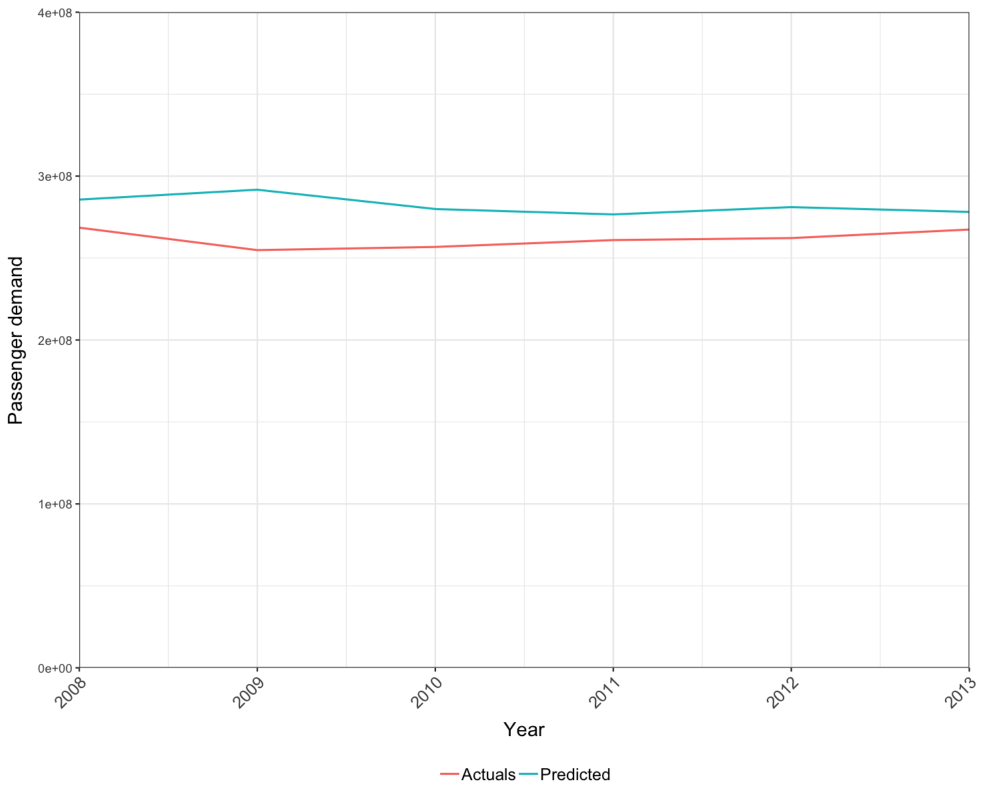

Figure 3.7: Observed and predicted total number of passengers throughout validation years (i.e. 2008-2013) for medium-haul city-pairs

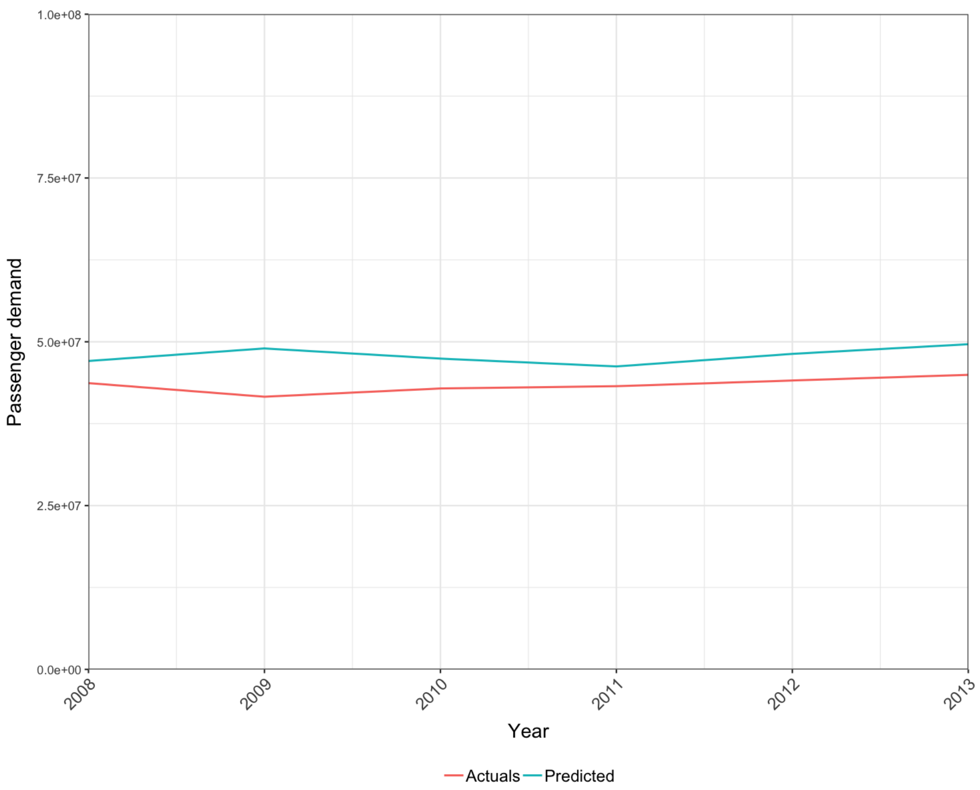

Figure 3.8: Observed and predicted total number of passengers throughout validation years (i.e. 2008-2013) for long-haul city-pairs

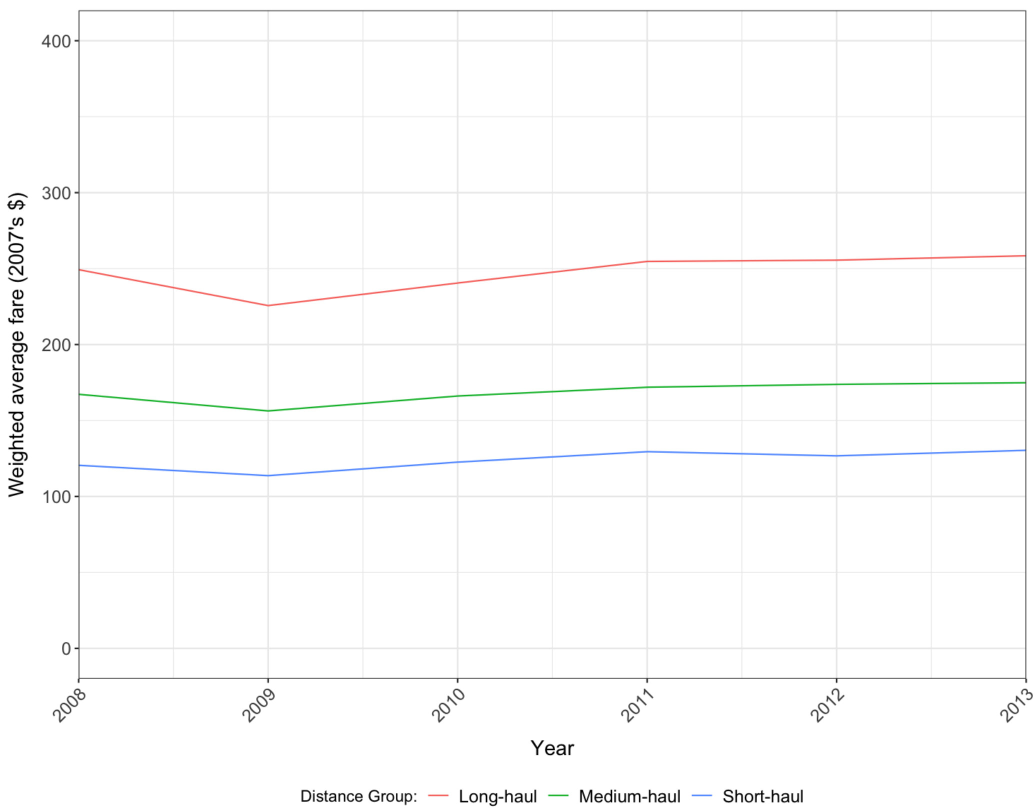

Figure 3.9: Weighted average fare by distance group at the market level.

System-wide passenger demand for the validation years is computed by aggregating the predictions obtained for each of the three models. Figure 3.10 presents the comparison between observed and predicted number of passengers throughout validation years (i.e. 2008-2013) for all city-pairs considered in this study. Figure 3.11 presents the observed and predicted system-wide total number of passengers throughout the validations years.

Table 3.4 shows the percentage error between predicted and observed total passenger demand for the different distance groups and at the aggregated level throughout the validation years (i.e. 2008-2013). Percentage error range from 2% to 10.6% with the exception of 2009 data for medium-, long-haul O&D demand models and at the aggregate level. As explained earlier, this can be explained by the economic crisis taking place during 2009 and that lowered market airfares. Generalised cost has a negative impact on passenger demand (i.e. as the model estimated results show in Table 3.3 and specified in the literature (Dray et al., 2014)). Although the economic recession also led to lower income levels, and therefore one would expect a decrease in passenger demand, the reduction of fares, and therefore generalised cost, has initially a greater impact on passenger demand. As seen in Figure 3.12 shows, in 2009 weighted average airfares lowered in comparison to the rest of validation years, prompting the models to estimate higher passenger volumes for 2009. Another aspect to consider is that the models slightly over-predict especially low demand O&Ds and particularly for the the long-haul markets (Figure 3.5), which is expected because given the way the model is formulated demand cannot go below 0, i.e. a city-pair with yearly demand of 200 can only under-predict by 200, but can over-predict by an unlimited amount.

Year	short-haul	medium-haul	long-haul	system-wide
2008	7.8%	6.3%	7.7%	6.8%
2009	8.8%	14.5%	17.7%	14%
2010	3.5%	9.0%	10.6%	8.4%
2011	2.0%	6.0%	7.0%	5.6%
2012	6.3%	7.2%	9.2%	7.3%
2013	4.0%	4.0%	10.4%	4.8%

Table 3.4: Percentage error between predicted and observed total passenger demand for the different distance groups and at the aggregate level throughout the validation years (i.e. 2008-2013).

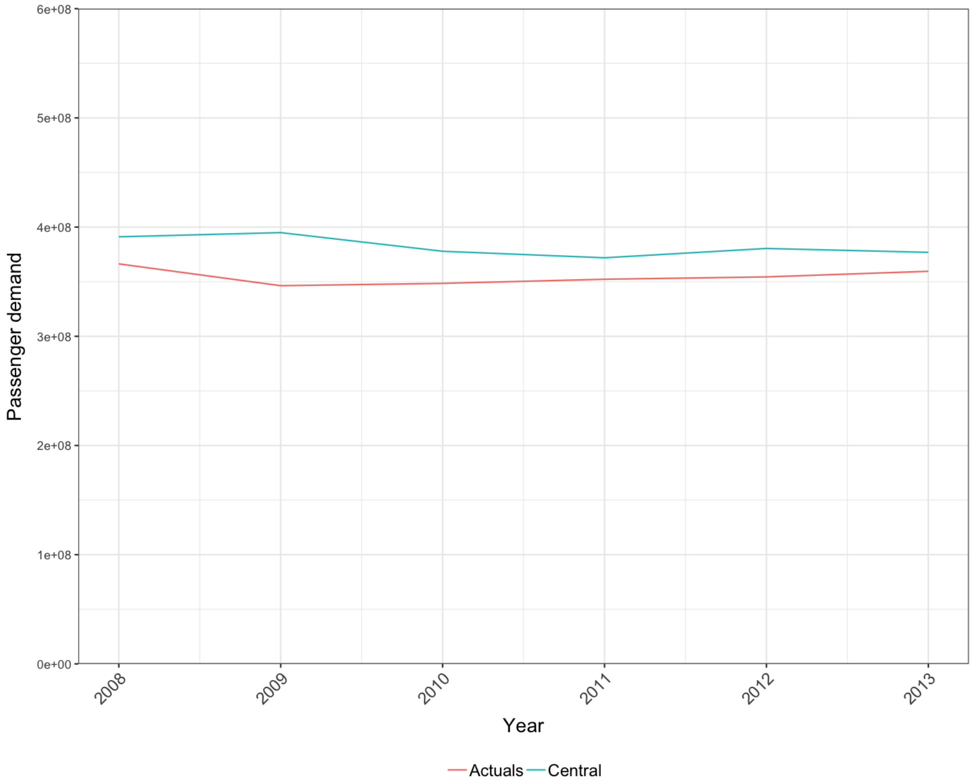

Figure 3.10: Observed and predicted number of passengers throughout validation years (i.e. 2008-2013) for all city-pairs considered in this study.

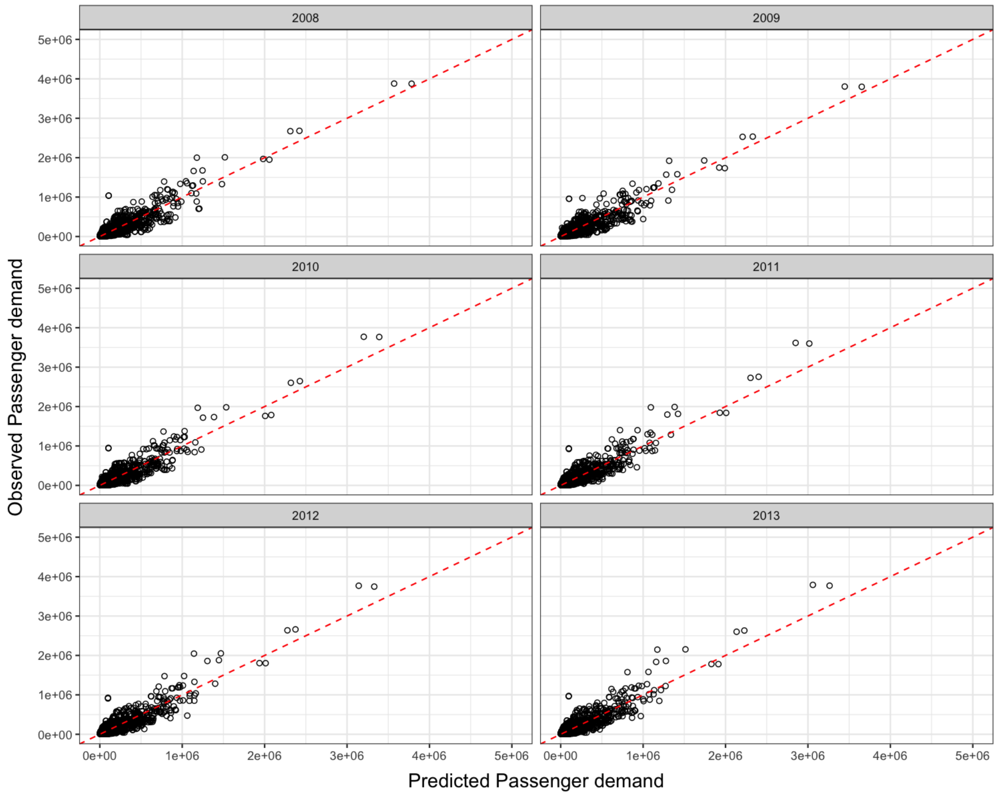

Figure 3.11: Observed and predicted system-wide total number of passengers throughout validation years (i.e. 2008-2013).

For each of the validation years adjusted R^2 is computed when considering the aggregated results (i.e. predicted system-wide passenger demand) (Table 3.5). As a comparison, adjusted R^2 obtained from a previous attempt in which a single O&D passenger demand model was developed are also included in Table 3.5 to demonstrate the significant improvement achieved by splitting the problem by distance groups. Adjusted R^2 obtained during training and validation are also higher that some of those found in the literature (Dray et al., 2014), although model specification are not like for like.

Model	2008	2009	2010	2011	2012	2013
3- distance group model	0.833	0.833	0.827	0.814	0.830	0.804
Single model	0.525	0.531	0.537	0.537	0.538	0.536

Table 3.5: Adjusted R^2 obtained across the validation years (2008-2013) when considering the aggregated results from the 3 O&D models developed by distance group. For comparison, adjusted R^2 from a previous attempt is also included.

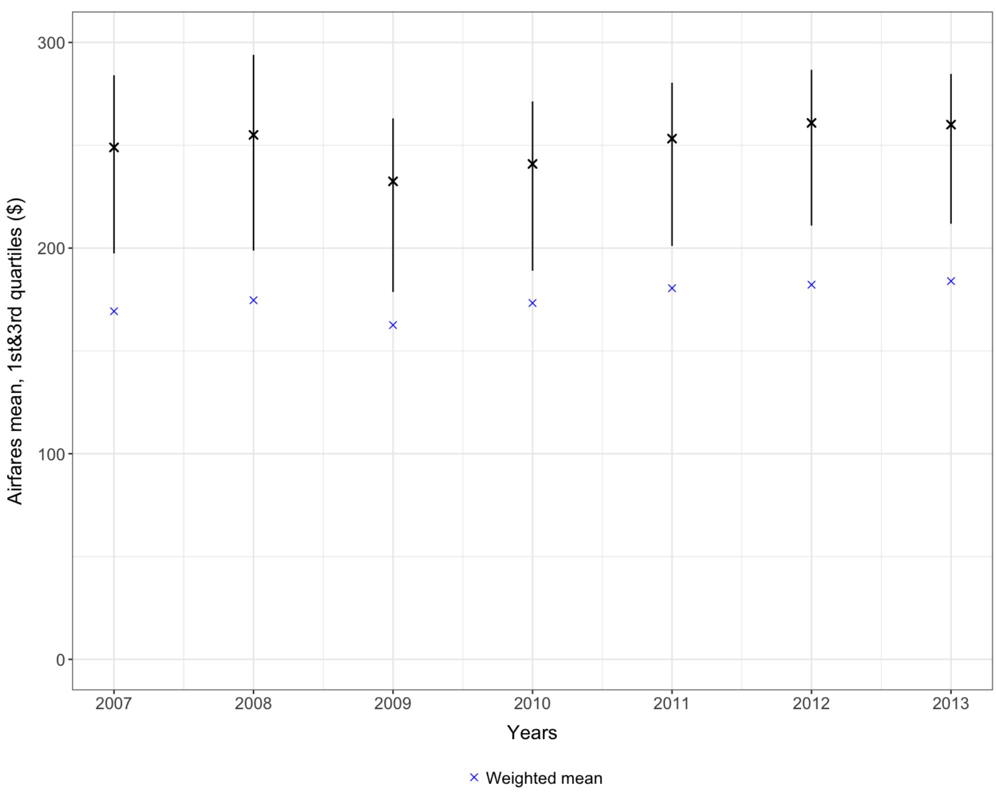

Figure 3.12: Average and weighted average fares for the US domestic network from 2007 to 2013.

3.5 Module 2: Airport connectivity model and itinerary choice model

The second stage of the modelling framework looks at modelling the distribution of passenger demand between city-pairs across the available itineraries serving each city-pair. This stage is split into two sections. First, airport connectivity is analysed so that all available itineraries between any given city-pair can be compiled. Second, once all available itineraries are compiled, an air itinerary choice model is applied so that passenger demand between city-pairs can be distributed across the available itineraries (i.e. routes) serving each city-pair. Both these models are discussed below.

3.6 Module 2: Airport connectivity model

The aim of the airport connectivity model is to estimate whether an airport-pair would be connected by a scheduled flight in the future, with the results of this connectivity model used to compile all available itineraries that can serve any specific O&D. Studying airport connectivity allows to study the evolution of the network, since it assumes that the US ATS is not static and therefore airport-links can be added or removed to/from the network.

The work presented in this dissertation regarding airport connectivity is inspired by the work done by Kotegawa (2012) as discussed in the literature review (Chapter 2.2.3). Kotegawa (2012) developed a network restructuring model using a combination of Support Vector Machine and logistic regression as modelling techniques and only considered network theory variables as input variables. Although Kotegawa's work (2012) was a leading research in the field, accuracy of the models ranged between 20% and 40%. Taking his work as inspiration, this dissertation explores the application of a classification model to study the evolution of the US air transport network using input variables that go beyond network theory metrics.

In order to predict whether an airport-pair would change their connectivity status -i.e. whether they will be removed/added from/to the network- a classification model is used, in particular a logistic regression model (explained in more detail below). This will predict the likelihood of previously connected airport-pairs being disconnected in the future, as well as the likelihood of unconnected airport-pairs being connected in the future. These input variables can be split into two groups: network theory metrics; and aviation-related metrics.

3.6.1 Network Theory metrics to study airport connectivity within the US

Network theory is the study of complex interacting systems which are usually represented by graphs that model the reality of the network. A graph G, representing a network, consists of a set of nodes N, a set of links L and a mapping function f, which maps links into pairs of nodes (Lewis, 2009). The complex theory bases its study in the relationships among the social structure represented by G when a time-varying element is considered. Therefore, it studies the dynamics of the network that affect its behaviour.

The air transportation system fits perfectly into this description. When using network theory to represent the air transport system, nodes are represented by airports and links are represented by aircraft operations between airports[10]. The study of the nodal properties defines what is known as the topology of the network, which helps to understand the network dynamics affecting its behaviour. In the air transportation system, these nodal properties characterise airports (i.e. nodes) in terms of importance within the system, so that one would expect to have an impact towards airport-pair connectivity. For estimation purposes, nodal properties are computed using data from the previous year to the one being estimated.

In network theory, the full representation of the air transport system is done through the adjacent matrix A and its weighted corresponding version (A^w). Elements for the adjacent matrix are binary - i.e. '1' if an airport-pair is connected by a flight, '0' otherwise-; whereas its weighted version can take any natural number - for example, number of aircraft operations between two airports. Practically, the adjacent matrix is an indication of which airport-pairs are connected by air; while the weighted adjacent matrix is an indication of the air traffic levels between those connected airport-pairs.

For clarification purposes, given a set of 3 airports (e.g. JFK, LAX and IAD),

[10]For future reference airport and node are used indistinctly in the rest of the thesis. Similarly, the use of 'links' are also used to indicate airport-pairs connectivity.

adjacent matrix (A) below indicates that flights operate between JFK and LAX (and vice-versa[11]) and LAX-IAD; whereas JFK-IAD is not connected (i.e. there are no airlines operating this route). If only considering the adjacent matrix, only information regarding the number of connections an airport can be known. The weighted adjacent matrix gives an idea of the frequency in which those operations occur. The weighted adjacent matrix (A^w) below shows that number of operations between JFK and LAX doubles those between LAX and IAD, prompting the belief that JFK-LAX is a route with greater passenger demand than LAX-IAD.

$$A = \begin{array}{c} \\ JFK \\ LAX \\ IAD \end{array} \begin{array}{c} \begin{array}{ccc} JFK & LAX & IAD \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

$$A^w = \begin{array}{c} \\ JFK \\ LAX \\ IAD \end{array} \begin{array}{c} \begin{array}{ccc} JFK & LAX & IAD \end{array} \\ \begin{pmatrix} 0 & 100 & 0 \\ 100 & 0 & 50 \\ 0 & 50 & 0 \end{pmatrix} \end{array}$$

From all possible nodal properties that exist, node degree, node weight, eigenvector centrality and clustering coefficient have been chosen for in this research. This choice is made taking into account Kotegawa's research (2012), in which those four topological parameters gave the best results when used as input data for forecasting connectivity likelihood among airport-pairs.

The degree of a node (k) gives a sense of importance and it is computed by counting the total number of connections a node has with other nodes. Given the assumption of undirected network, the node degree is calculated as described in Equation 3.5.

$$k_i = \sum_j A_{ij} \tag{3.5}$$

where k_i refers to the node degree of airport i, A is the un-weighted adjacent matrix and j refers to the rest of airports apart from i that exist in the air transport system. Considering the example above with the 3 airport set, JFK degree is 1 ($k_{JFK} = 1$)

[11]Note, the US air transport system is assumed to be an undirected network; therefore if a connection exist between airports i and j, a connection will also exist between j and i.

because it is only connected to LAX; whereas LAX degree is 2 ($k_{LAX} = 2$) because is connected to JFK and IAD.

Although node degree refers to node importance it does not provide with complete information about the role of a node since it makes no differentiation between two nodes with the same number of connections. In order to address this deficit, node weight is computed. Node weight s gives an idea of the role of that node within the network quantifying also its importance. For the case of the air transport system, the node weight accounts for the total number of flights operated from a specific node among all its connections with the rest of the nodes. It is computed by adding all the columns entries through a given row in the weighted adjacent matrix as described in Equation 3.6 .

$$s_i = \sum_j A^w_{ij} \tag{3.6}$$

where s_i refers to the node weight of airport i, A^w refers to the weighted adjacent matrix so that A^w_{ij} corresponds to the number of operations performed between airport i and j; and j refers to the rest of airports that exist in the air transport system. In this case, JFK weight is 100 ($s_{JFK} = 100$), as the sum of flights amongst all the routes departing from JFK; whereas LAX weight is 150 ($s_{LAX} = 150$), 100 from LAX-JFK route and 50 from LAX-IAD. This example clearly shows the complementary information the adjacent matrix adds when studying the network. Looking only at the adjacent matrix, one could conclude that JFK and IAD are airports with the same level of importance since both have only 1 connection. However, weighted adjacent matrix shows how JFK's passenger demand doubles that of IAD, and therefore JFK has a more influential role in the network than IAD.

Measures such as node degree and node weight give a sense of how well connected an airport is. This is an effect of how many airlines operate in it. An airport in which a higher number of airlines are operating will most likely serve a higher number of routes. Consequently, this will reflect on higher values for these airport's topological properties, its degree and weight.

So far the nodal properties presented focus on the properties of the network determined individually by the node itself. However, nodal properties can be also computed at a collective level by taking into account the importance of the neighbouring nodes, hence, their influence in the network. The two nodal properties computed at a collective level are eigenvalue centrality and clustering coefficient.

Considering the air transport system, eigenvector centrality (EVC) assumes that the importance or popularity of an airport is proportional to the sum of centralities of the neighbouring airports to which is connected to -i.e. EVC assumes that an airport's importance is not only based on its own connections but also on the number of connections of the airports it connects to. Consequently, an airport, whose connections have a high node degree, will have a higher eigenvalue centrality than another airport whose connections have low node degree. This concept supposes that centrality of node i is represented by a linear combination of its connecting nodes' degree and their respective weights. Equation 3.7 shows the mathematical expression for eigenvector centrality where λ is a constant. The compact form of expression 3.7 is $Ax = x\lambda$, which is equivalent to the familiar problem of finding eigenvectors, where x refers to the eigenvector, λ to the eigenvalue and A to the adjacent matrix.

$$x_i = \lambda^{-1} \sum_j A_{ij} x_j \tag{3.7}$$

where x_i refers to the eigenvector centrality[12] of airport i and j refers to the set of other airports in the network.

Clustering coefficient (CC) is the last topological metric considered in this dissertation. CC is a measure that quantifies how many times a given node forms triangular sub-graphs with their adjacent nodes. In other words, it is a measure of local cliquishness of the network, giving an idea of the node robustness. The higher the CC of a given node is, the more robust this node is since a higher number of alternative paths are more likely to exist if any of the existing links fail. CC is defined as expression 3.8 as the number of triangles centred on node i divided by the number of triples centred on that node.

$$CC_i = \frac{1}{k_i * (k_i - 1)} \sum_{j,k} A_{ij} A_{ik} A_{jk} \tag{3.8}$$

where CC_i is the cluster coefficient of node i, k_i refers to the degree of node i, A refers to the un-weighted adjacent matrix and j and k are the rest of the nodes in

[12]Note that to follow a naming convention and avoid confusion with reference to input variables, EVC will be used to refer to eigenvector centrality in this dissertation.

the network that are checked to form triangles centred on node i.

When computing nodal properties associated to a given airport-pair to be used as explanatory variables, two specifications for these variables are used. The two specifications are as follows: nodal properties have been input into the model as two separate variables, one for each of the edge nodes forming a given airport-pair; or a combined variable has been computed. The data transformation used to compute the combined nodal property associated to airport-pair between airports i and j is defined by Equation 3.9 using degree (k) as an example[13].

$$k_{ij} = \sqrt{k_i * k_j} \tag{3.9}$$

3.6.2 Example application of network theory to characterise the US Air Transport System

Before investigating the use of network theory metrics to study the evolution of the US air transport network, the analysis of the topological properties of the US airport transportation system is presented below. This analysis is done considering the four nodal properties used in this research (i.e. node degree, node weight, clustering coefficient and eigenvector centrality). In this study, the US ATS network is considered an undirected network - i.e. links have no direction associated to them.

Figure 3.13 shows the degree distribution of the US ATS[14]. Figure 3.13 clearly shows the US Air Transportation network following a power-law degree distribution, by which most of the nodes have only few links while only few of the nodes have a high degree. This is a clear representation of a hub-and-spoke network where the majority of airports are small connecting to a low number of other airports; and very few large airports acting as major hubs connect many smaller airports among them. From this evidence one could suggest that the US ATS is a scale-free network[15].

When looking at eigenvector centrality (EVC) and node weight (s_i), Figures 3.14 and 3.15 show the relationship between node degree and node weight and eigen-

[13] Equation 3.9 shows the combined degree term between airports i and j for explanation purposes, however the same expression is applied to the rest of network theory variables.

[14] Node degree (k_i) is defined as the count of links that a given node has and a single link implies one return flight.

[15] A scale-free network is characterised by links following a power-law distribution, by which some nodes have a very high number of links while the rest of nodes have a low number of links (http://eaton.math.rpi.edu/csums/papers/FoodWebs/barabasisciam.pdf).

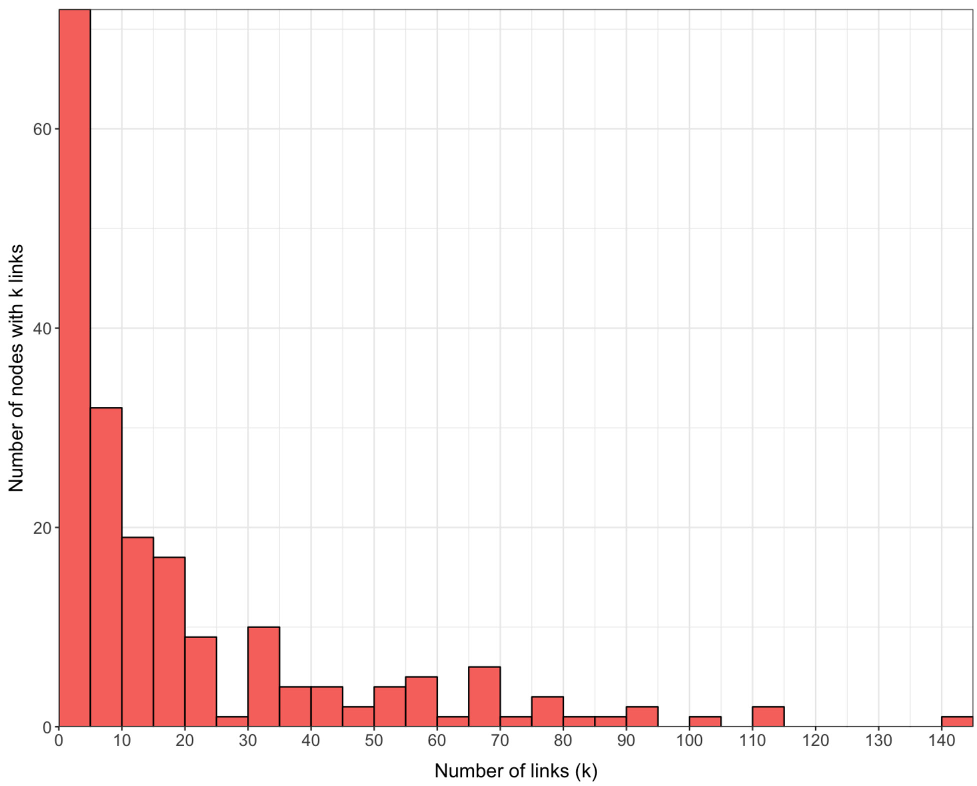

Figure 3.13: Degree distribution.

vector centrality respectively. As expected, the greater the degree of an airport is, the greater its weight is. The same positive relationship can be seen between eigenvector centrality and degree of an airport. Weight of an airport is computed through the weighted adjacent matrix (A^w), which refers to the number of flights going outwards from a given airport. As Figure 3.14 shows the more connections a given airport has the larger number of flights can be expected to be operating from that airport. As per *EVC*, an airport whose connections are in turn highly connected to many other airports will incur a high eigenvector centrality. Therefore, the importance of an airport is not only based on its degree but on the degree of the airports that is connected to. Similarly to weight, one could expect that the higher the degree of a node, the higher the *EVC*.

In contrast to the positive relationship between node degree with node weight and *EVC*, *CC* seems to have the opposite behaviour as shown in Figure 3.16. To understand *CC* better, a walk-through example is presented below using Rick Husband Amarillo International airport (i.e. IATA code is AMA) as node i. *CC* refers to the number of triangles formed by a given node and its respective connecting

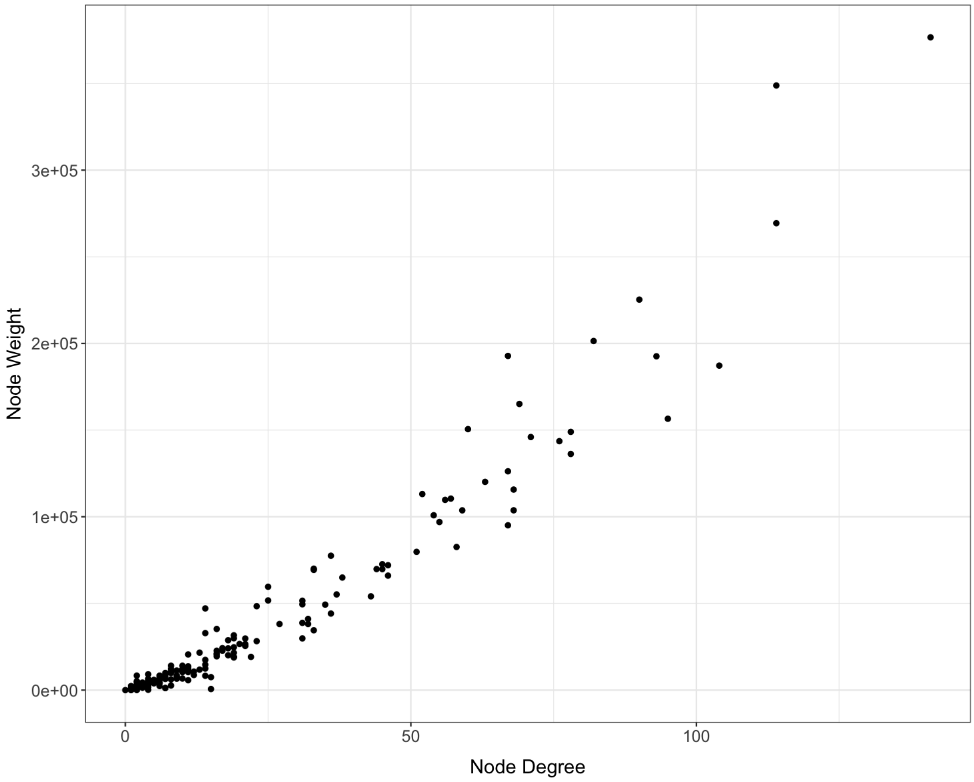

Figure 3.14: Node weight (s) against node degree (k) .

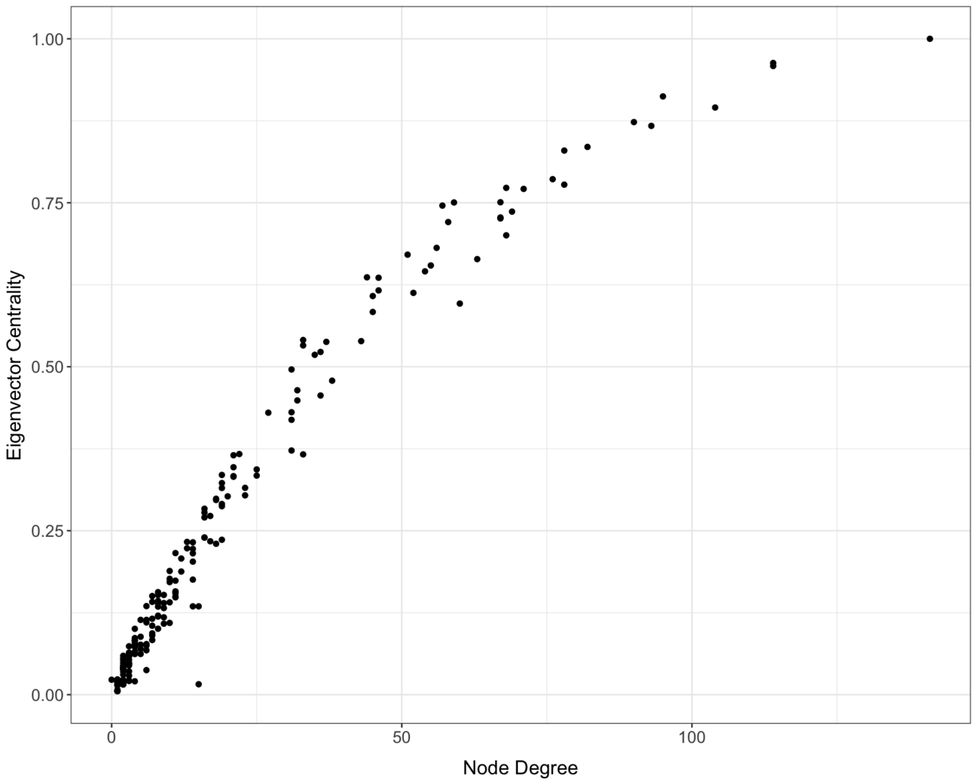

Figure 3.15: Eigenvector centrality (EVC) against node degree (k).

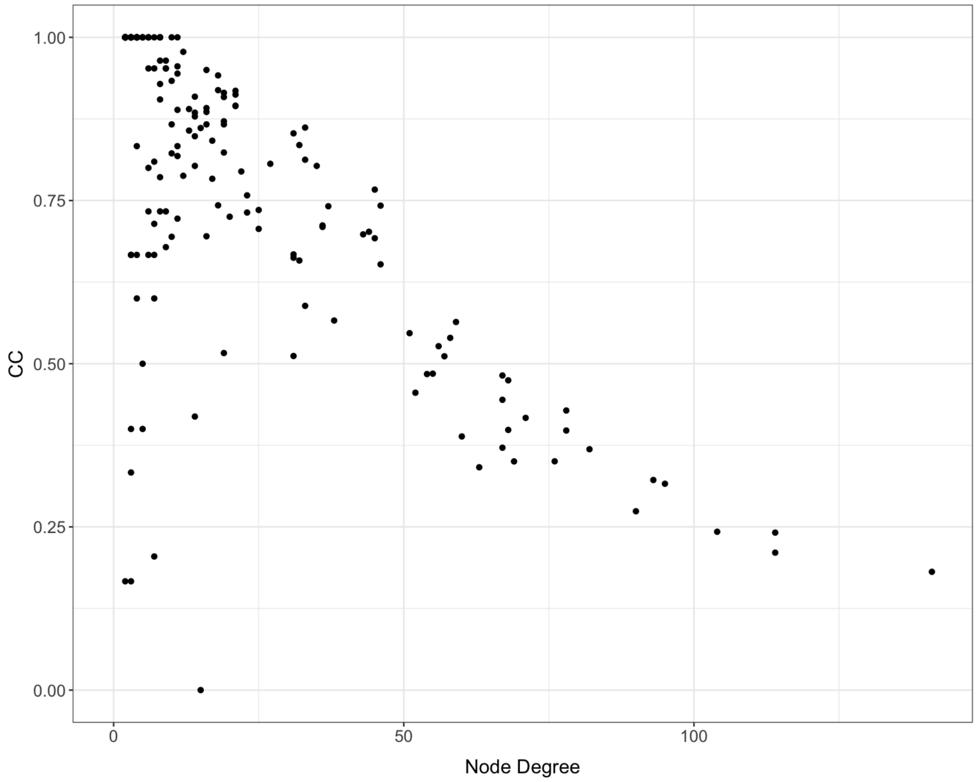

Figure 3.16: Clustering coefficient against node degree for all the nodes considered in this study.

nodes divided by the total number of possible triangles that could be formed. CC captures the degree to which the neighbours of a given node link to each other. The more densely interconnected the neighbourhood of a node is, the higher its local clustering coefficient is.

Taking AMA airport as node example, data from 2007 shows five links exist between AMA and other airports (Table 3.6), therefore degree of AMA airport is 5 (i.e. $k_{AMA} = 5$). The numbers of possible triangles that AMA could form with its neighbouring nodes -i.e. if all connected nodes that are linked to AMA airport would at the same time be connected to each other - is computed by applying Equation 3.10. This results in 20 possible triangles.

$$\text{Num triangles}_i = k_i * (k_i - 1) \tag{3.10}$$

This is represented by the adjacent matrix A_{AMAj}, where j is the set of airports connected to AMA (3.17).

ORIG	DEST	Connected?
AMA	DFW	1
AMA	DAL	1
AMA	LAS	1
AMA	IAH	1
AMA	ABQ	1

Table 3.6: Airport-pairs with AMA as origin airport.

$$A_{AMA} = \begin{array}{c c} & \begin{array}{ccccc} DFW & DAL & LAS & IAH & ABQ \end{array} \\ \begin{array}{c} DFW \\ DAL \\ LAS \\ IAH \\ ABQ \end{array} & \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Figure 3.17: Adjacent matrix of AMA airport with its neighbouring nodes.

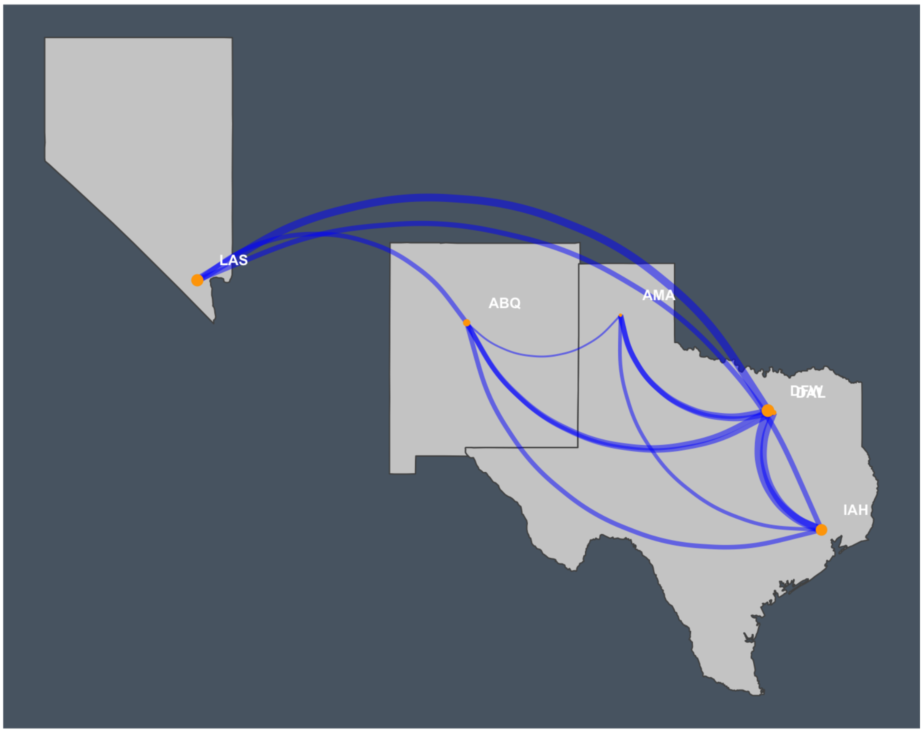

Figure 3.18: AMA airport's network. Size of a node is representative of the node degree.

Considering only the nodes that AMA is connected to, the actual number of triangles that currently exists need to be calculated. In order to do that, the connectivity amongst the nodes connected to AMA need to be investigated. The assumption taken of the US ATS being an undirected network implies that all nodes

AMA is connected to will at the same time be connected to AMA airport. In order to illustrate this, AMA connectivity is represented by its adjacent matrix (A_{AMAj}) (Figure 3.17, where j is the set of airports connected to AMA). Given Figure 3.17, and the assumption that all nodes are connected to AMA airport, the total number of existing triangles is 16. Therefore $CC_{AMA} = 16/20 = 0.8$. Figure 3.18 shows a graphic representation of the AMA airport network.

From the example above one could argue that CC might be capturing low-costs' point-to-point strategy, rather than hub-and-spoke flag carrier's strategy. In line with what is found when exposing the AMA airport example, Figure 3.16 (i.e. which shows the relationship between clustering coefficient and node degree for all airports considered in this study) clearly demonstrates the fact that nodes with higher degrees do not necessarily have greater clustering coefficients.

3.6.3 Aviation-related variables to study airport connectivity within the US

The set of aviation-related variables included in the airport connectivity models are distance between airports and a set of two dummy variables indicating whether one, two or none of a pair of airports are hub airports[16].

Investigation with a larger set of variables was also performed. However the inclusion of these 3 variables proved to be the best performing combination. Other variables previously considered to influence airport connectivity include city-pair characteristics such as population and O&D passenger demand, a dummy variable capturing whether for a given O&D any airport-pair connection existed the previous year or fuel price. City-pair characteristics were expected to capture the need for new links (or vice-versa); information regarding previous years connectivity was introduced to the modelling approach to capture competition since adding a new connection to a city-pair with already existing connecting routes might imply added difficulty to gain market share; and fuel price was included to capture external factors that would influence airlines' operational costs, and therefore airline decisions regarding route planning. A brief summary on previous attempts can be found in Appendix D.

[16]This set of dummy variables was also included in two of the O&D demand model and are noted as h_1 and h_2

3.6.4 Airport connectivity model specification

Initially, the actual airport-pair connectivity of the US ATS between years 2007 - 2013 is analysed to understand the evolution that the US network has experienced throughout this year range. Table 3.7 gives a summary of the connectivity between US airport-pairs from 2007 and 2013 as well as the average values across that same period. In total a sample of 20,098 airport-pairs per year have been included in the analysis presented in Table 3.7. The information presented in Table 3.7 is as follows:

- *Connected airport-pairs* is the number of airport-pairs considered to be connected;
- *Connected* -> *Connected* refers to the number of airport-pairs that stayed connected given that they were connected the previous year;
- *Connected* -> *Unconnected* refers to the number of airport-pairs removed from the network with respect to the previous year; Unconnected airport-pairs is the number of airport-pairs not connected;
- *Unconnected* -> *Unconnected* refers to the number of airport-pairs that stayed unconnected given that they were not connected the previous year;
- *Unconnected* -> *Connected* refers to the number of new links that every year are added to the network;

Year	connected airport-pairs	Connected -> Connected	Connected -> Unconnected	Unconnected airport-pairs	Unconnected -> Unconnected	Unconnected -> Connected
2007	2123	1936	56	18858	17070	187
2008	2130	2060	63	18851	17056	70
2009	1938	1891	239	19043	17072	47
2010	1954	1878	60	19027	17235	76
2011	1965	1895	59	19016	17225	70
2012	1994	1915	50	18987	17205	79
2013	1987	1911	83	18994	17179	76
Mean	2010	1926	87	18971	17148	86

Table 3.7: Airport-pair connectivity between 2007 and 2013. An airport-pair is considered connected if there are at least 52 flights operating between them[17].

[17]Walk-through example: in 2006 there were 1,992 connected airport-pairs (1936+56); from 2006 to 2007 the net airport-pairs added into the network were 131 (187-56) and therefore, the total number of connected airport-pairs in 2007 were 2,123 (1,992 + 132).

It is interesting to see that the number of new links is roughly the same as the number of links being removed from the air transport system. The exception are two: 2007 when the number of links added is more than three times the number of links being removed; and 2009 when there was the opposite effect (i.e. one could think 2009's effect was due to the economic crisis). These results differ from Kotegawa's study (2012) because he had a different size of dataset and the number of years used to compute those values were 1990-2009 as opposed to 2007-2013 used in this research[18] .

Since new links between airport-pairs are similar to the number of removed links, one could assume that the same airport-pairs that get connected at one point might be the same ones that get disconnected after certain period of time. The presence of airport-pairs connected for a short period of time -i.e. 1 or 2 years- could be the effect of airports launching special prices for airlines to start operating at these airports. From all possible 20,098 airport-pairs in the network, a total of 1,985 airport-pairs are connected across the entire period (i.e. from 2007 to 2013); whereas 18,448 airport-pairs are unconnected across the same period. Consequently, only 848 links change their connectivity status throughout the period range considered.

Table 3.8 presents the number of times a link gets removed against the number of times that link gets added to the network. The values presented in Table 3.8 are the aggregated number of airport-pairs that have had at least one change in connectivity throughout the years between 2007 and 2013.

			Unconnected -> Connected # added links		
			0	1	2
Connected -> Unconnected	# removed links	0	NA	246	0
		1	257	272	35
		2	0	24	14

Table 3.8: Connectivity grid: removed links against added links. Number of airport-pairs removed against number of airport-pairs added to the network. Numbers are computed across the range years 2007 - 2013.

For example, between 2007 to 2013, 257 airport-pairs have gone from being

[18]Kotegawa (2012) developed two modelling frameworks, one integrated with the FAA Airspace System Performance Analysis Capability (NASPAC) and another one to be integrated to the NASA/Purdue Fleet-level Environmental Evolution Tool (FLEET). The former considered 140 and 94 links to be added and removed respectively; whereas the latter considered 89 and 50 links to be added/removed respectively.

connected to disconnected. Similarly, 246 have gone from being unconnected to connected at some point across the same period. Therefore, 503 (i.e. 246 + 257) airport-pairs have changed their connectivity status only once between 2007 and 2013.

At the other extreme we have airport-pairs that kept connecting and disconnecting throughout the period range. For example, 14 airport-pairs have been removed and added from/to the network twice. This behaviour across a 6 year period means that these links have been only connected on alternate years. Similarly, 24 airport-pairs have been removed twice and added once; 35 airport-pairs have been added twice and removed once; and 272 were were removed and added once. One would assume some of these airport-pairs might follow a behaviour in-line with the special offers' assumption aimed to increase the use of specific airports.

Looking into more in detail for the 272 subset of links removed and added once, Table 3.9 shows the distribution of the number of years these 272 airport-pairs were connected. From Table 3.9, the majority of airport-pairs are connected for 1 or 2 years across the 7 year period, which would be aligned with the assumption of some of those links appearing into the network due to promotions to operate at those airports.

	Connected years						
	1	2	3	4	5	6	7
Number of airport-pairs	80	105	13	23	15	18	18

Table 3.9: Distribution of connected years for those links that change their connectivity status twice between 2007 and 2013 -i.e. distribution of the 505 airport-pairs from Table 3.8-.

From the connectivity analysis of the US air transportation network presented above the following considerations during the estimation process are drawn:

- Airport-pairs that have a high rate of connectivity change[19] -for example, FLL - CAL (Fort Laureldale, FL - Akron, OH) was connected intermittently across the 7 year period between 2007 and 2013- are excluded from the estimation process. Therefore, a total of 345 airport-pairs are excluded, which are those shown in Table 3.8 that have at least two connectivity changes. Considering the remaining 503 connectivity changes bring the average of links removed and

[19]One connectivity change is defined when the connectivity of an airport-pair is changed from one year to the following. For example, an airport-pair going from being unconnected to being connected the following year.

added to the network down to 35 and 32 respectively as presented in Table 3.10.

- The small percentage of links removed from Table 3.8 (345/20,098) and the residual network connectivity (Table 3.10) serve as evidence to assume the US air transport network as a mature system with a relatively steady-state in terms of connectivity changes.

Year	connected airport-pairs	Connected -> Connected	Connected -> Unconnected	Unconnected airport-pairs	Unconnected -> Unconnected	Unconnected -> Connected
2007	1916	1863	44	17006	16962	53
2008	1913	1892	24	17009	16985	21
2009	1864	1843	70	17058	16988	21
2010	1855	1833	31	17067	17036	22
2011	1860	1832	23	17062	17039	28
2012	1885	1843	17	17037	17020	42
2013	1892	1849	36	17030	16994	43
Mean	1886	1850	35	17036	17003	32

Table 3.10: Airport-pair connectivity summary after excluding the 345 airport-pairs due to their unstable connectivity. Mean values across years 2007-2013.

Having eliminated one source of scatter, several modelling approaches are investigated, with the one presented in this section demonstrating the best accuracy and performance. The modelling approach presented below splits the problem into two and as such two different models are needed: one aiming to estimate links that are added into the network the following year -i.e. named link addition model; and another one aiming to estimate links that will be removed from the network the following year -i.e. named link removal model.

The decision to have two different models (i.e. one for link addition and one for link removal) follows an extensive attempt to use a single model for predicting connectivity changes between airport-pairs. Other variants also included the exploration of a larger set of aviation-related variables as inputs, for example the O&D passenger demand, but this did not improve the modelling accuracy - rather the opposite. Initial attempts using a single model aimed to predict whether airport-pairs would be connected or not, rather than whether there would be a connectivity change, and hence the models had a high accuracy on predicting disconnected airport-pairs (i.e. the majority class) but failed to identify those connected well. Model formulation had to be changed so that the aim of the model focused on predicting connectivity changes rather than connectivity status. Since different factors might influence the addition and removal of a link to/from the network having a single model with the aim of predicting any type of connectivity change (i.e. addition and removal) would not be able to capture both types of connectivity changes in one, and therefore the problem was divided into two.

Removing O&D passenger demand from the model, however, introduces a caveat on the predictive capabilities of the connectivity model, since links formed by airports without any connection elsewhere would be impossible to predict. However, given the assumption that the US ATS is a mature system with a relatively steady-state demand as it was pointed out earlier -i.e. and shown by the low variability of the network connectivity presented in Tables 3.7 and 3.10, the number of links formed by completely unconnected airports is expected to be low. This assumption goes in line with the assumption that airlines' objective is to optimise their existing networks, by which links between airports that have no existing connections would be extremely rare.

The problem that represents the link addition and removal cases is considered a classification problem. In the first case (i.e. link addition), the problem is whether airport-pairs that are currently not connected would be connected in the future. In the second case (i.e. link removal), the problem is whether airport-pairs that are currently connected would be disconnected in the future. Both model responses can be seen as a binary response: for the link addition model '1' if the link is added into the network, '0' otherwise (remains unconnected); and for the link removal model '1' if the link is removed from the network and '0' otherwise (the link remains connected).

For both models (link addition and link removal), a logistic regression model has been used. Logistic regression takes the form of Equation 3.11 in general form, where $h_\theta(x)$ represents a probability, Z is the linear combination of input variables, θ^T (Eq. 3.12) represents the set of coefficients to be estimated and x is the set of input variables. Logistic regression is a statistical technique that trains a probability curve based on historical events, therefore the resulting value of applying Equation 3.11 in the area of airport connectivity is the probability of an airport-pair changing its connectivity status (i.e. a value between 0 and 1, with values closer to 1 indicating high probability of an airport-pair changing its connectivity status, either going from connected to unconnected or from unconnected to connected).

$$h_\theta(x) = \frac{1}{1 + exp^{-Z}} \tag{3.11}$$

$$Z = \theta^T x = \theta_0 + \theta_1(x_1) + \theta_2(x_2) + .. + \theta_n(x_n) \tag{3.12}$$

Different sets of input variables have been investigated for both link addition and link removal models. However, in this section only those with the highest performance accuracy are discussed. For both models, link and addition model, a logistic regression equation taking the general expression of Equation 3.11 has been used, with the only difference being the set of input variables. For the link removal model, only CC has been retained as input variable (as presented in Equation 3.13), with this model specification producing the most accurate predictions amongst all models investigated. In this case, the cluster coefficient term associated to a specific airport-pair is added into the model as a combined variable by applying Equation 3.9.

$$P_{removal,ij} = \frac{1}{1 + exp^{-(\theta_0^R + \theta_1^R (CC_{ij,t-1}))}} \tag{3.13}$$

where $P_{removal,ij}$ refers to the probability of link between airports i and j being removed from the network, $CC_{ij,t-1}$ refers to the cluster coefficient term associated with the pair of airports i and j and computed applying Equation 3.9 related to the previous year and θ^R is the set of coefficients to be estimated[20].

Parameters estimated for the link removal model are presented in Table 3.11, with all coefficients statistically significant at 95% confidence level[21]. From these results, each one-unit change in clustering coefficient variable will increase the log odds of an airport-pair being remove from the network (i.e. likelihood of an airport-pair changing its connectivity from connected to unconnected) by 1.91. The log odds is the logarithm of the odds ratio, which is the probability of the desired outcome (i.e. in this case an airport-pair being removed from the network) being

[20] Note that since there are two logistic models the superscript R is used to indicate the set of coefficients for the link removal model; while the superscript A is used for the link addition model.

[21] Link removal model cannot be tested for multicollinearity since the model contains only a single variable and multicollinearity is an issue affecting two or more variables that are highly but not perfectly correlated.

true divided by the probability of desired outcome not being true. Since logistic regression models generate a probability, log odds are used to interpret the constant effect of the predictors variables (i.e. in this case only CC) on the likelihood of an outcome occurring -i.e. which for the case of the link removal model is an airport-pair currently connected being removed from the network.

	Coefficient
(Intercept) (θ_0^R)	-5.37^{***}
CC (θ_1^R)	1.91^{***}
Num. obs.	1907

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 3.11: Estimated model results for the link removal model.

The positive coefficient associated with the clustering coefficient variable obtained when estimating the link removal model is in line with the relationship seen and discussed when analysing the US topological properties (Section 3.6.2) by which nodes with higher degree will tend to have lower clustering coefficients. Consequently, links between airports with low node degree (i.e. airports with a low number of connections) will typically have a higher probability of being removed from the network than those airport-pairs between airports with high node degrees.

Airports with high clustering coefficients are typically small airports and considering the hub-and-spoke routing structure that characterises the US air transportation system, links associated with small airports would be operationally easier to remove than those links associated with larger hubs. Also, the profitability of these small-airport links might not be enough to maintain operations in such airports, especially during economic downturn. For example, Wittman and Swelbar (2013b) showed that small- and mid-sized airports have been largely affected by cuts in commercial aviation in the US, through looking at reductions in the number of scheduled domestic flights in US's largest airports compared to the reduction at smaller airports during the recession years (i.e. 2007 - 2012). The reduction was 8.8% for larger airports whereas a reduction of 21.3% was seen in small- and mid-size airports.

This results differ to those found by research done by Kotegawa (2012), by which node weight was found to be the variable resulting the highest accuracy. Differences might arise due to the different set of airports used by Kotegawa (2012) and that therefore would infer slightly different network characteristics.

For the purpose of predicting which links will be added to the network (i.e. link addition model), EVC with respect to the previous year, distance between the airport-pair and a set of two dummy variables indicating whether none, one or both airports within an airport-pair are hubs proved to be the best combination of input variables. In this case, EVC is added into the model as two separate variables, one for airport i and one for airport j (e.g. EVC_i and EVC_j). Link addition model takes the form of Equation 3.14 where $P_{add,ij}$ refers to the probability of airports i and j being added to the network, ECV_i and ECV_i refers to the eigenvector centrality associated with airports i and j respectively, d_{ij} refers to the distance between airports i and j, and $h1$ and $h2$ are a set of dummy variables indicating whether both or none airports i and j are hub airports respectively.

$$P_{add,ij} = \frac{1}{1 + exp^{-(\theta_0^A + \theta_1^A(EVC_{i,t-1}) + \theta_2^A(EVC_{j,t-1}) + \theta_3^A(d_{ij}) + \theta_4^A(h1_{ij}) + \theta_5^A(h2_{ij}))}} \tag{3.14}$$

The above results are broadly in line with the literature (Kotegawa, 2012). Kotegawa's link addition model used a combination of Support Vector Machine (SVM) and logistic regression: the former was applied to find a subset pool of possible new link candidates first (i.e. reducing the search space); while the latter was applied in this subset of link candidates to obtained the probability of those changing their connectivity status. Eigenvector centrality was used as input variable in the first stage (i.e. SVM), whereas node degree was used in the logistic regression model. Although the modelling approaches are different, results are in line with existing work since eigenvector centrality is linked to a node's degree, and a measure associated with the influence of the node.

Estimated model results for the link addition model are presented in Table 3.12, with all coefficients statistically significant at 95% confidence level and no sign of multicollinearity[22]. From results presented in Table 3.12, each one-unit change in EVC_i and EVC_i will increase the log odds of an airport-pair being added into the

[22]VIF values computed for each of the variables is less than 2.7.

network via a flight service by 1.99 and 1.53 respectively. In contrast, distance has a negative effect to the log odds by -1.1. The interpretation of dummy variables within a logistic regression model is slightly different than continuous variables: going from only one of the airports (of an airport-pair) being a hub airport to none of the airports being a hub will increase the log odds of an airport-pair being added to the network by 0.69; whereas going from none or only one hub airport to both airports being a hub will increase the log odds of an airport-pair being connected by 3.87.

	Coefficient
(Intercept) (θ_0^A)	-8.19 ***
$EVC_{i,t-1}$ (θ_1^A)	1.69 ***
$EVC_{j,t-1}$ (θ_2^A)	1.53 ***
$dist\_ij$ (θ_3^A)	-1.10 ***
hub1 (θ_4^A)	0.69 ***
hub2 (θ_5^A)	3.87 ***
Num. obs.	17015

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.12: Estimated model results for the link addition model.

Eigenvector centrality measures an airport's importance while considering at the same time the importance of the airports it is connected to. EVC reflects the influence score for strategically connected airports, by which an airport with few connections could have a very high eigenvector centrality if those few connections were themselves very well connected, which would imply that there is a high flow of passenger demand making the addition of a new link most likely to be profitable. Results obtained associated with the dummy variables indicating whether within the airport-pair none, one or both airports are hubs are in line with the well-established hub-and-spoke network structure that characterises the US ATS; results show that links between two hubs are more likely to be added than those between a hub and no-hub airports or between two no-hub airports. In line with the mention finding above regarding small- and mid-size airports being largely affected by cuts, airlines will prefer to add a new route between airports in which resources are already in place and high volumes of demand are expected. Finally, the negative relationship between distance and the addition of a new link is again in line with the hub-and-spoke network structure in the US, in which long-haul markets would tend to be served by one-stop itineraries rather than by adding a new non-stop flight. The results obtained are in line with the characterisation of the US network as having a mainly hub-and-spoke structure, which might mean that the model might fail to predict the evolution of the sub-network structure that some of the low-cost carriers in the US follow (i.e. point-to-point).

3.6.5 Airport connectivity model validation

To evaluate the predictive power of the link addition and link removal models, both models are validated using actual data for years from 2008 to 2013. Considering all possible airport-pairs that the set of 337 airports considered in this study can form, along with their respective information used in both models (i.e. CC and EVC from the previous year, distance and hub information), the set of airport-pairs are split between connected and unconnected. Equation 3.13 is applied to the subset of connected airport-pairs; whereas Equation 3.14 is applied to the subset of unconnected airport-pairs.

Since applying Equations 3.13 and 3.14 results in a probability value (i.e. between 0 and 1), the resulting value needs to be mapped to a binary response (i.e. 0 or 1). In order to perform this mapping, a classification threshold (or decision threshold) needs to be decided. This threshold is usually problem-dependent, and any resulting value above that classification threshold would be classified as 1; whereas resulting values below will be classified as 0.

Taking the consideration of the US ATS as a mature system, where only small changes in network connectivity are expected, an alternative approach to that classification threshold has been used. From Tables 3.7 and 3.10 is clear that there is little shift in airport-pair connectivity, with an average of 87 airport-pairs per annum disconnecting and an average of 86 airport-pairs per annum forming new connections for the year range 2007-2013. The models per-se would not have a limit on link addition and/or removal. Therefore, an endogenous limit[23] is used to decide a limited number of airport-pairs changing their connectivity status instead of using a probability threshold. The choice of the limit is described shortly. The reasoning

[23] Endogenous limit in this context is defined as a constant value threshold imposed on the number of positive events that will be considered.

behind this decision is based on:

- It is considered that there is a diminishing return by adding a new link in a market that is already saturated. This assumes that an airline is less likely to add a new link in a market in which already operates and that is already generating a revenue. Note that the number of O&D markets served by an airline takes into account all the O&D markets served by its alliance's partner.
- Slow fleet growth limits the number of links that an airline can add to its network, needing in most of the cases to remove an existing link to be able to open a new one.

The consideration of using a limit value follows the same approach used by Kotegawa (2012). Kotegawa's work (2012) showed that on average 94 links are removed every year within the US ATS, while 140 new links are added on average every year into the US network when considering 304 nodes. These thresholds refer to the average annual values for link removal and addition between 1990 and 2009. As mentioned earlier, Kotegawa's study (Kotegawa, 2012) differs from the study presented in this document in the numbers of nodes considered as well as the years in which the study is performed. The research presented in this dissertation considers 337 nodes and the number of links added and removed on average every year are 86 and 87 respectively (Table 3.7).

Another aspect to consider when studying the capacity network evolution is the saturation within a given market, which refers to the city-pair as opposed to a link which refers to one of the available airport-pairs serving that city-pair. A market is assumed saturated when it reaches the maximum number of competitors (i.e. airlines) for a profitable market. A theoretical study done by Hansen and Liu (2015) shows that the number of airlines operating in a market is affected by the form of frequency dependence (e.g. s-curve) which is assumed in itinerary utilities. This consideration is based on the diminishing return of adding a new link in a markets that is already saturated. Therefore, for those saturated market only the action of link removal would be possible.

Similarly, link addition mainly occurs within markets in which no direct link or only one link exists. To illustrate this, Figure 3.19 shows the number of links that O&D markets had the year before a link was added, while Figure 3.20 shows the number of links that O&D markets had the year before a link was removed. These figures include data between 2007 and 2013. As Figures 3.19 and 3.20 show below, most of the capacity network evolution occurs within markets with no-direct or one link, for link addition, and within markets with one or two links, for link removal. Based in Figures 3.19 and 3.20 the threshold of saturated market for link addition is set up at 3 connections for the previous year except for 2008 in which the threshold is 5; while the threshold of saturated market for link removal is set up at 5 connections for the previous year.

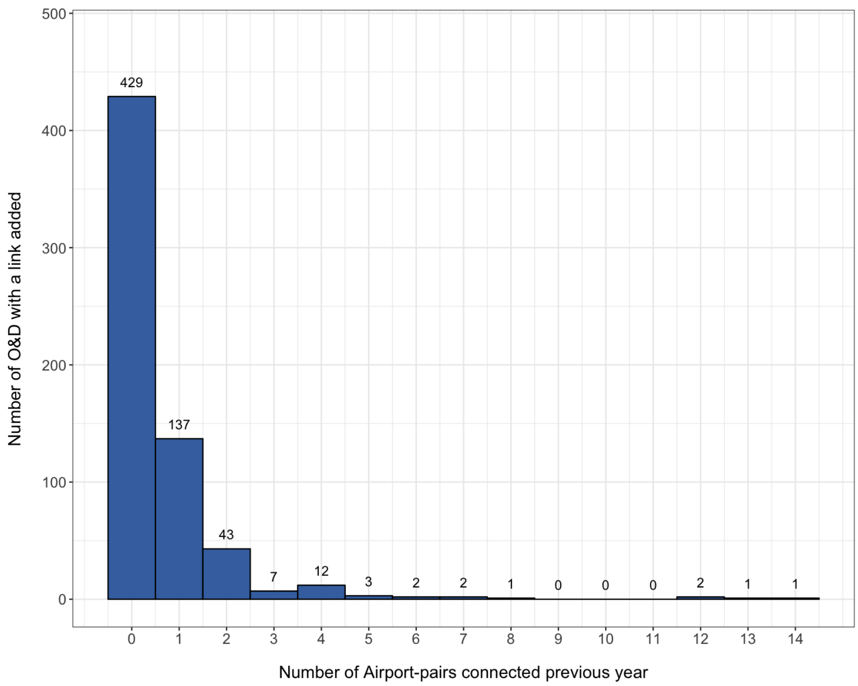

Figure 3.19: Distribution of number of links connected the previous year for those markets that experienced a link addition.

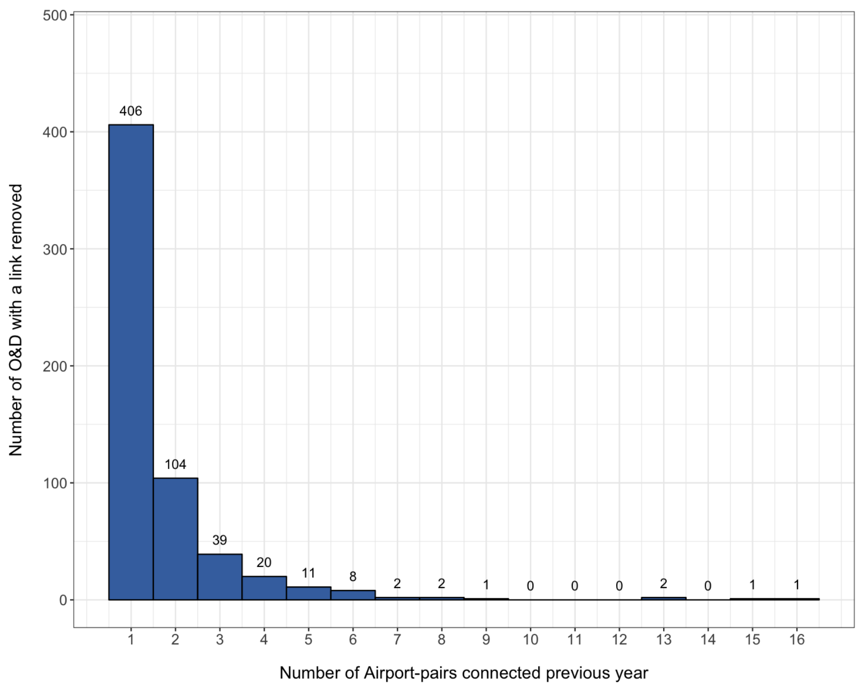

Figure 3.20: Distribution of number of links connected the previous year for those markets that experienced a link removal.

The number of new links to be added into the system is set up to be 86 while the number of links to be removed from the network is set up to be 87 (i.e. average number of links added and removed annually between 2007 and 2013). The exception is 2009, for which a 200 limit has been considered for links removed to take into account the economic crisis and the effect it had on the aviation industry.

To evaluate the model accuracy for classification problems, a set of model performance measures are considered as follows: specificity, sensitivity and precision (i.e. metrics related to how well the model predicts positive or negative cases) plus false positive rate and false negative rate (i.e. measures of the models' error). Each of these metrics are defined as per below:

- Specificity (True Negative Rate - TNR): measures the proportion (%) of actual negatives that are correctly identified as such. For the case of link removal, actual negatives will refer to those links that have not been removed from the network -i.e. without a change on their connectivity-; for the case of link addition, actual negatives will refer to those links that have not been added to the network and instead remain unconnected.

- Sensitivity (True Positive Rate - TPR): measures the proportion (%) of actual positives that are correctly identified as such. For the case of link removal, actual positive will refer to those links removed from the network; for the case of link addition, actual positives will refer to those links that have been added to the network.

- Precision: proportion (%) of predicted positives that are actually positive over the total number of predicted being positive. Note that precision is used for model evaluation rather than accuracy since in this case study, where the connectivity changes is small and is limited to a threshold, accuracy (i.e. which indicates the fraction of correct predictions) would tend to be high.

- False Negative Rate or Miss Rate (FNR): measures the proportion (%) of actual positives that are incorrectly identified not as such.

- False Positive Rate or Fall-out (FPR): measures the proportion (%) of actual negatives that are incorrectly identified as positive.

A summary of performance metrics from validation of the estimated link removal model is presented in Table 3.13. TNR refers to True Negative Rate, and FPR and FNR refer to false positive rate and false negative rate respectively. Similarly, performance metrics summary for the estimated link addition model is presented in Table 3.14. Performance metrics for link addition are the same as the ones for link removal with the exception of computing true positives (i.e. TPR - is the number of links predicted to be added into the network that are actually being added to the network) instead of true negative rate.

Year	Actual links connected	Actual links removed	TNR	Precision	FPR	FNR
2008	2060	63	7%	11.1%	88.9%	4%
2009	1891	238	70%	29.4%	70.6%	6.9%
2010	1878	60	6%	10%	90%	4.5%
2011	1895	59	6%	10.2%	89.8%	4.4%
2012	1915	50	16%	32%	68%	3.9%
2013	1911	83	16%	19.3%	80.7%	3.9%
Mean	1925	92.2	20.2%	18.7%	81.3%	4.6%

Table 3.13: Performance metric summary from the validation of the estimated link removal model.

Year	Actual links disconnected	Actual links added	TPR	Precision	FPR	FNR
2008	17056	70	13%	18.6%	81.4%	0.5%
2009	17070	47	9%	19.1%	80.9%	0.5%
2010	17232	76	11%	14.5%	85.5%	0.5%
2011	17225	70	12%	17.1%	82.9%	0.5%
2012	17204	79	14%	17.7%	82.3%	0.4%
2013	17179	76	13%	17.1%	82.9%	0.4%
Mean	17161	69.70	12%	17.4%	82.7%	0.5%

Table 3.14: Performance metric summary from the validation of the estimated link addition model.

Results of both models present an improvement compared to when a single model was applied, increasing precision levels to just below 20% for link addition and about 20% for the link removal model. Although these results are an improvement from previous attempts, precision levels are still significantly low. Since results from the literature show a similar trend of relatively low precision levels -i.e. Kotegawa's (2012) results had an accuracy of just under 20% and 40% for the link addition and removal respectively -, it is believed links may have individual random factors affecting the connectivity that might be impossible to formulate mathematically. Given the fact that the number of connectivity changes in the US is relatively low and that external factors might not affect the entire network in the same way, such as small airports being typically affected more than large ones by cuts in commercial aviation, the complexity of the problem might not be possible to formulate with these two aggregate models looking at the entire network, but different sub-networks, such as the point-to-point network structure followed by LCC, might need a more specific model linked to their dynamics[24]. A comparison with Kotegawa's work can be also done based on the FPR, by which his work obtained 90% and 60% for the link addition and link removal models respectively; with an achieved FPR of 82.3% and 81.7% for the link addition and removal models in this research, an improvement can be found for the former model, which suggests the potential of including a set of variables beyond those associated with network theory, as done by Kotegawa, to improve these type of models aim to predict connectivity changes.

[24]Figures in Appendix D.2, which show a graphic representation of the links added and removed in 2011 and 2012 as example, clearly show there is not an obvious trend on link addition and removal but rather a quite even distribution of those links changing connectivity across the entire network.

A further analysis is performed to check whether the combination of these two models allows the prediction of the evolution trend of the entire network, with the caveat that these models are poor at predicting the exact airport-pairs connectivity change. However, if the models are able to predict at least the development of the network structure, they will be useful at the aggregate level when studying the network evolution trend. In order to test this hypothesis, a second validation is performed by applying the estimated models to the entire network and generate predictions of the US Air Transport network evolution, starting off from actual data in 2007 and then using the generated predictions up to 2013.

The predicted network degree and network eigenvector centrality evolution between years 2008 and 2013 are compared to their actual values in Figures 3.21 and 3.22. Predicted and actual network topology metrics seem to follow similar trends, which would confirm that the models presented in this study could be useful for the purpose of predicting the development of the network structure at the aggregate level over time.

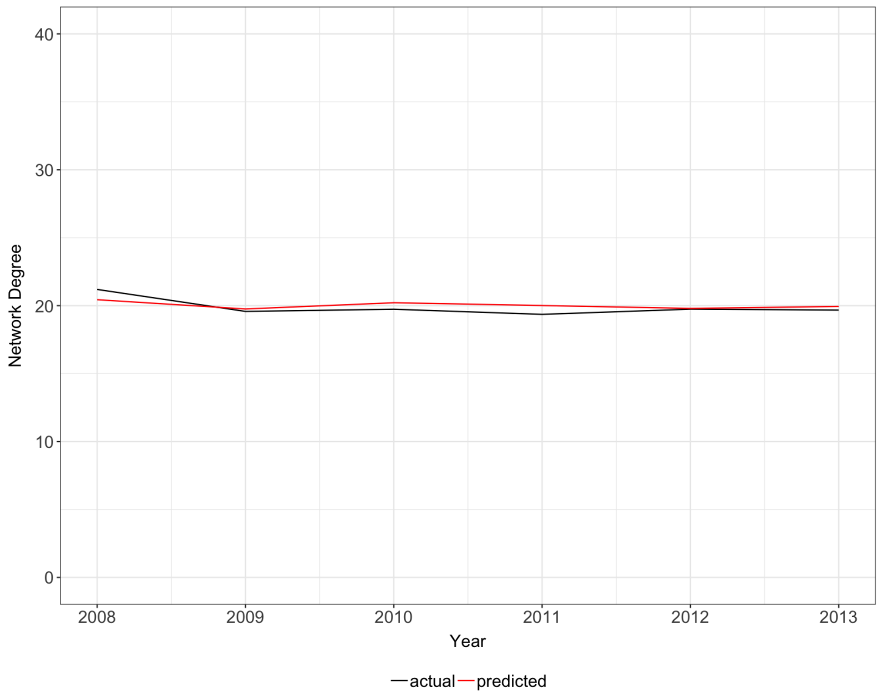

Figure 3.21: Actual and predicted network degree evolution between 2008 and 2013.

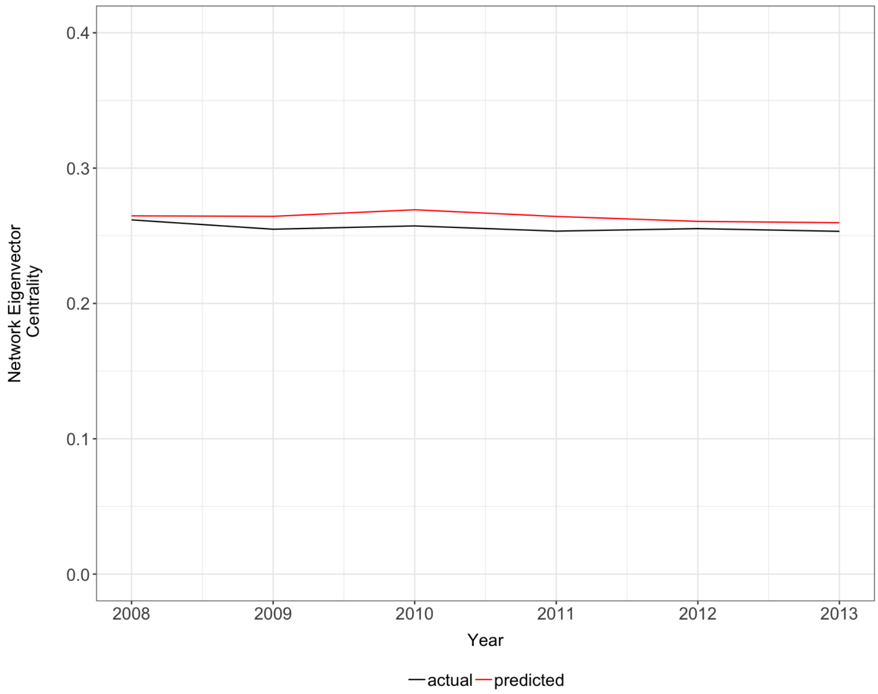

Figure 3.22: Actual and predicted network eigenvector centrality evolution between 2008 and 2013.

3.7 Module 2: Itinerary choice model using multinomial logit

The aim of the itinerary choice model (or passenger choice model) is to estimate the number of passengers that choose each of the available itineraries between two cities. Two methodologies have been used in this study: multinomial logit and neural network. In this section the first methodology is described. For the purpose of simplicity, it is referred to as MNL (i.e. multinomial model) in the rest of the thesis.

3.7.1 MNL model specification

The multinomial logit model is based on the random utility theory such as that the probability of decision maker n of choosing alternative i in a choice set $m = 1,2,3,..,j$ is defined as Equation 3.15 (Ben-Akiva et al., 1985).

$$P_{in} = \frac{e^{V_{in}}}{\sum_{j \epsilon m} e^{V_j}} \tag{3.15}$$

where V is the utility function for choice i. V_i is a linear function of the input variables and assumes that each vector of attributes characterising an alternative can be reduced to a scalar value, which expresses the attractiveness of each alternative. Consequently, it is expected that the choice maker will choose the alternative with the highest value, maximising his/her utility. Expression 3.16 shows the general expression of V where X is the vector of K explanatory variables and β represents the set of parameters to be estimated.

Since in this research, the choice making problem is at the aggregate level -i.e. the objective is to estimate the aggregate passenger demand for any itinerary - the probability estimated is at the same time the itinerary share for that given itinerary. And therefore there is not a specific decision maker n.

$$V_i = \sum_{k=1}^{K} \beta_{ik} \cdot X_{ik} = \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot X_{i3} + \cdots + \beta_k \cdot X_{ik} \tag{3.16}$$

Multinomial logit models are usually estimated using a maximum likelihood estimation (MLE) process. However, in choice behaviour modelling software capability becomes a critical concern when dealing with a significant number of choice sets and when only an aggregated demand for each choice is available. In the case of itinerary choice modelling within air transport, the number of choice sets is large (i.e. more than a thousand for this study case), making the number of modelling tools available to deal with such problems scarce. Consequently, most of the studies that have dealt with this type of data have limited their scope to a small number of choice sets (Weidner, 1996; Coldren et al., 2003; Atasoy and Bierlaire, 2012; Ghobrial and Soliman, 1992), early work from this dissertation also being one of them (Busquets et al., 2018).

Previous effort done on topic, which used a MLE estimation process, divided the US network in to 5 different regions, as done by Coldren et al (2003): four Continental time zones (Central, East, Mountain and West) and a region for Alaska and Hawaii. This specific O&D market grouping was an attempt to capture similarities among all city-pairs and to resolve the software limitation issue that was encountered when attempting to analyse the entire US network at once.

Given these regions, 18 entities[25] were defined, and one model was produced for each of this combinations. Each of these models would estimate the itinerary share of all city-pairs within that specific region combination. Even considering these city-pairs subsets, computational capabilities became an issue for region combinations with a large number of city-pairs and model estimation results were poor.

To overcome software limitations and with the aim of achieving a workable single discrete choice model applied to the entire US network (i.e. all O&D considered in this study), an alternative solution for the estimation process is used: the Berkson-Theil approximation method. The Berkson-Theil method offers an alternative solution for such problems by transforming a logit choice model using MLE to a least square regression model. Least square regression models can be estimated with any statistical software and any data size without having to limit the scope of the problem.

This transformation was initially proposed by Berkson (1953) only considering the case of two choices, and was further developed by Theil (1969) to cover those cases when more than two choices are available (a multinomial logit model). Examples in the literature (using this methodology to estimate itinerary shares at the aggregate level) are limited. Hsiao and Hansen (2011) used the Berkson-Theil approximation method to model itinerary shares for 213,917 city-pairs; Carrier and Weatherford (2014) also used this methodology applied to aggregated booking data, with results obtained using the Berkson-Theil method being very close to those estimated when using MLE; and finally Li et al. (2017) performed a comparison analysis between the Berkson-Theil and MLE methods in aggregated choice data with multiple choice sets, showing that the Berkson-Theil method is an effective approach in dealing with this type of aggregation level and big data.

Considering the itinerary choice problem in this research, Berkson-Theil defines expression 3.17 for itinerary i in a city-pair:

$$ln(P_i) - ln(P_j) = ln\left(\frac{P_i}{P_j}\right) \tag{3.17}$$

where itinerary j is the reference itinerary in this specific city-pair. Selection of

[25] Considering all 16 possible combinations of the Continental time zones – e.g., Central-Central (C-C), Central-East (C-E), Central-Mountain (C-M), Central-West (C-W), [..], West-Mountain (W-M), West-West (W-W) –; as well as an entity for Alaska and Hawaii to Continental US and an entity for the Continental US to Alaska and Hawaii.

the base itinerary is arbitrary, consequently the itinerary with the highest passenger demand is selected as the base itinerary in each city-pair. The probability of choosing any given itinerary is its itinerary share - i.e. ratio between the number of passenger choosing that itinerary and the total number of passengers in the city-pair-. Consequently, the probabilities of choosing itineraries i and j can be easily calculated. According to Equation 3.15, expression 3.17 is equal to V_i - V_j, which can be re-written as expression 3.18. From expression 3.18, β can be estimated by a least squares regression model. WLS is the estimation technique used in this research for applying the Berkson-Theil method and the models are estimated in R. The numbers of passengers are used as weights.

$$ln\left(\frac{P_i}{P_j}\right) = V_i - V_j = \sum_{k=1}^{K} \beta_{ik} \cdot (X_{ik} - X_{jk}) \tag{3.18}$$

Several model specifications are tested for the itinerary choice model, and eventually three input variables are included in this model: journey fare (the average fare paid in dollars by passengers on a given itinerary); journey time (the average journey time in hours including connection times on a given itinerary); and the number of airlines serving a specific itinerary; all for the year 2007.

Besides the computational limitation concern within the area of itinerary choice modelling, journey fare has proven to be the most problematic of explanatory variables. This is due to the large number of fares available and the difficulty of determining fares available to individual choice makers. Consequently, some studies omit fares as input variables and others use average fares, with the later resulting in most cases in endogeneity bias (Hsiao and Hansen, 2011).

Endogeneity occurs when a correlation exists between an explanatory variable and the error term -i.e. unobserved factors- in a model. In the case of air itinerary choice models, prices are endogenous because they are influenced by demand, which is itself influenced by prices. This may lead to higher average fares on more popular routes as a result of airline pricing-and-yield management systems adjusting fares based on changing demand, which are not captured when average fares are used. Omitting the simultaneity of supply and demand systems may result in erroneous results. As a result, estimated coefficients might be biased upward leading to values of times that are too high (2011; 2017). Previous attempts in the present research to model passenger choice without considering fare endogeneity also led to erroneous results, with estimated coefficients being too high (see Appendix E).

One of the methods used to correct fare endogeneity is to use a 2-stage control-function (2SCF) using instrumental variable (IV). A instrumental variable is a variable that does not belong in the demand equation, but is correlated with the endogenous variable. The instrumental variable used in this research is the product of distance and unit jet fuel cost (in 2007 dollars per gallon) as done by Hsiao and Hansen (2011). This variable captures the cost of offering the service, and therefore has an effect on airfares, but it is expected not to have a direct effect on market shares.

As the name indicates, the two-stage control function (2SCF) method is split into two. The first-stage of the model is an ordinary least-square regression with airfare as the dependent variable, represented by Equation 3.19.

$$fare_i = \alpha IV_i + \gamma' x_i + \mu_i \tag{3.19}$$

where $fare_i$ is the airfare associated with alternative i; IV_i is the instrumental variable for alternative i; α is the coefficient associated with the instrumental variable; γ is the vector of coefficients associated with all exogenous variables, excluding airfares; x_i is the vector of explanatory variables used in stage 2 with exception of airfares; and μ is the error term for stage 1 regression.

Explanatory variables for this first stage regression model include the instrument variable -i.e. product of distance and unit jet fuel cost- along with the set of all other exogenous variables used in the itinerary share model, with exception of airfare. The second-stage of the model (Equation 3.20) uses as input variables the forecast airfare predicted during the first stage of the 2SCF methodology, the other exogenous regressors excluding airfares as well as the residuals from the first stage of the method -i.e. difference between the actual and the predicted airfare from the first stage regression.

$$V_i = \delta_i + \beta_{fare} fare_i + \beta_i' x_i + \epsilon_i \tag{3.20}$$

where δ_i is the residual from stage 1, the difference between actual and predicted airfares from stage 1; β_{fare} is the coefficient associated with airfare from stage 2;

β' is the vector of coefficients associated with all exogenous variables for stage 2, excluding airfare; x_i is the vector of explanatory variables used in stage 2 with exception of airfares; and ϵ is the error term.

Note that at the second stage, we use the Berkson-Theil method (Equation 3.18) therefore all variables are differences at the attribute level between the alternative and the reference alternative. The model specification for the second stage is as presented in Table 3.15, which includes as input variables journey fare, journey time and number of airlines.

Note that due to the characteristics of the input data, in which a large number and distinct choice sets exists - i.e. one for each city-pair - but only an aggregate demand for each choice is available, a set of alternative specific constants (ASC) are also included in the model. The *ASCs* relate to the itinerary's level of service and consist of one ASC corresponding to non-stop itineraries and 25 ASCs each corresponding to one of the connecting hub airports for the one-stop itineraries[26]. Note that in Table 3.15, *HUB* input variable is represented as a character vector that takes the IATA code of the connecting airport for one-stop itineraries and *Non-stop* value for non-stop itineraries.

During the estimation process, the first-stage of the 2-stage control function (2SCF) model, which aims to project itinerary fares, is an ordinary least-square (OLS) regression; whereas the second stage, following the Berkson-Theil method, is a weighted least-squares (WLS) regression using number of passengers as weight.

The validity of the instrument used in this research (i.e. in this case the product between itinerary distance and jet fuel cost) depends on two requirements:

- Instrument relevance: Instrument variables need to be highly correlated with the endogenous regressors.

- Instrument exogeneity: instrument variables need to be uncorrelated with the error term generated in the second outcome of the 2-stage control function model. This requirement requires 3 strong theoretical arguments to be true:

  - Exclusion restriction: no direct effect of the instrument variable on the

[26]Following the scope of this dissertation (Section 3.1.2) only non-stop and one-stop itineraries are considered in this study. One-stop itineraries are limited to be connecting in one of the 25 hub airports considered in this study.

Variable	Coefficients	Explanatory variables
constant	ASC_{NS}	1 x HUB$_\text{i}$=="Non-stop"
	ASC_{ATL}	1 x HUB$_\text{i}$=="ATL"
	$\vdots$	$\vdots$
	ASC_k	1 x HUB$_\text{i}$=="k"
Journey fare	β_{fare}	fare$_i$
Journey time	β_{time}	time$_i$
Num. of airlines	$\beta_{airlines}$	num_airlines$_i$
Residual	δ_i	fare$_i$ - $\hat{fare}_i$

Table 3.15: Specification table of the utility function for stage two of the 2SCF model estimated using Berkson-Theil method.

dependent variable, which in the case of this research would mean no effect on itinerary shares.

- Rule out any reverse effect of the dependent variable on the instrument variables.

- Convincingly describe why the instruments influence the endogenous regressors.

The first requirement can be empirically tested during the first stage of the regression by t-testing the null hypothesis that there is no relation between the instrument variable and journey fare - i.e. α from Eq. 3.19 is equal to 0. The second requirement needs a strong theoretical argument and can generally not be tested, especially in the case when the number of instrument variables (L) used is the same as the number of endogenous variables *(k)* (i.e. L=K), case known as just-identified model (Schmidheiny, 2018). For the case when there are more instrumental variables than endogenous variables (i.e. L>K), known as an over-identified model, different methodologies exist to test the validity of the models[27].

[27]Sargan (1958) noted that for linear models the residuals of the instrumental variables regression can be used to test for instrument exogeneity; Amemiya (1978) proposed a two-stage minimum-chi squared estimator for the simultaneous equations Probit model; Guevara (2006) proposed a simpler test for discrete choice models based on ratio log-likelihood testing called Direct Test.

Input data set characteristics

The base year to train the model is 2007. The input dataset is formed of 2007's air itinerary choices for the US air transportation network that serve the set of city-pairs considered in this study. For any given city-pair, an itinerary is defined as the flight or combination of flights that connect the origin city with the destination city. The choice for each itinerary is computed by the aggregated passenger demand for that itinerary - i.e. annual number of passengers - and each itinerary has associated a number of attributes as defined in the previous section.

The dataset also contains a large proportion of low-demand itineraries, which may skew the results if included in the model. Two selection criteria are being applied to exclude these less important itineraries while retaining the representativeness of the sample, similar to the criteria also applied by Li et al. (2017). First, for city-pairs with a large set of available itineraries, only the top 9 itineraries are included in the estimation process. Secondly, any itinerary that accounts for less than 1% of the total city-pair demand is also omitted. The application of these criteria reduces the number of passengers considered by 5.9%. For estimation purposes, city-pairs with only 1 choice itinerary are also omitted since there are no alternative options.

For training purposes, the US network is considered undirected and therefore the aggregated passenger demand for any itinerary is computed as the average number of passengers between both directions of the same itinerary - i.e. given an itinerary i serving a city-pair between city o and city d, number of passengers would be the mean between $N_{i,od}$ and $N_{i,do}$ (N denotes number of passengers).

	Dataset
No. of city-pairs	2,641
No. of itineraries	17,859
Maximum city-pair passenger demand	3,590,010
Minimum city-pair passenger demand	5,790

Table 3.16: Key characteristics of the input dataset.

Given the data processing applied, the key characteristics of the dataset used to estimate the itinerary share model are presented in Table 3.16. Note the characteristics presented in Table 3.16 only consider those city-pairs used to train the model.

Since only a subset of 25 airports are considered a feasible option to form one-stop itineraries, the distributions of journey fare and time across the several itinerary options are presented in Figures 3.23 and 3.24 respectively. Surprisingly, the average journey fare for non-stop itineraries is not one of the most expensive, with most of the one-stop itineraries having an average journey fare higher. With the majority of markets within the domestic US network being of a medium-haul distance (Table 3.1) it is expected that most of the passengers would travel non-stop[28], suggesting higher competition brings airfares to lower levels. Also, long-haul flights tend to be more expensive than those of a short and medium range. As expected, journey time is much smaller for the case of non-stop itineraries; with one-stop itineraries connecting at LAX airport having one of the highest average journey time, which is expected since this airport would be typically used as connecting hub for long-haul itineraries such as those that connect Hawaii with mainland US.

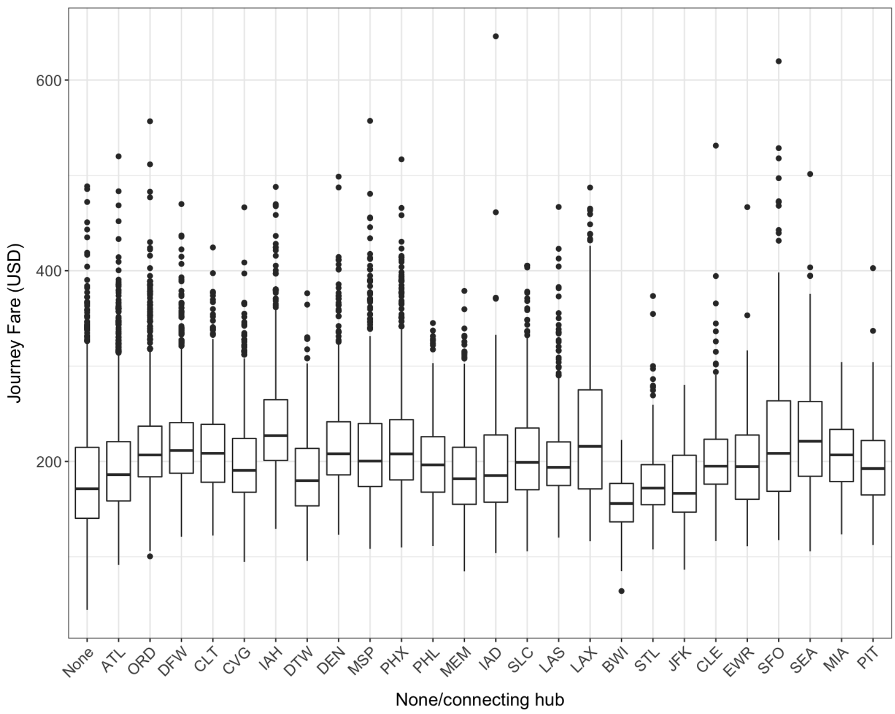

Figure 3.23: Journey fare distribution across the feasible itineraries based on the connecting airport.

[28]Considering the data used in this research around 80% of the domestic passengers choose non-stop itineraries.

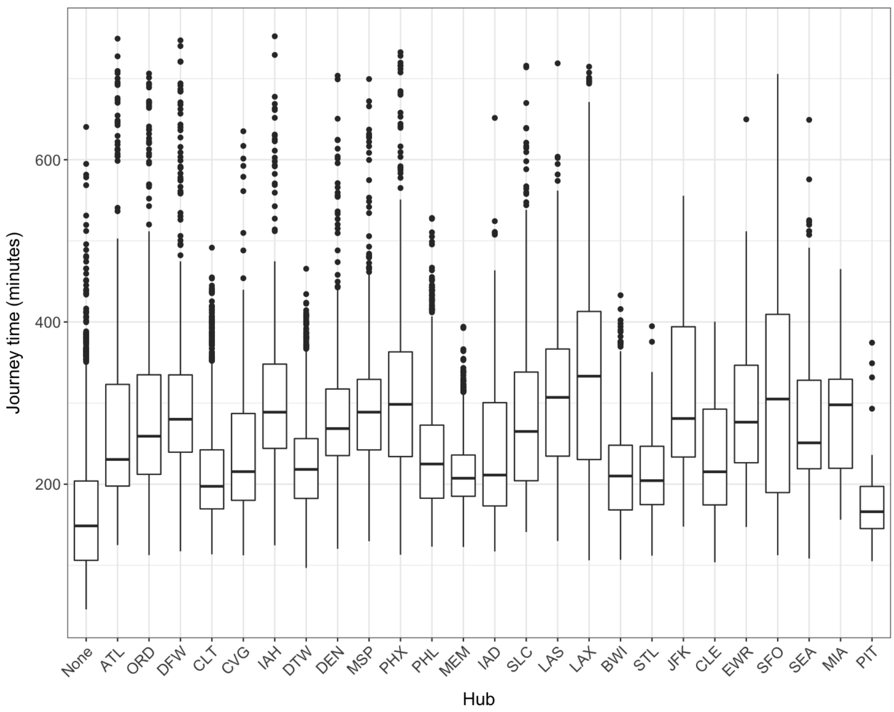

Figure 3.24: Journey time distribution across the feasible itineraries based on the connecting airport.

3.7.2 MNL model results

Results from the estimation process for the itinerary choice model using a 2-stage control function model with the Berkson-Theil approximation method are presented in Table 3.17 for the 1st stage OLS regression modelling journey fare, and Table 3.18 for the 2nd stage WLS regression modelling itinerary shares.

	Coefficients
Itinerary Cost (α)	0.013 ***
Journey time (γ_1)	4.346 ***
Number of airlines (γ_2)	-1.168 ***
Adj R-squared	0.27
F-statistic	2255 (p-value: < 2.2e-16)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.17: Estimation results for the 1st stage of the control function model aim to predict journey fares.

The results of the first-stage OLS regression aimed to model journey fare show that the estimate for the parameter associated with the instrument variable

(*Itinerary Cost*) used to control endogeneity is statistically significant at the 99% confidence level and no presence of multicollinearity[29]. This allows us to reject the null hypothesis that there is no relationship between the instrument variable and journey fare. In addition, the 1st-stage F-statistic is well above the critical value of 10 indicates having a sufficiently strong instrument variable. The threshold of 10 is recommended as a rule of thumb by Staiger and Stock (1994) when using 2-SLS. And research done by Guevara and Navarro (2013) also suggest that similar threshold can be used in the case of control function models in logit models.

The residuals (δ) from Equation 3.19 are retained and included, without transformation, as an additional variable in the utility function of the itinerary choice model. As shown in Table 3.18 (under the "Control Function" model), the parameter estimate associated with this residual is statically significant at the 99% confidence level, which confirms the presence of endogeneity, and specifically that the instrument variable is correlated with journey fare and is thus valid.

Results obtained for the second stage show all estimated parameters to be statistically significant at the 99% confidence level with the exception of the alternative specific constant refer to SEA airport as connecting hub. The model is also tested for multicollinearity, with all VIF values lower than 3 apart from journey time (9.57) and journey fare (11.7), which are still well below the threshold.

[29]VIF values obtained for each variable are less than 5.9.

	Coefficients
None (ASC_{None})	-0.161 ***
ATL (ASC_{ATL})	-0.407 ***
BWI (ASC_{BWI})	-0.650 ***
CLE (ASC_{CLE})	-1.106 ***
CLT (ASC_{CLT})	-0.618 ***
CVG (ASC_{CVG})	-1.110 ***
DEN (ASC_{DEN})	-0.550 ***
DFW (ASC_{DFW})	-0.263 ***
DTW (ASC_{DTW})	-0.664 ***
EWR (ASC_{EWR})	-1.176 ***
IAD (ASC_{IAD})	-1.115 ***
IAH (ASC_{IAH})	-0.375 ***
JFK (ASC_{JFK})	-0.372 ***
LAS (ASC_{LAS})	-1.029 ***
LAX (ASC_{LAX})	-0.657 ***
MEM (ASC_{MEM})	-1.073 ***
MIA (ASC_{MIA})	-0.427 ***
MSP (ASC_{MSP})	-0.492 ***
ORD (ASC_{ORD})	-0.622 ***
PHL (ASC_{PHL})	-0.920 ***
PHX (ASC_{PHX})	-0.764 ***
PIT (ASC_{PIT})	-1.729 ***
SEA (ASC_{SEA})	-0.124
SFO (ASC_{SFO})	-0.693 ***
SLC (ASC_{SLC})	-0.789 ***
STL (ASC_{STL})	-1.342 ***
Journey fare' (β_{fare})	-0.013 ***
Journey time (β_{time})	-0.527 ***
Number of airlines ($\beta_{airlines}$)	0.136 ***
Residual (δ)	-0.001 ***
Adj R-squared	0.7

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.18: Estimation results for the second stage of the control function model, when applying Berkson-Theil method.

All estimated coefficients are of the expected sign with coefficients obtained for journey fare and time being negative and the parameter related to number of airlines positive. More expensive and longer journeys would tend to attract less passengers if competing with other itineraries that are cheaper and/or shorter. The presence of a higher number of airlines could reflect a higher frequency on that specific itinerary, making it more attractive for passengers to choose against other itineraries. It is also an indicator of more popular routes where competition amongst airlines may result on lower journey fares, attracting a higher number of passengers. Overall, results show that most of the variation will come from journey fare and time, followed by number of airlines, with the set of ASCs accounting for everything else that fare and time cannot capture.

Coefficients obtained for the set of ASCs show how non-stop itineraries would attract most of the passengers compared to other itineraries. For one-stop itineraries the most attractive is DFW; whereas connecting airports PIT and STL have the two highest negative estimated parameters, suggesting that these are the least attractive connecting hubs amongst the available one-stop itineraries. This results would be in line with the fact that PIT used to be a major hub for US airways but in the early 2000s the airline began scaling down its operations, shifting towards PHL and CLT (i.e. with results in Table 3.18 showing these connecting airports as more attractive). STL airport is a focus city, which is defined as a destination from which an airline operates limited point-to-point routes to serve local markets rather than connecting passengers, for Southwest airlines and therefore it is expected to have less attractiveness for one-stop routes than other connecting hubs.

3.7.3 MNL model validation

In order to validate the model and assess its predictive powers, input data for years between 2008 and 2013 is used to generate output predictions of number of passengers across all available itineraries. Following the 2SCF model, parameter estimates in Table 3.17 are used to estimate journey fares for each itinerary available. The journey fares estimated along with the residuals, journey time and number of airlines are then used as input variables for the 2nd-stage model. For the second-stage, model estimation results in Table 3.18 are used as coefficients in equation 3.16, so that the utility and the probability of choosing each itinerary can be calculated. Consequently, the total number of passengers choosing each itinerary can be obtained by multiplying that probability and the total number of passengers of a city-pair. The adjusted coefficient of determination (adjusted R^2), which indicates how close predictions of passenger numbers are from the observed passenger demand, is then calculated.

Figure 3.25 shows observed against predicted number of passengers throughout the validation years (2008-2013). Average adjusted R^2 across all years is 0.857, which are relatively higher than those obtained in the literature (Li et al., 2017) and comparable to those obtained when using the most common formulation (i.e. multinomial logit using maximum likelihood estimation) to estimate air itinerary shares. The work done by Li et al. (2017), who used Berkson-Theil approximation method combined with a 1-stage OLS regression model, obtained R^2 values just above 0.6. The results obtained in this research ($R^2 > 0.8$) highlight the potential of using WLS over OLS when this approximation method is used as well as demonstrating that this methodology can be used to simplify some of the issues encounter when developing an air itinerary choice model using a multinomial logit formulation with maximum likelihood estimation process (e.g. computational limitations might lead to a decline in predicting capabilities of such models to a subset of cities).

Another way of assessing the model is through calculating value of time (VOT). VOT is the willingness of passengers to pay for the reduction of one hour of travel. The generalised formula to compute VOT is presented in Equation 3.21. VOT obtained from the 2nd-stage 2SCF model is presented in Table 3.19. VOT obtained is \$41.3/h, which is within the range of those found in the literature (Hsiao and Hansen, 2011; Atasoy and Bierlaire, 2012; Li et al., 2017). As a comparison, VOT computed from a previous modelling attempt using multinomial logit and the Berkson-Theil approximation method but without considering fare endogeneity, is also presented in Table 3.19 (noted as 1-stage WLS). As mentioned earlier, previous modelling attempts that did not consider fare endogeneity resulted in erroneous parameter estimates, leading to VOT values that are too large compared to those in the literature, which are expected to be under \$100/h (Hsiao and Hansen, 2011; Atasoy and Bierlaire, 2012; Li et al., 2017).

$$VOT_i = \frac{\partial V_i/\partial time_i}{\partial V_i/\partial fare_i} \tag{3.21}$$

	2SCF	1-stage WLS
VOT	\$41.3/h	\$296/h

Table 3.19: Comparison of Value of Time obtained for the itinerary share model using 2SCF model and using WLS regression

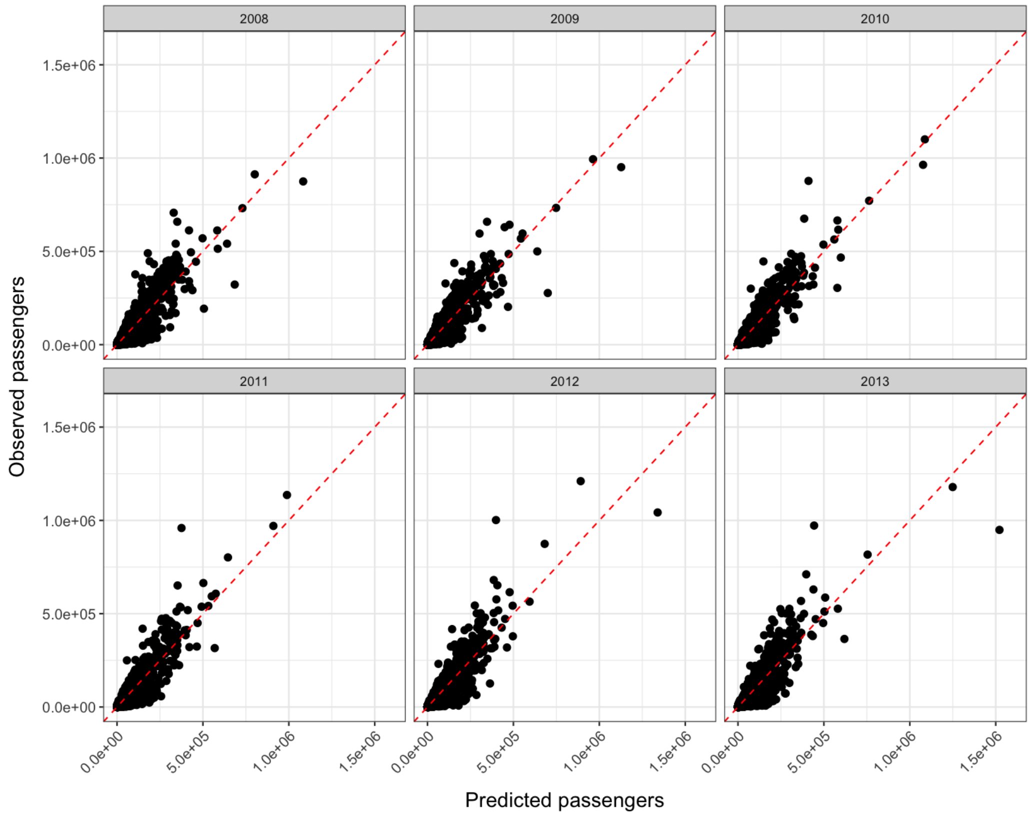

Figure 3.25: Observed against predicted number of passengers throughout the validation years.

3.8 Module 2: Itinerary choice model using neural networks

Artificial neural networks (ANN), also known as neural networks (NN), is the second technique investigated within the area of itinerary choice modelling. NN has not featured in existing research into air itinerary share modelling at an aggregate level and without using booking data. NN is a variety of deep learning and part of artificial intelligence (AI). Neural networks can be described as a parallel distributed processors made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use (d'Avila Garcez, 2014). As they currently exist, NN perform small, highly specific tasks. The neural network concept derives from the study of the human brain, which is in turn a biological neural network (BNN). The structure of an artificial neural network resembles that of a human brain. Besides the structure, two aspects enhance the similarity between ANN and the brain: knowledge is acquired through a learning process; inter-neuron connectivity strength, known as synaptic weights, are used to store the acquired knowledge.

Resembling the brain, neural networks consist of neurons inter-connected between them. They are structured in layers of similar neurons and most have at least an input layer and an output layer. Figure 3.26 shows a typical neural network architecture, consisting of an input layer, one hidden layer and one output layer. The strength of a neural network arises from the effects of the interconnection amongst the neurons. However, what happens throughout the hidden layers is not exactly known, which is one of the reasons why it is sometimes referred as a black box (Heaton, 2015).

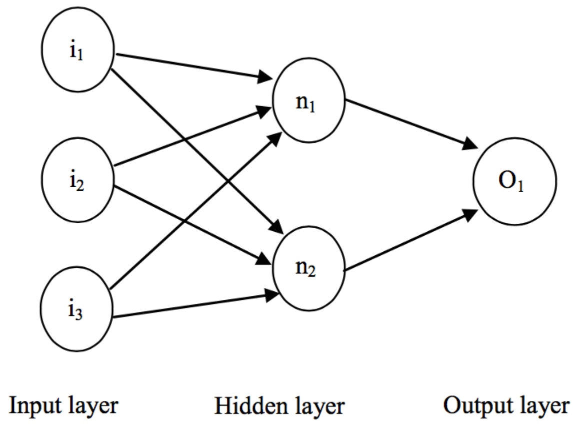

Figure 3.26: Example of a Neural Network.

In order to estimate the passenger choice model using neural networks, the topology of the neural network as well as the types of neurons need to be decided. Figure 3.26 shows the architecture of a neural network similar to the one used in this study to develop an itinerary choice model. The neural network in Figure 3.26 is formed of the following components:

- Input layer - formed of input neurons which accept data from the program into the network. There is one neuron for each input variable.
- Hidden layers - formed of hidden neurons, which process the input to form the output, however they are not directly connected to the incoming input data nor to the eventual output data. Therefore, hidden neurons can only receive input from other neurons and can only output to another neuron.
- Output layer - formed of output neurons, which provide data back onto the

program once it has processed the input data throughout the network. With the purpose of estimating the share of an itinerary i there is only need for one output neuron taking a value from 0 to 1.

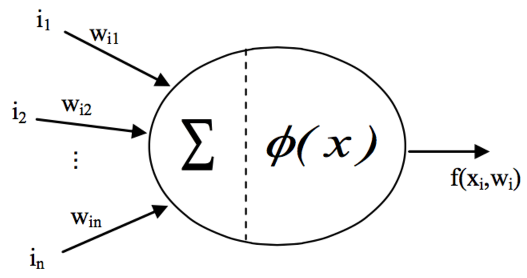

Figure 3.27: An artificial neuron.

Figure 3.27 presents a schematic of a single artificial neuron. As shown, the artificial neuron receives an input from one or more sources, denoted by $i_i(t)$. This input vector may be directly fed from the input neurons -i.e. input layer- or come from other neurons -i.e. a previous hidden layer-. Each of these inputs are multiplied by a weight. There must be one weight for each input. The neuron then calculates the weighted sum and supplies it into an activation function. Equation 3.22 summarises the calculation of a neuron's output, where x represents the value of input variable i, w is the weight corresponding to neuron i and ϕ is the activation function. The activation function aims to establish the bounds for the output of the neuron.

$$f(x_i, w_i) = \phi\left(\sum_i (w_i x_i)\right) \tag{3.22}$$

3.8.1 NN model specification

In order to define the topology, also known as architecture, of the neural network, several decisions need to be made. These include: defining the input vector, choosing an activation function, how many hidden neurons to use and the layer structure as well as deciding which learning algorithm to use to estimate the weights of the neural network model. The best way of choosing the most adequate architecture to generate a neural network is through an experimental process (i.e. trial and error)

(Heaton, 2015). For this stage, the software used to generate the neural network model to predict itinerary shares is Matlab, specifically the Deep Learning Toolbox.

First of all, the input layer is formed by as many input neurons as input variables. For the purpose of comparing both methodologies, the same model specification as chosen for the passenger choice model developed using the multinomial logit 2-stage control function is used for the neural network model. Therefore, 31 input neurons are used in total: 26 referred to the level of service -i.e. one for non-stop itineraries, plus 25 for one-stop itineraries through one of the considered hub airports; one referred to the predicted journey fare associated to an itinerary and one referred to the residual; one referred to the journey time; one referred to the number of airlines; and an extra one for the number of competing itineraries within a city-pair.

Regarding the hidden layer structure, according to the universal approximation theorem, a single-hidden-layer neural network can theoretically learn any pattern (Hornik, 1991). However, this was stated when deep learning was not available, and therefore neither were more complex representation of patterns in data (Heaton, 2015). Instead, two hidden layers are expected to allow more complex representation of patterns in data than a single hidden layer, while maintaining the training simplicity. Consequently, the chosen configuration for the neural network presented in this research is two hidden layers.

The number of hidden neurons for each of the hidden layers is decided through an experimental process. However, Heaton (2015) states a few rule-of-thumb methods that can be considered when designing this experimental process. The set of rules-of-thumb are as follows:

- The number of hidden neurons should be between the size of the input layer and the size of the output layer.
- The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.
- The number of hidden neurons should be less than twice the size of the input layer.

Considering the rules above, the experimental process is designed as follows:

for the first hidden layer a range between 10 to 30 neurons, in steps of 5 are tried; while for the second hidden layer, which always contains a smaller or equal number of hidden neurons than the first hidden layer, a range between 5 to 30 neurons is evaluated. In total a combination of 20 NN structures are evaluated. Each NN architecture is noted as $NN(n_{1st}/n_{2nd})$, where n_{1st} denotes the number of hidden neurons in the first hidden layer and n_{2nd} denotes the number of hidden neurons for the second hidden layer. Table 3.20 shows the different neural network topologies that have been considered during the development of an air itinerary share model using artificial neural networks.

	Hidden layer		Notation	Hidden layer		Notation
	1st	2nd		1st	2nd	
Number of hidden neurons	10	5	NN10/5	25	10	NN25/10
	10	10	NN10/10	25	15	NN25/15
	15	5	NN15/5	25	20	NN25/20
	15	10	NN15/10	25	25	NN25/25
	15	15	NN15/15	30	5	NN30/5
	20	5	NN20/5	30	10	NN30/10
	20	10	NN20/10	30	15	NN30/15
	20	15	NN20/15	30	20	NN30/20
	20	20	NN20/20	30	25	NN30/25
	25	5	NN25/5	30	30	NN30/30

Table 3.20: Neural Network topologies considered during the estimation process of an air itinerary choice model using artificial neural networks.

In terms of propagation function for all the neurons a weighted sum (Eq. 3.22), which is one of the most commonly used propagation function, is used. As activation function for the input and hidden layers the hyperbolic tangent ($tanh$) function, presented in Equation 3.23 is used over other activation functions as recommended by Heaton (2015) and because through the exploration phase this activation function resulted in the lowest mean square error when compared to a sigmoid function (further information regarding the exploration phase to determine the the NN model specification can be found in Appendix E). The hyperbolic tangent function is also sigmoidal (s-shaped), however output values range between -1 and 1, allowing more differentiation between negative input values than the logistic sigmoid function. The logistic sigmoid function will map all negative inputs close to zero, without differentiation; this might result in model parameters being updated less regularly than expected during training, and consequently, the neural network might get stuck during training. By allowing further mapping differentiation through using the hyperbolic tangent function (i.e. strong negative inputs will be mapped to more negative values and only near-zero inputs will be mapped to near-zero outputs), the neural network is less likely to get stuck during training.

$$\phi(x) = tanh(x) \tag{3.23}$$

This is one of the most widely used activation functions and outputs values in the range between -1 and 1. The hyperbolic tangent function is applied in the input and hidden layers. Therefore, input variables are normalised to take values between [-1,1] for continuous variables, and discrete values of -1 or 1 for dummy variables.

For the output layer a single neuron is defined (n_o). Since the output is the passenger share for a given itinerary (i.e. a value between 0 and 1) a sigmoid activation function is used so that no transformation is needed. The sigmoid function is represented by Equation 3.24 and the output values range from 0 to 1.

$$\phi(x) = \frac{1}{1 + e^{-x}} \tag{3.24}$$

Training a NN model is an iterative process in which the weights set feeding into the neurons will commonly start with random values until the optimal solution is found. To compare the different architectures and find the optimal neural network configuration, 10-fold cross validation is performed for each of the different neural network architectures (Table 3.20) and then compared. The best performing configuration will be the one chosen.

From previous exploration of NN application to air itinerary share models, Levenberg-Marquart algorithm (LMA) is chosen as training algorithm. Previous exploratory analyses include backpropagation and backpropagation with momentum as alternative training algorithm[30]. However, results with LMA outperformed those

[30] For further information regarding the exploratory study of NN using backpropagation and backpropagation with momentum, please refer to Appendix E

from the other two algorithms in terms of low mean square error during validation, low number of epochs taken to complete the training process and lowest training time.

LMA is a hybrid algorithm that is based on Newton's method and on gradient descent (backpropagation), hence it combines the strengths of both, making it a very efficient training method for neural networks. Gradient descent is guaranteed to converge to a local minimum, however it is slow. Newton's method is fast but it often fails to converge. LMA introduces a damping factor (λ) to interpolate between the two, creating the hybrid method.

The training parameters to be adjusted for this purpose are a learning rate (ε) and momentum (α). Learning rate determines how quickly the model is adapted to the problem (i.e. smaller learning rate will require more training iterations given the smaller changes made to the weights each update). A learning rate between 0.2 and 0.9 is the most commonly used in the literature (Kamiyama et al., 1992). Momentum is a learning property that causes the weight change to continue in its current direction, even if the gradient indicates that the weight change should reverse direction. A high momentum might help to move away from a local optimum and most likely find a different optimum. For the damping factor (λ), an initial value of 0.001, a decrease factor of 0.1 and an increase factor of 10 with a maximum λ value of 1e10 is used as suggested in the literature (Heaton, 2015).

3.8.2 NN model results and validation

During the training, the dataset is divided into training data (70%), validation data (15%) and test data (15%). For each topology and training algorithm 10 training runs are computed (i.e. 10-fold cross validation), for which training, validation and test datasets are chosen randomly. From the experimental process, the best performing NN configuration is then estimated using all input data. There are three criteria for selecting the best performing NN architecture: mean square error during training; number of epochs taken to complete the training process[31]; and time taken to train the algorithm.

Figures 3.28, 3.29 and 3.30 present the average training time, number of epochs

[31] The number of epochs refers to the number of complete passes through the training dataset during training.

and mean squared error obtained during the validation process across the 10-fold runs when training the 20 different NN models for air itinerary share estimation. From these 3 figures one can see there is not much difference across the different NN topologies, with exception of few of them taking more time to train. Consequently, the neural network architecture chosen to estimate an air itinerary share model is *NN30/10*, which seems to have a slightly lower mean square error than the rest.

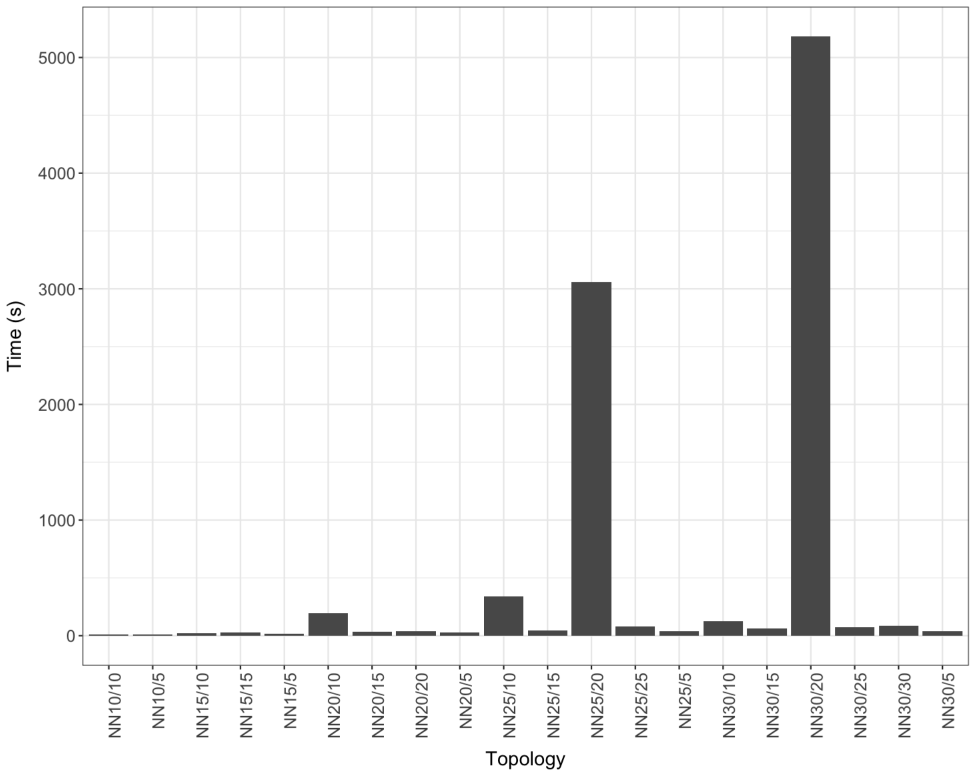

Figure 3.28: Comparison of average training time for all neural network topologies.

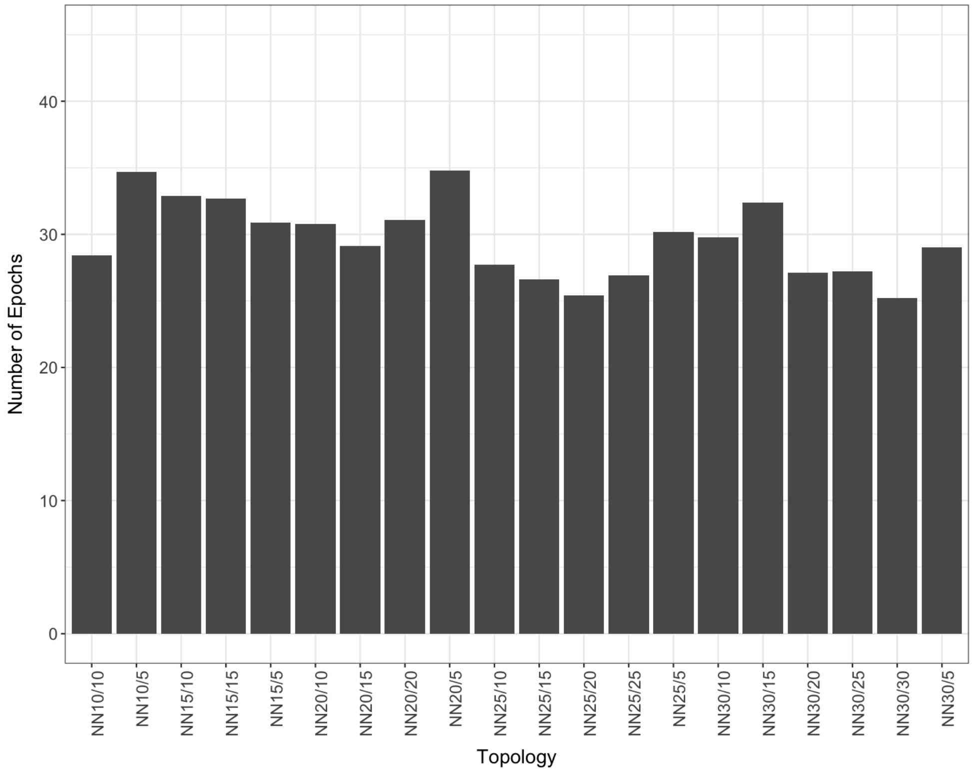

Figure 3.29: Comparison of average number of epochs during training for all neural network topologies.

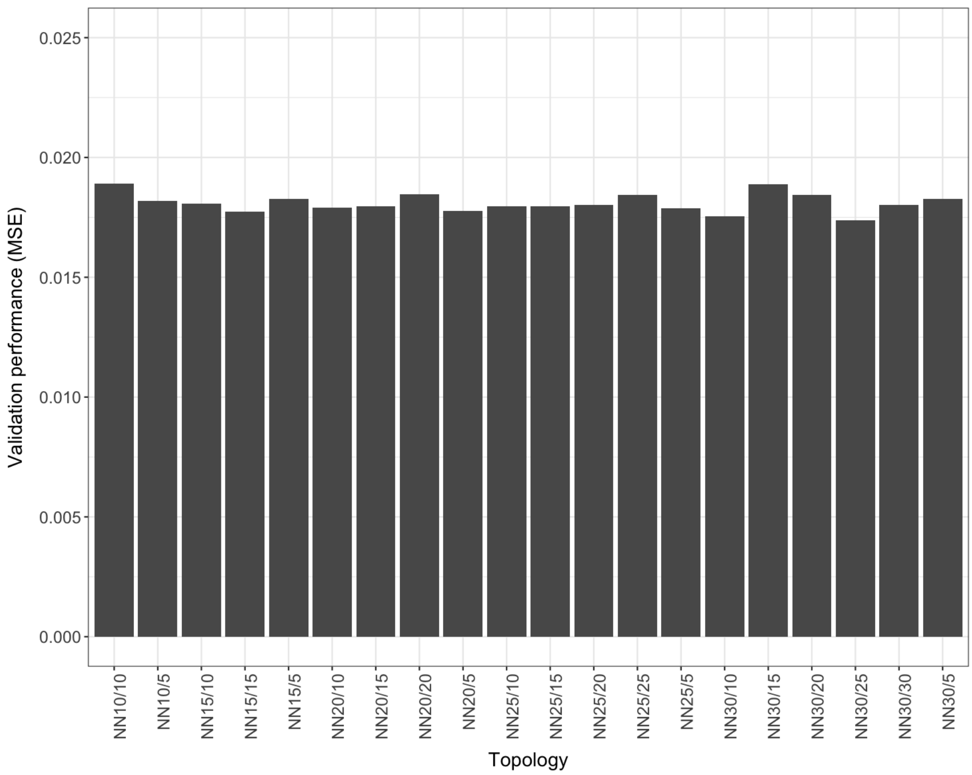

Figure 3.30: Comparison of average Mean Square Error (MSE) of the validation dataset for all neural network topologies.

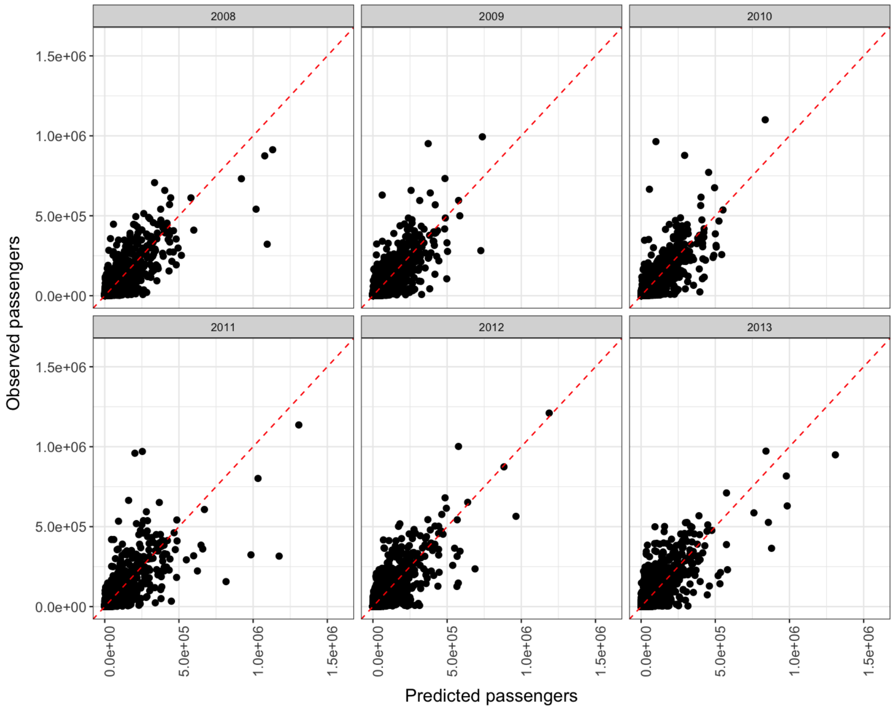

Figure 3.31: Observed against predicted number of passengers throughout the validation years using the estimated neural network model i.e. *NN30/10.*

Similarly to the rest of models presented in this dissertation, validation for this model is done using data for years from 2008 to 2013. Figure 3.31 presents observed passenger demand per itinerary against predicted passenger numbers across the validation years -i.e. between 2008 and 2013-. Overall, predictions from the NN model seem to be slightly more scattered than those obtained from the MNL model, especially for those itineraries with lower demand, which the NN model seem to predict a higher share. Adjusted R^2 is also computed as a measure of predictive power for the neural network model. The average Adjusted R^2 obtained across all the validation years is 0.691, which is lower than that obtained for the multinomial logit model (0.857). However, considering this is a leading research in this topic, results obtained show the potential that machine learning techniques could have when applied for the purpose of estimating itinerary shares. Further exploration of this results is done in the following section through a comparison of both methodologies (i.e. multinomial logit and neural network).

3.9 Module 2: Comparison of the multinomial logit model and the neural network model

Comparison of predictive powers of both models is done by comparing the adjusted R^2 obtained for each of the validation years (2008-2013) (Table 3.21). Overall, the MNL model outperforms the NN model, which can be also seen when comparing Figures 3.25 and 3.31, with predictions closer to the observed values for the MNL model. It is interesting that, for 2012, performance of the NN model seems to be improved with respect to the rest of the years; however, for the MNL model the predictive power for 2012 is the second lowest obtained throughout the validation years. This might be an indication that the NN model is influenced differently by the input variables than the MNL model, and therefore a different specification might be needed, such as considering a different input set of variables.

Models are also compared at the aggregate level by calculating annual revenue passenger kilometres (RPK). Comparison for the validation years is shown in Figure 3.32 where actual RPK are also included. Results in Figure 3.32 show how both models follow the same trend as actual RPK throughout the year range considered. However, both model predictions are slightly higher than the observed trends, with the MNL results slightly closer to those actuals than the NN results. Since the input data is the actual data for all the years, the difference in output is assumed to be due to both models predicting a smaller proportion of non-stop passengers than observed, with the MNL model being slightly better at getting this proportion correct. This hypothesis is confirmed through Figure 3.33, which clearly shows the differences between both models, with the NN model tending to forecast a larger proportion of one-stop passengers. It might be the case that the NN model is more susceptible to changes in journey fares and time than the MNL model, which follows a smoother trend more in line with the observed proportion of non-stop passengers.

These results highlight a potential line of research for further work[32], by which different model specifications could be investigated to understand which are the most influencing factors that affect the neural network model as well as whether any other factors that have not being taken into account might also influence the NN model

[32] Neural network models function as 'black boxes' and do not have the degree of interpretability that statistical models have, and therefore at this stage it is not possible to identify which factors influence the most the NN model.

results, such as the inclusion of fuel prices.

Model	2008	2009	2010	2011	2012	2013	Average
MNL	0.849	0.870	0.874	0.867	0.846	0.838	0.857
NN	0.730	0.709	0.713	0.619	0.722	0.657	0.691

Table 3.21: Comparison of adjusted R^2 obtained for each of the validation years when using the MNL model and the NN model.

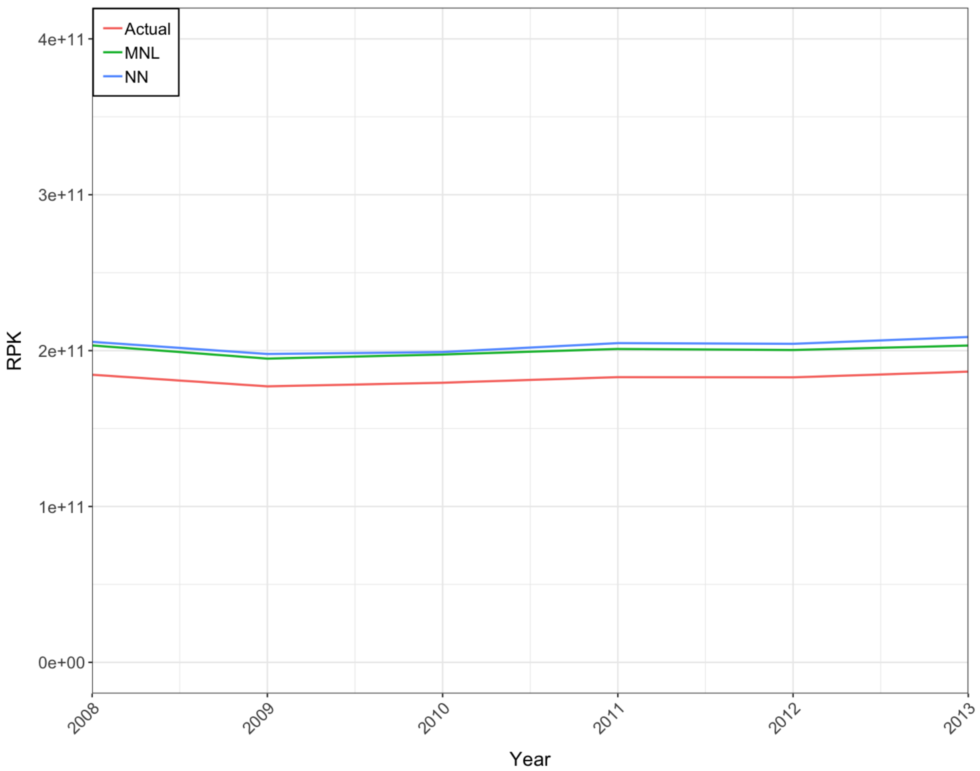

Figure 3.32: RPK computed for the validation years (2008-2013) based on validation results obtained from the MNL and the NN model. Actual RPK values are also included.

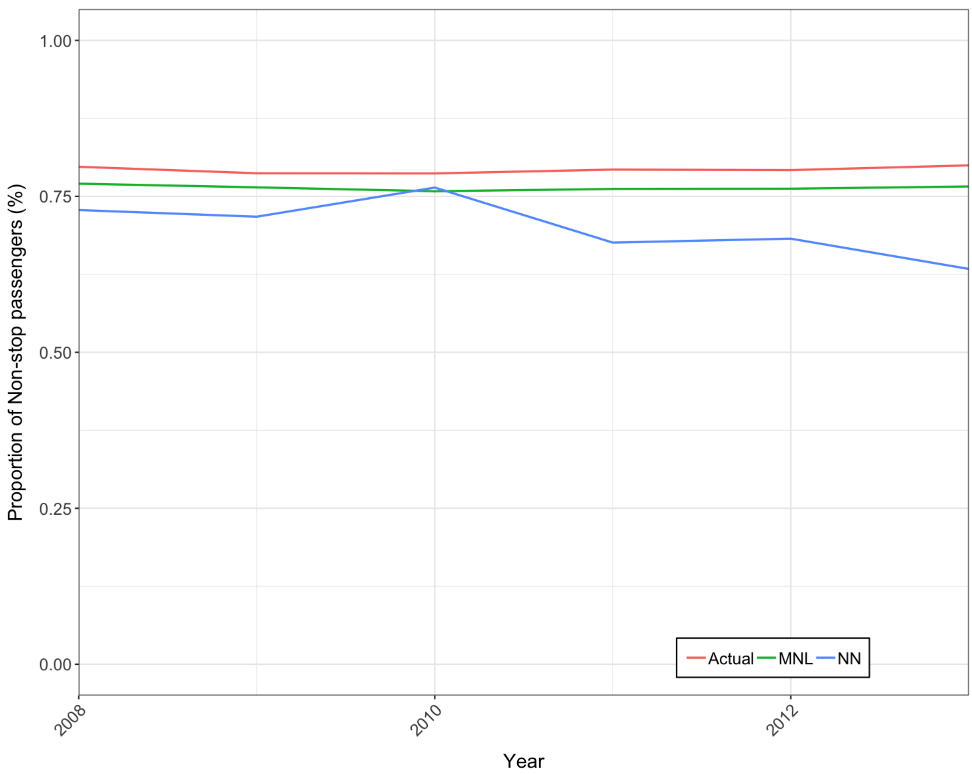

Figure 3.33: Predicted and actual proportion of non-stop passengers throughout the validation years (2008-2013).

3.10 Module 3: Flight frequency model

The third and last stage of the modelling framework is to estimate air traffic levels between airport-pairs (i.e. at the segment level). This is done by using a model that projects the number of flights between two airports within the US ATS, given the segment passenger demand estimated in the second stage of the modelling framework presented in this dissertation. Since the output from the itinerary choice model is number of passengers choosing a specific itinerary, which in turn might be formed of more than one flight (in the case of one-stop itineraries), an intermediate step is necessary to convert itinerary passenger demand into segment passenger demand.

For non-stop itineraries, the compilation process is straightforward, since all passengers will be assigned to a single flight segment. For one-stop itineraries the compilation process breaks down the itinerary into the two flight segments forming that itinerary and assigns the total number of passengers predicted for that itinerary to both legs. Once all itineraries are broken down to segment level, passenger demand is aggregated at the airport-pair level.

An early attempt developed within this research to model air traffic levels considered a model specification that included an auto-regressive term (i.e. the number of flights that were operating on a segment the previous year) as one of the input variables. Following this previous attempt, the investigation has been extended to an alternative model specification without an auto-regressive term. In this section both specifications will be presented and validated. For reference, in the following sections the model containing the auto-regressive term as one of the input variables is referred as Model 1-2SLS, whereas the model without the auto-regressive term is referred as Model 2-OLS.

3.10.1 Model specification: Model 1-2SLS

For the model including the auto-regressive term (Model 1) the estimation process used is a 2-stage least squares (2SLS). The use of a 2-stage estimator is due to possible endogeneity issues arising from the inclusion of the auto-regressive term, since it is expected that there will be correlation between this variable and the error term. This is because there is an issue of simultaneity, where the auto-regressive term influences a change to the dependent variable (i.e. flight frequency), but at the same time a change in the dependent variable also influences a change to the auto-regressive term. The 2SLS estimator with instrumental variables (IV) methodology is a widely used method to correct endogeneity, as previously proved when used to estimate the itinerary choice model to solve fare endogeneity. Instrumental variables are chosen so that they do not belong in the main equation, but are correlated with the endogenous variable. In this case, the selected instrumental variables are airport-pair degree and eigenvector centrality, which are two of the network theory metrics presented in Section 3.6 and which capture the importance and the influence of an airport-pair.

As the name indicates, the two-stage least squares method is split into two. The first-stage of the model is an ordinary least-square (OLS) regression, with the traffic from the previous year ($traffic_{t\text{-}1}$) as the dependent variable, represented by Equation 3.25.

$$Traffic_{ij,t-1} = \alpha IV_{ij} + \gamma' \dot{x}_{ij} + \mu_{ij} \tag{3.25}$$

where $traffic_{ij,t-1}$ is the air traffic levels associated with airport-pair ij for period *t-1*, IV_{ij} refers to the set of instrumental variables associated with the airport-pair ij, α is the set of coefficients associated with the instrumental variables, γ is the vector of coefficients associated with all exogenous variables, excluding traffic and μ is the error term for stage 1 regression.

Explanatory variables for this first stage regression model include the instrument variables -i.e. airport-pair's degree and eigenvector centrality- along with the set of all other exogenous variables with exception of traffic from the previous year.

$$Traffic_{ij,t} = \beta_{traffic,t-1} traffic'_{ij,t-1} + \beta' x_{ij} + \epsilon_{ij} \tag{3.26}$$

where $\beta_{traffic,t-1}$ is the coefficient associated with traffic from stage 1, $traffic'_{ij,t-1}$ is the predicted traffic between airport-pair ij for period *t-1* during stage 1, β' is the vector of coefficients associated with all exogenous variables, excluding traffic for period *t-1* and x_{ij} is the vector of all exogenous variables excluding traffic for period t.

The second-stage of the model (Equation 3.26) uses, as input variables the traffic from previous year predicted during the first stage of the 2-SLS methodology and the other exogenous regressors, excluding actual flight frequency from previous year.

Tables 3.22 and 3.23 presents the model specification for stage 1 and 2 to estimate air traffic levels between airport-pairs[33]. Note $Traffic'_{ij,t-1}$ in stage 2 has a prime to indicate that it is not the observed value for traffic but the predicted one. Explanatory variables are normalised using equation 3.27. As presented in Table 3.22, along with the auto-regressive term the rest of input variables used at this stage are: passenger demand difference with respect the previous year between airports i and j; and the total number of airports in the cities where the origin (o) and destination d airports are located.

Passenger demand corresponding to the current year for the pair of airports i and j was also considered, but results were not as good as when using passenger demand difference year-on-year. Also, when considering a different data transfor-

[33]Note that Number $\mathrm{Airports}_{\mathrm{od,t}}$ refers to the sum of number of airports in the pair of cities o and d where the pair of airports i and j are located.

mation for the number of airports, such as product or square root of the product of number of airports, these alternatives yielded similar results as when using the simple sum of the number of airports.

Coefficients	Explanatory variables
$\gamma_{ij}^{N_{YoY}}$	$\text{Passenger}_{\text{ij, t}}$ - $\text{Passenger}_{\text{ij, t -1}}$
$\gamma_{od,t}^{NumArpts}$	Number $\text{Airports}_{\text{od,t}}$
$\alpha_{ij,t-1}^{DEG}$	$\text{DEG}_{\text{ij, t-1}}$
$\alpha_{ij,t-1}^{EVC}$	$\text{EVC}_{\text{ij, t-1}}$

Table 3.22: 1st stage model specification.

Coefficients	Explanatory variables
$\beta_{ij,t-1}^{Traffic'}$	$\text{Traffic'}_{\text{ij, t-1}}$
$\beta_{ij}^{N_{YoY}}$	$\text{Passenger}_{\text{ij, t}}$ - $\text{Passenger}_{\text{ij, t -1}}$
$\beta_{od,t}^{NumArpts}$	Number $\text{Airports}_{\text{od,t}}$

Table 3.23: 2nd stage model specification.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.27}$$

To test for endogeneity and to check the validity and relevance of the instrument variables used during the 2-SLS estimation process, a set of specification tests are performed: a test for weak instruments; a Hausman test for endogeneity; and a test for the validity of the instruments.

To test for weak instruments in the 2-SLS equation, the joint significance of the instruments in the 1st stage model (Equation 3.25) is tested. The null hypothesis is that coefficients of the instrument variables are equal to 0 (i.e. H_0: $\alpha = 0$). If the *p-value* obtained is large enough to reject the null hypothesis, it can be confirmed that the instrument variables used are not weak. Another way of looking at instrument weakness is through the *F-statistic*, which if being greater than 10, allows the null hypothesis to be rejected as explained in Section 3.7 (Staiger and Stock, 1994; Guevara and Polanco, 2013).

To check for the validity of the instruments used, a Sargan test (1958) for over-identifying restrictions is performed. Note that this test can be performed only because the number of instruments (L) used is greater than the number of endogenous variables *(k)* (i.e. $L>K$) as opposed to the case when the itinerary choice model was presented (Section 3.7). The null hypothesis is that the covariance between the instruments z and the error term e is zero (H_0: $cov(z,e) = 0$). Therefore, rejecting the null hypothesis indicates that at least one of the extra instruments is not valid.

Without the presence of endogeneity, ordinary least squares will be efficient and consistent whereas IV-estimation will be efficient but inconsistent. However, in the presence of endogeneity, IV-estimation will be both efficient and consistent, while OLS will be inefficient. Consequently, it is necessary to test that the variable assumed to be endogenous is indeed endogenous. This problem is addressed by the Wu-Hausman test for endogeneity which checks the consistency of the OLS estimates under the assumption that the IV-estimates is consistent. Rejecting the null hypothesis means that OLS is not consistent, suggesting that endogeneity is present. This is equivalent to looking at whether the residual obtained in the first stage (i.e. the difference between the predicted and actual flight frequency for the previous year) is statistically significant if added into the second stage equation as was done in Section 3.7 for the multinomial logit model estimated using 2-SCF methodology.

The obtained coefficient estimates for the Model 1-2SLS, specified through Equations 3.25 and 3.26 are presented in Tables 3.24 and 3.25 corresponding to results for the first and second stage respectively. Stage 1 estimation model results (Table 3.24) show that coefficients associated with the two instrument variables used to predict flight frequency from previous year (i.e. airport-pair degree and eigenvector centrality) are statistically significant at the 95% confidence level. This allows the null hypothesis to be rejected, that there is no relationship between the instrument variables and flight frequency from previous year. In addition the *F-statistic* obtained at this 1st stage is well above the critical value of 10, a sign of having sufficiently strong instrument variables.

	Coefficients
(Intercept) (β_0)	0.056 ***
$\mathrm{N_{YoY}}$ ($\gamma_{ij}^{N_{YoY}}$)	-0.037 ***
Num. Airports ($\gamma_{od,t}^{NumArps'}$)	0.057 ***
$\mathrm{DEG_{ij}}$ ($\alpha_{ij,t-1}^{DEG}$)	1.075 ***
$\mathrm{EVC_{ij}}$ ($\alpha_{ij,t-1}^{EVC}$)	-0.550 ***
Adj R-squared	0.37
F-statistic	319

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.24: Coefficient estimates obtained for the first stage of the Model 1-2SLS.

	Coefficients
(Intercept) (β_0)	-0.086 ***
$\mathrm{Traffic'_{t-1}}$ ($\beta_{ij,t-1}^{Traffic'}$)	0.913 ***
$\mathrm{N_{YoY}}$ ($\beta_{ij}^{N_{YoY}}$)	0.229 ***
Num. Airports ($\beta_{od,t}^{NumArps}$)	0.009 ***
Adj R-squared	0.988
Tests	**p-value**
Weak instruments test	<2e-16***
Wu-Hausman test	1.4e-06***
Sargan test	0.6

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.25: Coefficient estimates obtained for the second stage of the Model 1-2SLS.

The inability to reject the null hypothesis of the Sargan test allows us to confirm that the instruments used are valid, while rejecting the null hypothesis of OLS consistency in the Wu-Hausman test suggests that endogeneity is present, related to flight frequency from previous year. All VIF tests indicate absence of multicollinearity[34]. Coefficient estimates obtained for the second stage of the 2-SLS regression show that the auto-regressive term is the most significant variable on the number of flights an airport-pair has associated, followed by the difference in number of passengers year-on-year.

[34] VIF values computed for all the variables are less than 1.5.

3.10.2 Model specification: Model 2-OLS

Predictive power for the model described above (i.e. Model 1- 2SLS) is expected to be high as the adjusted R^2 obtained during training shows. Considering the low variability of flows on a mature system like the US, air traffic levels are expected to be similar to those from the previous year and therefore predictions will tend to be quite accurate. Even in edge cases, such as when a new segment is added to the network, the model might be able to identify air traffic levels well (i.e. because the previous year flight frequency is done via 2SLS the estimate of previous year frequency in this case is not zero and the prediction may not actually be that bad). However, the model fails to identify the influence that other factors might have to air traffic levels, such as aircraft type restrictions or competition, due to the strong influence the auto-regressive term has. In order to further investigate other factors that might influence the air traffic levels amongst airport-pairs within the US air transportation system, a model specification that does not contain the auto-regressive term is also investigated.

In this case the set of input variables defining the model specifications are as follows:

- Passenger demand - Passenger demand associated with a specific airport-pair, airports i and j.
- Hub information - two dummy variables that capture whether origin and/or destination airports are hubs or not. The naming convention used is as follows:
  - *hub1* - 1 if both airports are not hubs; 0 otherwise.
  - *hub2* - 1 if both airports are hubs; 0 otherwise.
- Longest runway length (feet) - given the longest runway for each of the airports forming the airport-pair, the shortest is taken. This variable represents the limitation of aircraft types that can operate in a given segment-pair because larger aircraft types require longer runways.
- Distance (miles) - distance between airport-pairs.
- Load Factor - Average load factor associated to a specific segment.

Different combinations of input variables have been considered, as well as the option of including the number of low-cost carrier operations on a given segment. However, only the best performing model is presented in this section. Flight frequency between airport i and j is calculated using a simple one-equation linear regression model of the form of

$$Traffic_{ij} = \beta_0 + \beta_1 N_{ij} + \beta_2 RW_{ij} + \beta_3 dist_{ij} + \beta_4 LF_{ij} + \sum_d \beta_i D^d_{ij} \tag{3.28}$$

where $Traffic_{ij}$ is the total number of flights between airports i and j, N_{ij} is the passenger demand between airports i and j, RW_{ij} is the shortest runway between the two longest ones of airport i and j, $dist_{ij}$ is the distance between airport i and j, LF_{ij} is the average load factor for the airport-pair ij, D_{od} are the set of 2 dummy variables capturing whether airport i and/or j are hub airports or not, and the set of β are the parameters to be estimated.

An OLS estimator us used for this model, and (similarly to Model 1-2SLS), all continuous variables are normalised by using Equation 3.27.

Coefficient estimates obtained (Table 3.26) show all coefficients statistically significant at the 95% confidence level and no sign of multicollinearity. Results also show how passenger demand at the segment level is the most influencing factor. It is interesting to see how runway length and distance seem to have a negative influence in terms of number of flights in a segment. Regarding length of runways, this might be influenced by the fact that long runway can handle large aircraft operations, decreasing the need for frequency since they can carry a larger number of passengers. In terms of distance, the further a segment is, the longer the flight time will be, and hence, less flight frequency is expected. It is also the case that longer distances will usually need aircraft with higher range which tend to be those that can carry a larger number of passengers. This will consequently decrease the number of flights needed to serve all the passenger demand. Finally, load factor coefficient is negative, which is a reflection of the fact that for the same number of passengers higher load factor means fewer flights.

	Coefficients
(Intercept) (β_0)	0.132 ***
N (β_1)	0.845 ***
RW (β_2)	-0.018 ***
dist (β_3)	-0.204 ***
LF (β_4)	-0.074 ***
hub1 (β_5)	-0.019 ***
hub2 (β_6)	0.015 ***
$Adjusted R^2$	0.877

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.26: Coefficient estimates obtained for the Model 2-OLS.

3.10.3 Model validation: Model 1-2SLS and Model 2-OLS

As done with previous stages, in order to validate the models and assess their predictive powers, actual data between 2008 and 2013 is used to generate predictions of air traffic levels across airport-pairs. For the Model 1-2SLS, estimated results presented in Tables 3.24 and 3.25 are applied as coefficients for Equations 3.25 and 3.26 respectively, so that air traffic levels between airport-pairs are predicted. For the Model 1-OLS, estimated results presented in Table 3.26 are used in Equation 3.28 to generate predictions of flight frequency for pair of airports within the US ATS.

The adjusted coefficient of determination (adjusted R^2), which indicates how close the predictions of air traffic levels are to the observed air traffic levels, is calculated and presented in Table 3.27. As expected, the Model 1-2SLS has a higher adjusted R^2 since the US air transportation system is a mature system and most of the routes are expected to be at steady-state and similar to those from previous years. However, results for the Model 2-OLS also show a good predictive power with an average adjusted R^2 of 0.884.

Figures 3.34 and 3.35 show the comparison between predicted and observed number of flights across the validation years for Models 1-2SLS and 2-OLS respectively. Comparing both Figures (3.34 and 3.35), one can again notice that Model 1-2SLS results are more accurate than those when Model 2-OLS is used since values are more spread horizontally.

Model	2008	2009	2010	2011	2012	2013	Average
Model 1-2SLS	0.985	0.980	0.985	0.986	0.989	0.989	0.986
Model 2-OLS	0.880	0.877	0.877	0.889	0.891	0.891	0.884

Table 3.27: Comparison of adjusted R^2 obtained for each of the validation years when using the Model 1-2SLS and the Model 2-OLS.

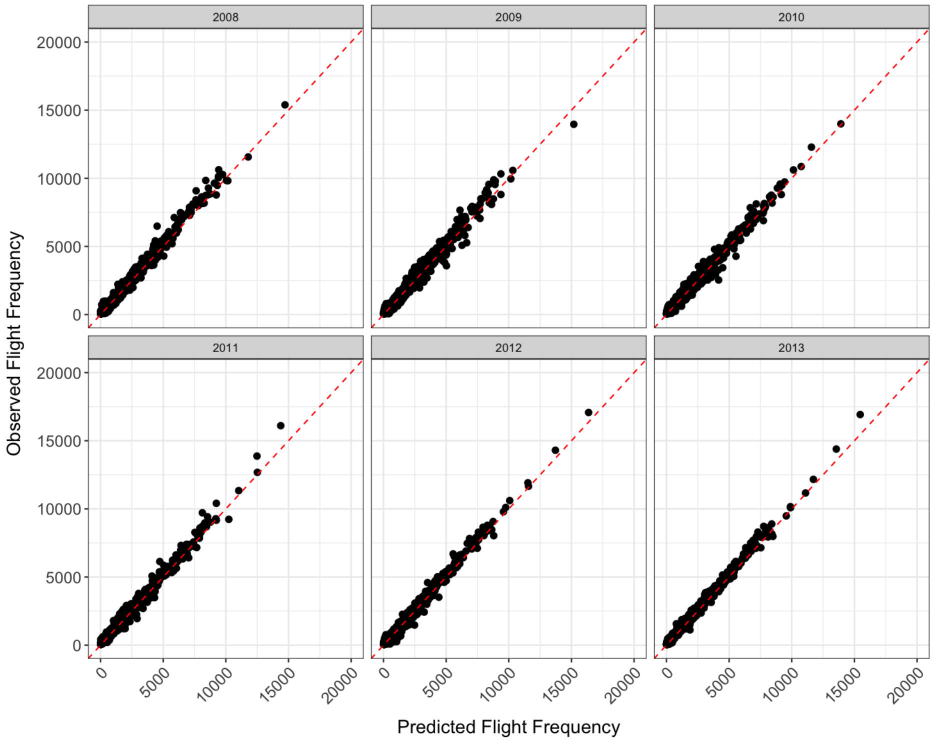

Figure 3.34: Comparison between observed and predicted air traffic levels across validation years.

Finally, the predictive powers of these 2 models have been compared for the edge case when a new link is added into the network. In such cases, the observed flight frequency from the previous year is 0; however, because this value is computed via the 2-stage formulation, in the Model 1-2SLS the flight frequency from the previous year value used is not 0. Considering only those edge cases (i.e. a new link is added), which are in total 264 data points across all the validation period, the mean square error is calculated as metric for comparison. MSE obtained from both models associated to this edge cases is presented in Table 3.28 and, surprisingly, the Model 1-2SLS has a lower MSE, which would suggest that the predicted flight frequency from the previous years helps to obtain a more accurate estimation of the number of flights for those new links being added into the network. However, the Model 1-2SLS offers the caveat of being over-influenced by the auto-regressive term, making

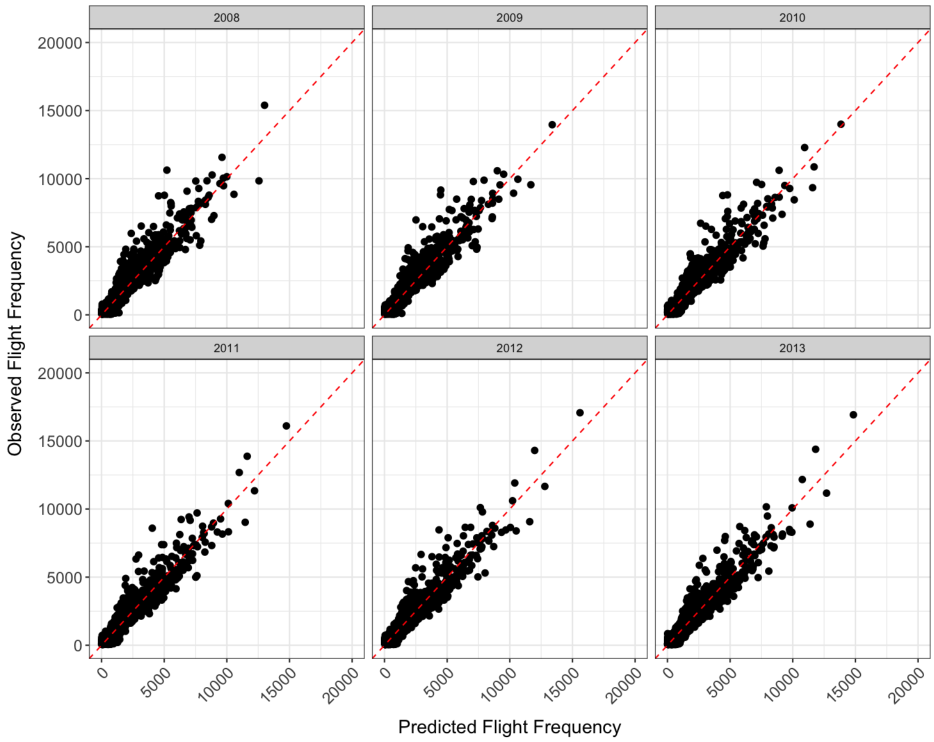

Figure 3.35: Comparison between observed and predicted air traffic levels across validation years for model OLS-3.

it difficult to highlight which other factors might affect air traffic levels.

	2-SLS	OLS
MSE	29,113	108,515

Table 3.28: MSE obtained from applying Model 1-2SLS and Model 2-OLS to those cases were a new link is added into the network.

3.11 Summary

The modelling framework presented in this dissertation looks at city-demand generation, itinerary demand assignment and air traffic estimation. This chapter describes the methodology and validation behind the main contributions contributions of this research: the use of network theory, with other variables, to analyse the evolution of the US air transportation system and the effect that such changes has to the rest of the network; the study of alternative techniques to develop an air itinerary share model at the most aggregate level; and to develop a single modelling framework including both airport connectivity and air itinerary share modelling. Findings obtained through the estimation process of the modelling framework presented in this chapter can be summarised as follow.

O&D passenger demand model

- As expected and widely stated in the literature, demographic characteristics (i.e. income and population) correlate strongly with O&D passenger demand. Although income elasticities obtained in this research are not as high as those reported in the literature, a lot of the variation in income elasticity has been found depending on specific model formulation, which is supported by the literature (2017, 2007).

- High generalised cost elasticity obtained is in line with those obtained in the literature and shows how higher fares will have a negative impact on O&D passenger demand.

- O&D passenger demand for different distance groups needs to be modelled differently. Previous attempts to model O&D passenger demand with a single model to fit all city-pairs within the US ATS yielded in poorer results. Splitting city-pairs into three distance groups (i.e. short-, medium- and long-haul) and estimating a set of model parameters for each of these distance groups has improved the predictive power of the models quite significantly and relatively larger R^2 (>0.8) than those found in the literature for the US ATS have been obtained (Dray et al., 2014).

Airport connectivity model

- Historical data shows a low variability of the US network connectivity, serving as evidence that the US air transport network is a mature system and relatively in steady state in terms of connectivity changes.

- From the application of network theory metrics to study the US air transportation system few conclusions can be drawn. The US ATS follows clearly a power-law degree distribution, by which most of the nodes have only few links while only few of the nodes have a high degree. This is clearly in-line with a hub-and-spoke network. Clustering coefficient (CC), which measures the density of the connections, is in most of the cases inversely proportional to node degree.

- Although several approaches have been explored, and link addition and removal models presented in this chapter represent an improvement with respect previous attempts, precision levels are still relatively low. Because the system is in steady state, the influence of hard-to-measure factors, such as airport incentives to airlines, is relatively greater than it would be in a growing system.

- With precision levels obtained of just below 20% for both link addition and removal models, results are in line with the literature (2012) for the link addition model and slightly worse for the link removal model (i.e. results in the literature show 40% accuracy). However, a lower false positive rate for the link addition model to that in the literature (i.e. 82.3% as opposed to 90%) suggests the potential of including a set of variables beyond those associated with network theory, as done by Kotegawa (2012), to improve these type of models aim to predict connectivity changes.

- Although the predictive powers of the estimated link addition and removal models predictive power are low, and therefore they are not suitable for predicting the exact airport-pair connectivity changes, results show that the models seem to be able to capture the evolution of the network at an aggregate level.

Itinerary choice model

- Two issues can strongly affect itinerary share models: computational limitations and fare endogeneity. The Berkson-Theil approximation has proved to be a good solution to solve the problem of computational limitations. The implementation of this methodology has allowed the development of an air itinerary share model without any run-time issues. Moreover, the use of this methodology with WLS estimation process, not yet used in the literature, implies an improvement of the predictive power (adjusted R^2 of 0.85) of this technique when used to model itinerary shares, becoming comparable to the most commonly used technique of multinomial logit with maximum likelihood estimation process. Fare endogeneity led to much higher value of times (VOT) than those found in the literature in previous attempts to model itinerary shares. The 2-stage control function method has proved to be a good methodology to solve fare endogeneity, reducing the VOT to a level comparable to those found in the literature (\$41.3/h).

- A neural network has been used as an innovative and alternative technique to model itinerary shares at the network level. Validation results for this model show an average adjusted R^2 value of 0.69, slightly lower than those obtained with more commonly used techniques, such as multinomial logit; however, results also show the potential of further investigating the application of this machine learning technique to model air itinerary shares since the NN model presents a slightly different response to changes in journey fare and times than those seen from the multinomial logit model.

Air traffic level model

- Including the auto-regressive term as one of the input variables to model the air traffic levels between two airports yields a model with high predictive powers (average adjusted R^2 of 0.99). However, in a steady-state network air traffic levels are expected to be similar to those from the previous year.
- Comparison between the Model 1-2-SLS, which includes the auto-regressive term, and the Model 2-OLS shows that the former outperforms the latter. However, predictive power for the Model 2-OLS is only slightly lower (0.885).
- Surprisingly, in the edge case of new links added into the network, the model with the auto-regressive term yields with a lower MSE, which would justify the selection of this model for future use. However, the 2-SLS model offers the caveat of being over-influenced by the auto-regressive term, making it difficult to highlight which other factors might influence air traffic levels.

Chapter 4

Example application of the modelling framework: US Air Transportation System

In the previous chapter (3) the validity of each of the sub-models within the modelling framework presented in this dissertation has been demonstrated. However, the validation process was performed in isolation without considering the interactions amongst the different models and using actual data for all validation years (between 2008 and 2013). In this chapter, the integrated modelling framework is validated by projecting air traffic levels from 2008 to 2025 and considering interaction amongst the sub-models (i.e. predicted O&D passenger demand will be used to generate predictions of itinerary shares).

Obtained projections are evaluated and assessed in two ways: first, to test the integrated model predictive power, projections generated for years between 2008 and 2018 are compared to actual values; second, overall projections (i.e. up to 2025) are analysed to look at how the modelling framework presented in this dissertation projects air traffic levels into the future and how sensitive it is to the influencing factors that have been considered within the modelling framework, such as changes in airfares. Projections are generated up to 2025 so that they are aligned with the FAA's projections generated in 2008 and the comparison can be made. However, it is theoretically possible to use this model to project to later time horizons.

Results for each stage of the modelling framework are presented separately along with a description of all considerations and assumptions taken to produce these projections. Two different set of projections are produced: one without considering the modelled evolution of the network -i.e. it does not consider the results of the airport connectivity model (case 1); and another set of projections where the results from the airport connectivity model are taken into account (case 2). Note that for this example application of the modelling framework, the model does not consider the impact of flight time changes due to network evolution to the overall demand (O&D passenger demand). This means that network evolution in case 2 will only have an impact in itinerary shares and traffic levels. Considering the amount of uncertainty on how other factors will also change (e.g. airfares) and that flight time is introduced in the model through generalised cost, it is assumed the impact might not be significant (i.e. a new shorter route might have a lower flight time but might imply a greater airfare, which would balance the changes). For example, adding a direct route between Arcata (north of California) and Los Angeles would decrease the average flight time by about 1.3 hours (i.e. current one-stop itineraries take approximately 4 hours via SFO); if market fare is assumed to balance the decrease in flight time in the generalised cost term from the demand equation, in this example airfare should increase by around \$52/h, which is well within the range of uncertainty in fare changes for moving to a direct route (i.e. value of time obtained in this dissertation was \$41.3/h); note that the current difference between average market fare of one-stop routes and when direct routes are added in the Arcata-Los Angeles market is actually \$45.3/h, which is consistent with the \$52/h obtained above. Besides, because most of the new links added are in city-pairs which have existing direct links already, the change in fare and time over existing links will only be small (most of the passengers on the new link would be expected to have shifted from other carriers).

For each of the cases two subsets of projections are generated based on the two models developed to model itinerary shares: the multinomial logit model and the neural network model. Also, for case 1 (without considering the modelled network evolution) air traffic levels are projected using both models developed to estimate flight frequency between airport-pairs: the Model 1-2SLS and the Model 2-OLS. For illustration purposes, Table 4.1 shows the different sets of projections that are being presented in this Chapter.

The rest of this chapter is structured as follows: projections for the O&D model are presented in Section 4.1, while Section 4.2 presents projections obtained from the airport connectivity model; Case 1 projections, without considering network evolution, are presented in Section 4.3, with information regarding projections for the itinerary share model and air traffic levels. Case 2 projections are presented in Section 4.4, which also includes projected itinerary shares and air traffic levels. Finally, a comparison between case 1 and case 2 projections is presented in Section 4.5.

Network evolution	Itinerary choice model	Airfare scenario	Air traffic model used
Case 1: No network evolution	MNL	Low	Model 2-OLS
		Medium	Model 2-OLS
		High	Model 2-OLS
		Low	Model 1-2SLS
		Medium	Model 1-2SLS
		High	Model 1-2SLS
	NN	Low	Model 2-OLS
		Medium	Model 2-OLS
		High	Model 2-OLS
		Low	Model 1-2SLS
		Medium	Model 1-2SLS
		High	Model 1-2SLS
Case 2: Network evolution	MNL	Low	Model 2-OLS
		Medium	Model 2-OLS
		High	Model 2-OLS
	NN	Low	Model 2-OLS
		Medium	Model 2-OLS
		High	Model 2-OLS

Table 4.1: Set of projections presented in this Chapter.

4.1 O&D passenger demand model

The O&D demand model projects true origin-ultimate destination passenger demand between a set of US domestic city-pairs. Based on the distance between the city-pairs, a model has been estimated for each of the three subsets defined as per below:

- City-pairs between 186 and 400 miles apart;
- City-pairs between 400 and 2113 miles apart;
- City-pairs more than 2113 miles apart;

In order to generate projections only those O&D with more than 1,000 passenger a year are considered, making an average of 4,238 city-pairs per year. From 2018 onwards, the set of city-pairs to have passenger demand projected has been considered to be the same as those in 2018.

For all the three O&D passenger demand models a linear regression model with logarithmic transformation is used (Equation 3.1). Model specification for each of the three models is the same with the exception of the set of dummy variables included for each distance group. Based on Equation 3.1, the input variables influencing changes in passenger demand for which projections are also needed are: population, mean household income per capita and generalised cost. The set of dummy variables included in each model are assumed to remain the same throughout all the projections years (2008-2025).

For population, projection estimates (Figure 4.1) have been extracted from the US Census Bureau (2014a) for years up to 2017 and from the several state governments websites for further years (see Appendix A for further details on the several sources consulted). While economic growth (Figure 4.2), which is defined as mean household income per capita, has been sourced from the US Bureau of Economic Analysis (BEA) (2014).

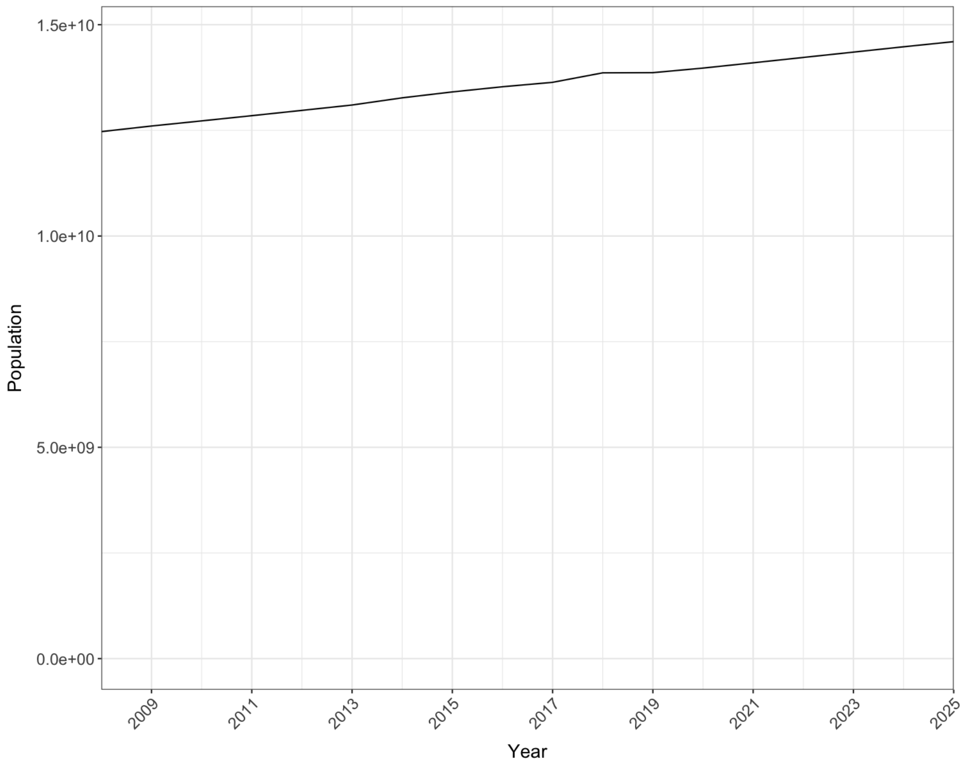

Figure 4.1: Population projections.

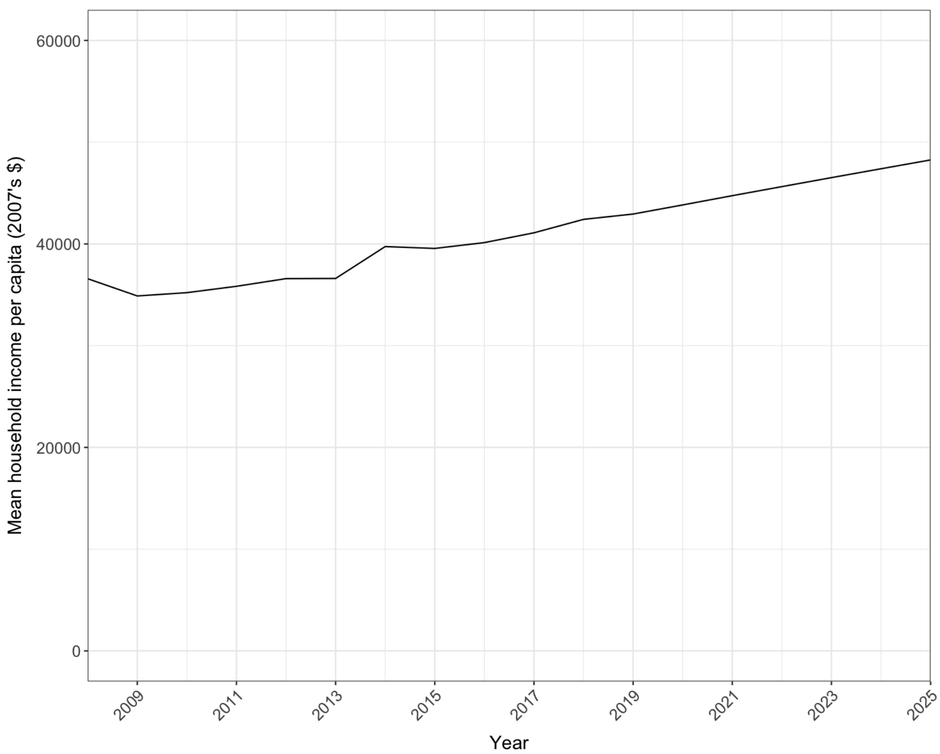

Figure 4.2: Mean household income projections.

Generalised cost is formed by the following information: airfares, flight time, delay and value of time. Flight time and delay are considered to not change throughout the years. The assumption of a constant delay assumes that airport capacity is added to maintain delays at existing levels in line with Dray et al. (2009). While this may be a poor assumption in the short term, it can be considered a better assumption in the long term. Value of time and value of delayed time is assumed to have an annual growth rate of 0.5% from 2015 onwards. This corresponds to the average annual growth rate between 2009 and 2014[1]. Finally, different scenarios have been considered for airfares.

Between 2008 and 2015, actual airfares data are used, which are obtained from DB1B (2014). To generate projections into the future, airfares have been linked to oil fossil fuel price projections[2] to create 3 different scenarios (i.e. low, central and high). Figure 4.3 shows actual oil fossil fuel prices up to 2015 and projections across the 3 different scenarios from 2016 up to 2040.

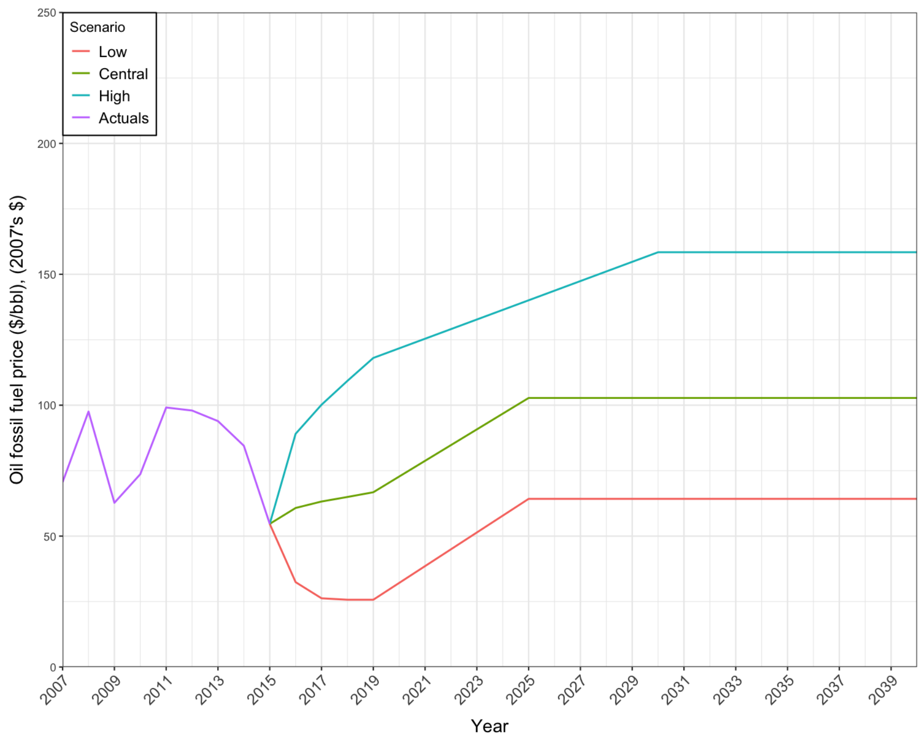

Figure 4.3: Oil fossil fuel price actuals and projections.

Figure 4.4 presents the comparison between airfares weighted mean and oil fossil fuel price between 2007 and 2013, which clearly shows that although airfares are affected by oil fossil fuel prices, the percentage change year on year in airfares is not as steep as it is in fossil fuel prices. Rate of change year on year for fares

[1] For further information regarding how the value of time and value of delayed time is calculated, please refer to Appendix B.

[2] Oil fossil fuel price projections have been obtained from DECC (2015) which are assumed global.

and oil fossil fuel prices is presented in Table 4.2. In order to project airfares for the three scenarios considered (and presented in Figure 4.4), a simple relationship between oil fossil fuel price change year-on-year and airfare change year-on-year has been assumed since airline economics are not part of the scope of the thesis and the underlying relationship between oil price and fares depends on multiple factors related to airline economics, such as level of competition, hedging strategies, non-fuel costs, etc. The simple relationship between airfare and oil fossil fuel prices, which is based on historical oil and fuel price variation over the modelling period, considers three cases associated with the impact that the year-on-year change on oil fossil fuel price has in airfares; these three cases are as per below (i.e. projected airfares for the period 2007-2013 have also been added into Figure 4.4):

- Positive Year-on-Year (YoY) change - Fares have been assumed to increase by 1/7 of the oil fossil fuel price change;
- No YoY change - Fares have been assumed also without a YoY change;
- Negative YoY change - Fares have been assumed to decrease by 1% with respect to the previous year fare.

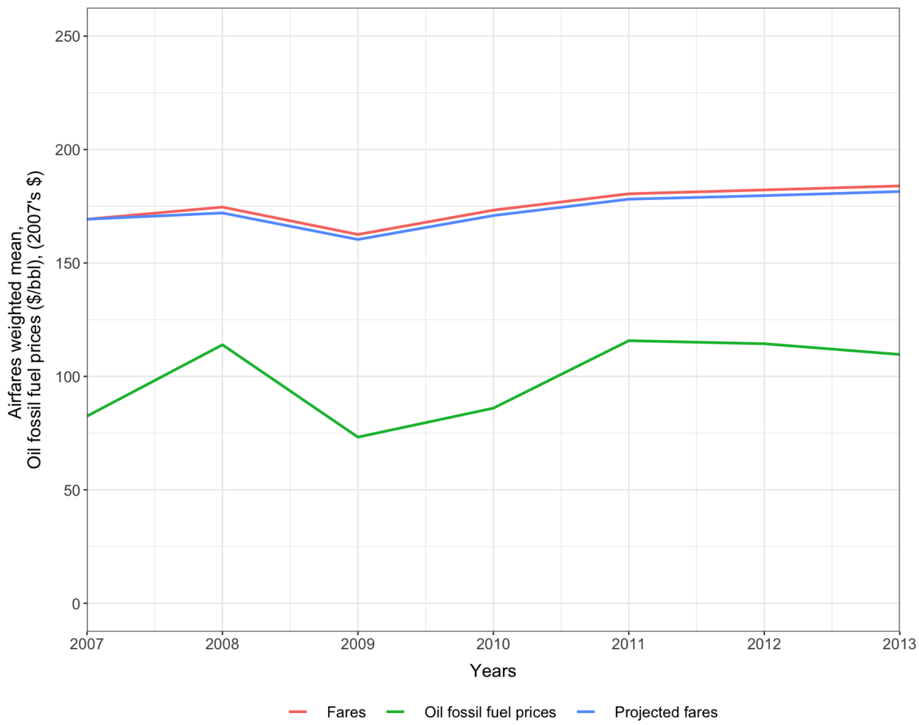

Figure 4.4: Comparison between airfare weighted mean and oil fossil fuel price between 2007 and 2013.

	Years					
	2008	2009	2010	2011	2012	2013
Fares weighted mean	3.2%	-6.9%	6.6%	4.2%	0.9%	0.96%
Oil fossil fuel price	38.2%	-3.6%	17.4%	34.6%	-1.1%	-4.1%

Table 4.2: Rate of change year on year comparison between oil fossil fuel and average airfares.

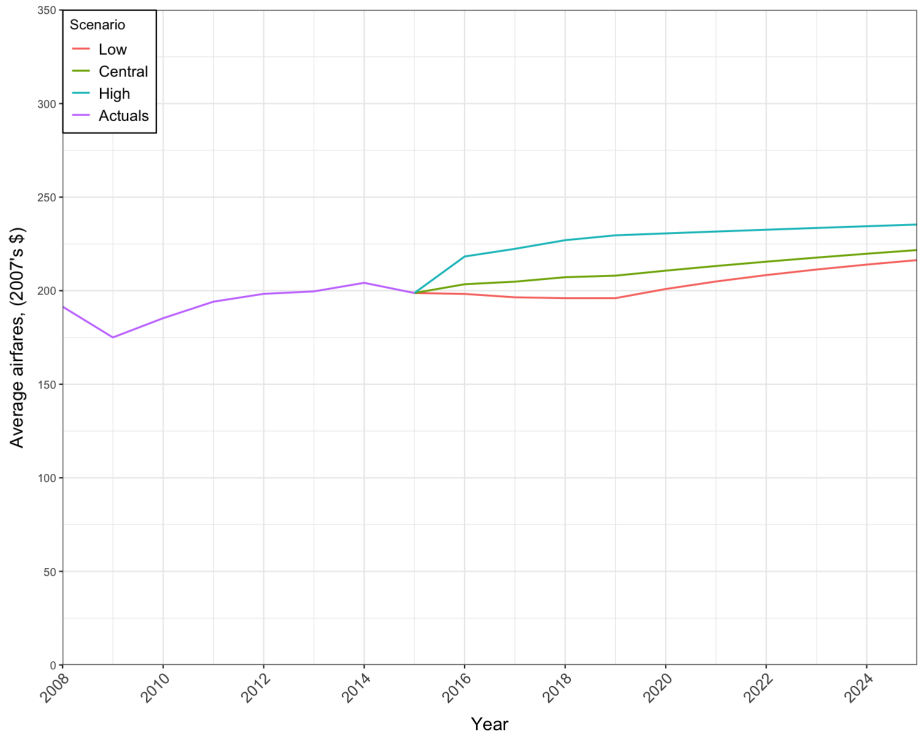

Figure 4.5: Mean airfare projections for low, central and high scenarios. Actual mean airfare up to 2015.

To generate projections, city-pairs are split into 3 groups based on the distance group they belong to (i.e. short-, medium- or long-haul). For each of the city-pairs within each subset O&D passenger demand is projected up to 2025. To obtain system-wide passenger demand, passenger projections for each city-pair are aggregated. Table 4.3 shows the percentage difference between predicted and observed total passenger demand for each of the models by distance group. From results shown in Table 4.3, it is clear that the economic crisis in 2009 has an impact on the models' results, as seen and discussed in Chapter 3 during the validation of this sub-model. During 2009's economic crisis, average airfares suffered a decrease (Figure 4.4) most likely to promote air travel amongst the population. Since passenger demand is inversely proportional to airfares, and in this case to generalised cost (Section 3.4), a decrease in airfares prompted the models to forecast a higher passenger demand than expected.

	short-haul			medium-haul			long-haul		
Year	Low	Central	High	Low	Central	High	Low	Central	High
2008		7.8%			6.3%			7.7%	
2009		8.8%			14.5%			17.7%	
2010		3.5%			9.0%			10.6%	
2011		2%			6.0%			7.0%	
2012		6.3%			7.2%			9.2%	
2013		4%			4.0%			10.4%	
2014		5.4%			9.9%			10.1%	
2015		6.2%			12.0%			12.2%	
2016	8.1%	4.2%	-5.7%	9.1%	5.6%	-3.4%	11.2%	7.5%	-2.3%
2017	11.8%	5.5%	-6.1%	9.1%	3.5%	-6.8%	14.0%	7.9%	-3.5%
2018	19.6%	10.7%	-2.8%	10.9%	2.7%	-8.8%	15.8%	7.7%	-4.9%

Table 4.3: Difference between projected and observed total passenger demand by distance group for years between 2008 and 2018.

Figures 4.6, 4.7 and 4.8 show the total projected O&D demand for city-pairs considered for each distance group model (short-, medium- and long-haul) respectively. Actual passenger demand levels are also included up to 2018.

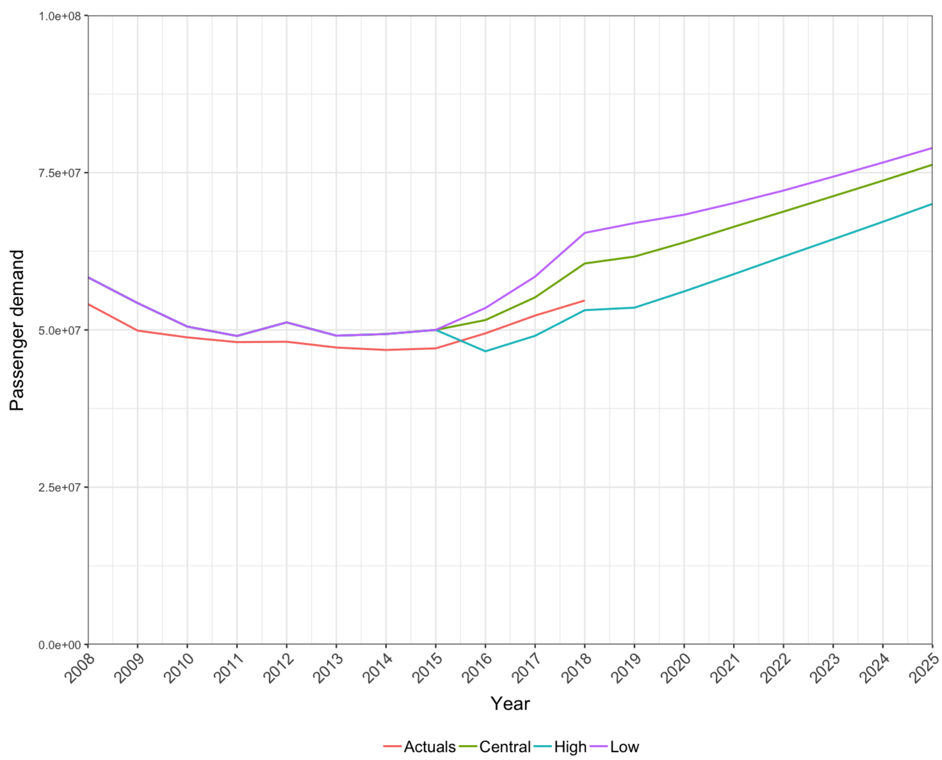

Figure 4.6: Projected O&D passenger demand for city-pairs that are between 186 and 400 miles apart for years between 2008 and 2025. Actual passenger demand levels up to 2018.

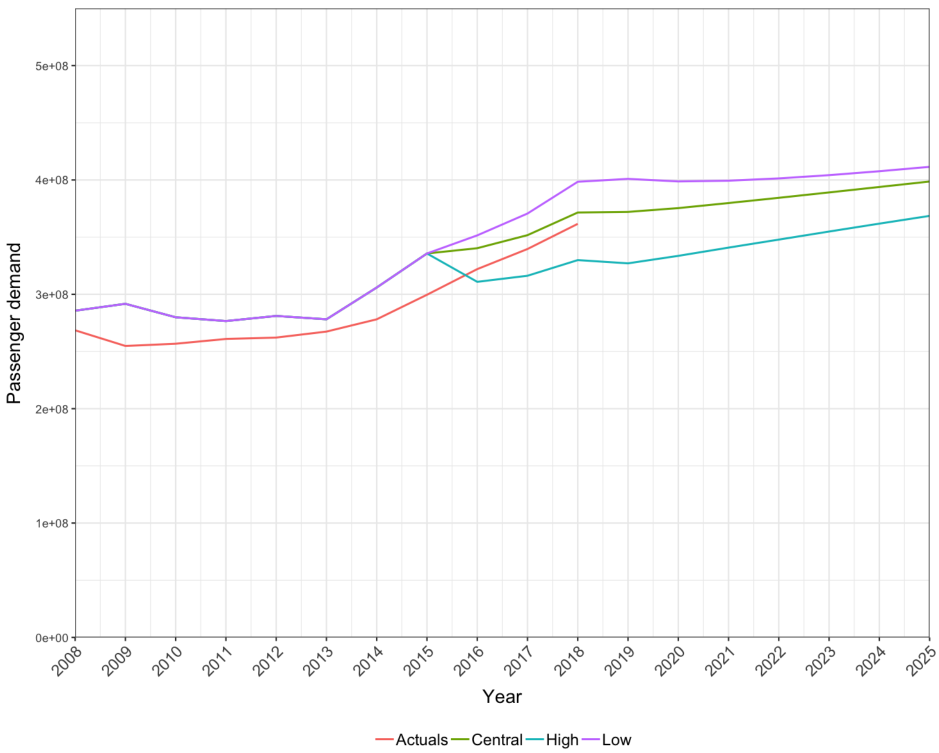

Figure 4.7: Projected O&D passenger demand for city-pairs that are between 400 and 2113 miles apart for years between 2008 and 2025. Actual passenger demand levels up to 2018.

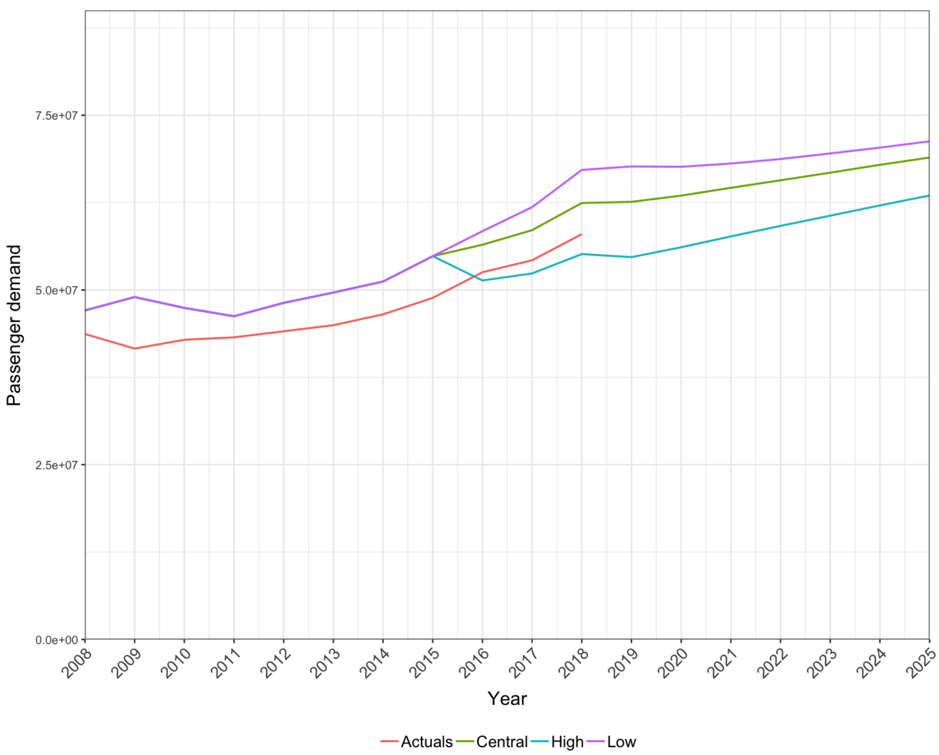

Figure 4.8: Projected O&D passenger demand for city-pairs that are more than 2113 miles apart years between 2008 and 2025. Actual passenger demand levels up to 2018.

As mentioned earlier, projections obtained from the three models presented above are aggregated to obtain the projected O&D passenger demand of the total US network system. Figure 4.9 shows the total projected passenger demand for the US domestic network for the three different scenarios considered based on airfares projections. Figure 4.10 shows the total observed passenger demand for the US domestic network against the projected passenger demand for years 2008, 2010, 2012, 2014, 2016 and 2018 and Table 4.4 shows the percentage error between projected and observed passenger demand up to 2018.

As seen with the projections obtained for each of the O&D demand passenger model by distance group, 2009 predictions tend to be much higher than those observed. In general models slightly over predict passenger demand, because given the way the model is formulated demand cannot go below 0, i.e. for a city-pair with annual demand of 1,000 can only under-predict by 1,000, but can over-predict by an unlimited amount leading to aggregate over-predictions for low-demand routes. However, results shown in Figure 4.9 show that predicted trend is similar to the one observed with exception of 2009. Results also show the models being sensible to changes in economy and airfares. For example, one could notice the slight peak of passenger demand in 2018, which could be the result of a combination between a peak in mean household income projections (Figure 4.2) in 2018 along with quite flat airfares projections for that same year (Figure 4.5).

Projections obtained imply an average yearly growth of passenger demand of 2% throughout the entire period (2008-2025); which are in line with recent forecasts published by the FAA (FAA, 2019), by which domestic passenger growth in the next 20 years is expected to average 1.8% per year. However, FAA's 2007 forecast projected a yearly average growth of 3.8% (FAA, 2007) for the years between 2008 and 2020[3].

[3]Note that the comparison might not be like for like, since FAA forecasts includes a larger number of airports as well as regional services.

Year	Low	Central	High
2008		6.8%	
2009		14.0%	
2010		8.4%	
2011		5.6%	
2012		7.3%	
2013		4.8%	
2014		9.43%	
2015		11.4%	
2016	9.3%	5.7%	-3.6%
2017	10.0%	4.3%	-6.4%
2018	11.9%	4.3%	-7.6%

Table 4.4: Difference between projected and observed total passenger demand for the US domestic network for years 2008-2018.

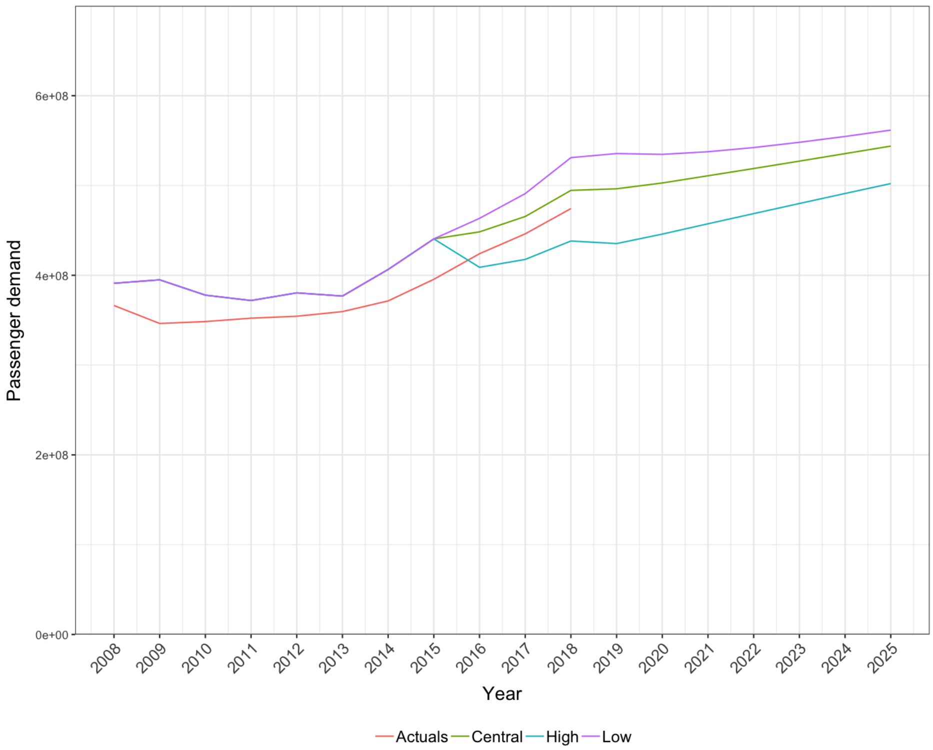

Figure 4.9: Projected O&D passenger demand for the US domestic network. Actual values up to 2018.

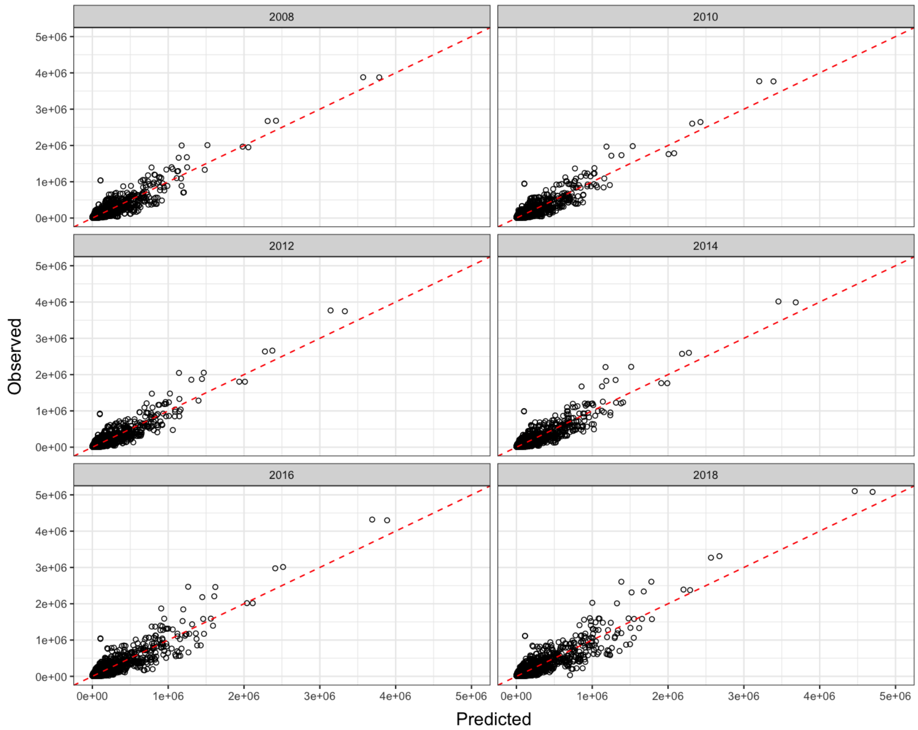

Figure 4.10: Observed against predicted O&D passenger demand for the US domestic network for years 2008, 2010, 2012, 2014, 2016 and 2018.

As stated earlier, for this example application of the modelling framework, the impact of network evolution, such as changes on average flight times (i.e. if a direct link is added in a market where only one-stop itineraries were available, flight time will most likely decrease) to the overall passenger demand is not considered and therefore this module is the only one that does not receive feedback from the other modules of the modelling framework. Considering the amount of uncertainty on how other factors will also change (e.g. airfares) and that flight time is introduced in the model through generalised cost, it is assumed the impact might not be significant (i.e. a new shorter route might have a lower flight time but might imply a greater airfare, which would balance the changes).

For example, the market Chicago-New York sees its average flight time being reduced by about 13 minutes in 2008 (i.e. compared to the actual value used) when considering the results obtained from the network evolution model for that year (due to changes in the connectivity of the available airport-pairs serving that market). This reduction of average flight time would imply passenger demand for this market to increase by 5% with respect to the currently predicted levels if and only if the rest of variables are kept the same. However, as mentioned earlier, the uncertainty around how the other factors affecting the overall demand, such as airfares or delay, would need to also be taken into account; as well as the fact that flight times for previously non-existing links is calculated by computing the product between average speed and market distance, which might be a too simple assumption. In this case example the reduction of flight time by 13 minutes is due to the new links added being associated with smaller airports (e.g. SWF-ORD), which are expected to have lower delay than those existing links between large airports (e.g. JFK-ORD); and therefore the overall passenger demand would be expected to not grow as much.

In this study, only scenarios regarding fares have been considered. However, it is known that also income would have an impact on passenger demand. Total number of passengers projected by 2025 for the US ATS is 543.8 million of passengers for the central scenario. This is equivalent to an increase of 39.1% with respect to 2008. Projected mean household income per capita used as input variable shows that mean household income in 2025 is 1.32 times that of 2008. On the one hand, if mean household income in 2025 would have been 1.2 times that of 2008, total O&D passenger demand would have been projected to be 466.6 million, considering the rest of variables unchanged. This would imply a decrease of 14% with respect to the initial projection and an increment of passenger demand with respect to 2008 of 19.3%. On the other hand, if mean household income in 2025 would have been 1.4 times that of 2008, total O&D passenger demand would have been projected to be 596,261,692 considering the rest of variables unchanged. This would imply an increase of 9.6% with respect to the initial projection and an increment of passenger demand with respect to 2008 of 52.5%. This calculations have been computed considering the linear relationship between mean household income and passenger, which is specific to the 3 distance group considered (i.e. short, medium and long)[4].

4.2 Airport connectivity model

The airport connectivity model projects evolution of US ATS network by projecting which airport-pairs (i.e. links) will be added to the network year-on-year and which will be removed from it. The problem is split between two models: one for link addition and one for link removal. The link addition model predicts airport-pairs

[4]Percentage of O&D demand by distance group is found to be as follows: short-haul holds 9.1%; medium-haul holds 78.3%; and long-haul holds 12.6%.

that are currently disconnected (i.e. there are not flights operating between airports i and j) and that will be added into the network the following year. The link removal model predicts airport-pairs that are currently connected (i.e. there are flights operating between airports i and j) and that will be remove from the network the following year (i.e. the operations will cease).

Both models use logistic regression as a model specification but they differ on the set of input variables considered for each. While the link addition model considers eigenvector centrality of the origin and destination airport, distance between the pair of airports and information regarding whether the origin and destination airport is a hub or not; the link removal model uses only a combined clustering coefficient representative of the airport-pair being under study.

Projections for input variables are not necessary at this stage since hub information and distance between airport-pairs are considered to remain the same throughout the projections years. Eigenvector centrality and clustering coefficient are computed from actual data for the base year (2007) and then computed from the projected connectivity from 2008 onwards since those two metrics only depend on airport degree - i.e. total sum of airports that an airport i is connected to.

To generate predictions of airport connectivity changes within the US ATS all possible airport-pairs that can be formed considering the 337 airports used in this study are split based on their connectivity status in the base year (2007). The two groups are: airport-pairs that are connected in 2007; and airport-pair that are not connected in 2007.

Based on the analysis of the evolution of the US air transportation system presented in Section 3.6, two thresholds in terms of number of links added and removed to/from the system are considered. Both thresholds are set to 87 up to 2013, with the threshold for link addition to increase to 108 from 2014 onwards and the threshold for link removal to decrease to 40 from 2014 afterwards. This decision is taken based on the average number of links added and removed from 2014 to 2018 (Table 4.5) which clearly shows a more differentiated pattern than those in the previous set of years (2007-2013) and presented in Table 3.7, which might be reflecting the start of the recovery from the economic recession that occurred in 2009.

Year	connected airport-pairs	Connected -> Connected	Connected -> Unconnected	Unconnected airport-pairs	Unconnected -> Unconnected	Unconnected -> Connected
2014	1,946	1,852	57	17,303	17,246	94
2015	1,942	1,863	83	17,307	17,224	79
2016	2,013	1,906	36	17,236	17,200	107
2017	2,074	1,970	43	17,175	17,132	104
2018	2,199	2,038	36	17,050	17,014	161
Mean		19,14	55	17,228	17,171	108

Table 4.5: Airport-pair connectivity between 2014 and 2018. An airport-pair is considered connected if there are at least 52 flights operating between them.

Similarly, the other aspect to consider when studying the capacity network evolution is the saturation within a given market. Based on Figures 3.19 and 3.20 (Chapter 3), which show the number of links that O&D markets had the previous year a link was added and removed respectively, one could agree that most of the capacity network evolution occurs within markets with no-direct or one link for link addition and within markets which one or two links for link removal. Based on this, the threshold of saturated market for link addition is set up to 2 and to 3 for link removal after 2013.

Considering the above assumptions, projections for airport-pair connectivity changes are generated from 2008 up to 2025. Table 4.6 presents the confusion matrix, which is used to visually compared observed number of airport-pair connected and disconnected against predicted number of airport-pair connected and disconnected, for years 2008, 2010, 2012, 2014, 2016 and 2018. Values presented in Table 4.6 are the cumulative values of the results obtained after applying the link addition and removal models. Although errors might seem low - i.e. number of predicted airport-pairs being connected which are actually not connected, and number of predicted unconnected airport-pairs which are actually connected -, number of allowed connectivity changes was limited, and therefore the error increase that can be seen through the years is a clear sign of low model precision as presented in Section 3.6. This is the result of error propagation, since from 2009 onwards input data for CC and EVC, which are part of the explanatory variables set, are being computed from projected values which as shown in Table 4.6 have low precision.

Similarly as done in Section 3.6 a further analysis is performed to check whether the results obtained by combining the link addition and removal model results allow

		Predicted	
		Connected	Unconnected
		Year = 2008	
Observed	Connected	14119	102
	Unconnected	139	1905
		Year = 2010	
Observed	Connected	13239	195
	Unconnected	349	1622
		Year = 2012	
Observed	Connected	12794	204
	Unconnected	417	1632
		Year = 2014	
Observed	Connected	12414	227
	Unconnected	486	1630
		Year = 2016	
Observed	Connected	12009	268
	Unconnected	563	1638
		Year = 2018	
Observed	Connected	11848	365
	Unconnected	613	1681

Table 4.6: Observed number of airport-pair connected and disconnected against Predicted number of airport-pair connencted and disconnected for years 2008, 2010, 2012, 2014, 2016 and 2018.

the prediction of the evolution trend of the entire network. In order to do that the projected network degree, network eigenvector centrality and network average path length[5] evolution are computed, with values for years between 2008 and 2018 being compared to their actual values (Figures 4.11, 4.12 and 4.13).

Projections for network degree, eigenvector and average shortest path (Figures 4.11, 4.12 and 4.13) are quite close to actuals up to 2015, where projections seem to diverge towards smaller values for all three metrics. Considering external factors, in 2015 there was a quite steep drop of fossil fuel prices, which in turn affected also airfares. This suggests that connectivity changes within the US network might be influenced by changes in fuel prices, making the study of how fossil fuel price changes

[5]The average path length in a system is calculated based on the shortest paths between all pairs of airports.

affect network evolution a potential line of research for future work.

Results obtained for the system-wide network theory metrics are slightly higher than the ones presented by Kotegawa's work (2012) for network degree (e.g. network degree was found to be about 17.1 in 2009 as opposed as around 20 in this research); but quite similar in terms of network eigenvector centrality (i.e. 0.28) and average shortest path (i.e. 2.29). The observed increment on average network degree would be a sign that the new links added to the network are associated with already well-connected airports, and therefore favouring a hub-and-spoke strategy. In contrast, the link addition and removal model projections seem to balance out by keeping a similar network structure across all the projections years, suggesting that other external factors, such as the mentioned changes in fuel prices or government cuts might also be influencing how connectivity changes occur.

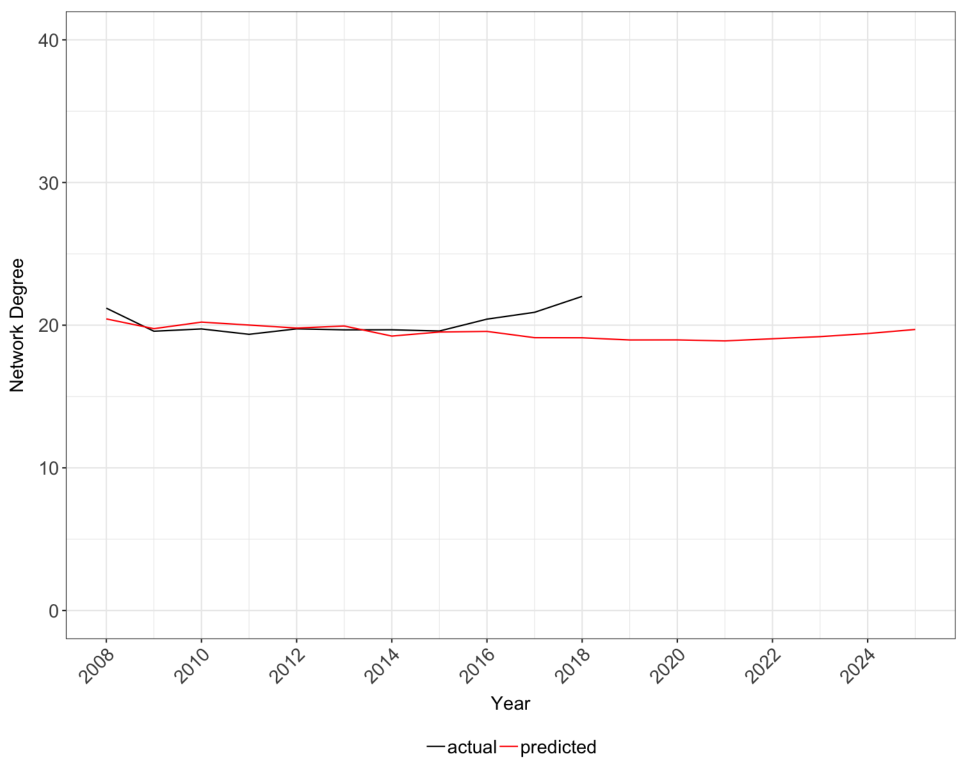

Figure 4.11: Projections for the US Air Transportation average network degree between 2008 and 2025. Actuals for years between 2008 to 2018.

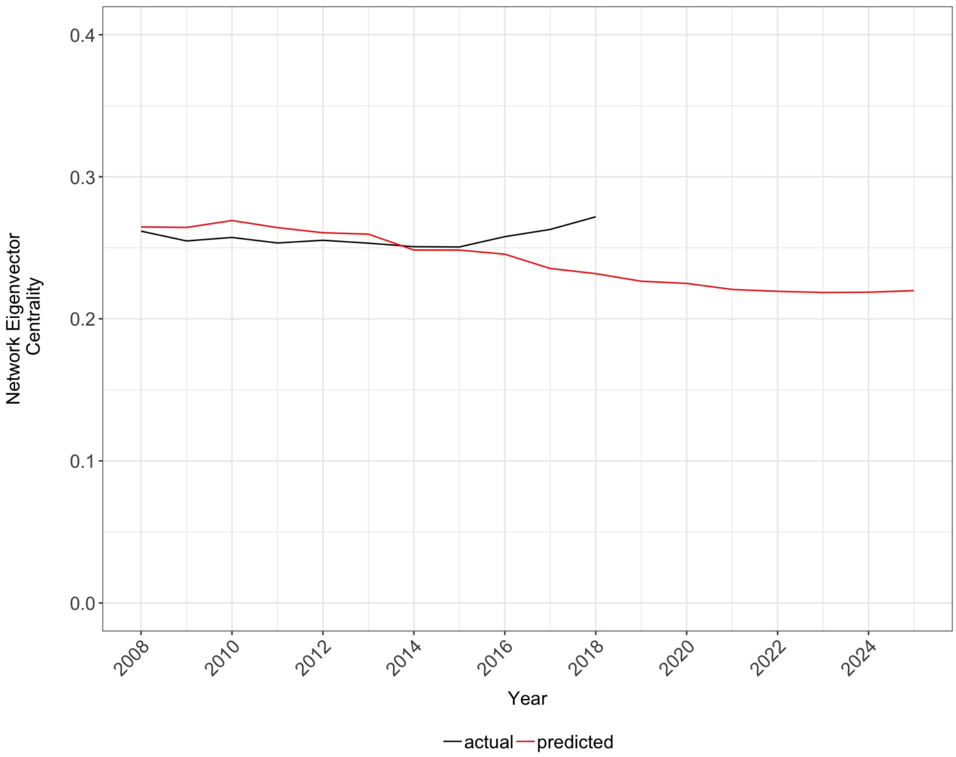

Figure 4.12: Projections for the US Air Transportation network average eigenvector centrality between 2008 and 2025. Actuals for years between 2008 to 2018.

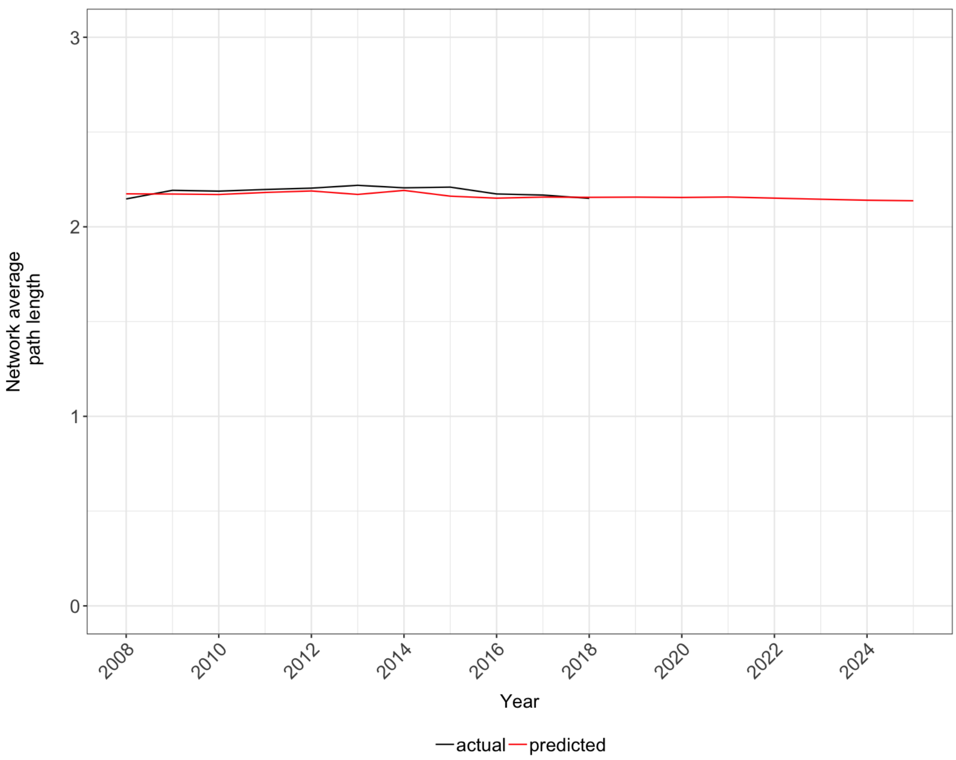

Figure 4.13: Projections for the US Air Transportation network average path length between 2008 and 2025. Actuals for years between 2008 to 2018.

4.3 Case 1: Evolution of the US network not considered

In this case results obtained from the airport-pair connectivity model presented in the above section are not considered and therefore itinerary availability will be considered to remained unchanged from 2018 onwards. This case involves generating projections from the itinerary choice model using both approaches presented in Chapter 3 (i.e. multinomial logit and neural network models) and to produce air traffic level projections up to 2025.

In order to generate projections for itineraries serving any O&D, two methodologies are used and presented in the following sub-sections: itinerary choice model using a multinomial logit through a 2-SCF methodology (Sections 4.3.1); and neural networks (4.3.2); both sets of projections will be compared and discussed in Section 4.3.3. The results of those two sets of projections will be used as one of the input variables for the third stage of the modelling framework which aims to project air traffic levels. In this case (i.e. only when evolution is not considered), two sets of results are also generated when projecting air traffic levels based on the model used: the 2-SLS model with auto-regressive term; and the OLS model without auto-regressive term; both presented in Sections 4.3.4 and 4.3.5 respectively.

4.3.1 Case 1 - Itinerary choice model: multinomial logit model

As input set, each O&D for which passenger demand has been projected during the first stage will be considered. For each of these O&D a maximum of 9 itineraries is compiled up to 2018, with itineraries afterwards considered to remain the same as 2018.

Similarly to the O&D passenger demand stage, 3 different scenarios have been estimated: low, medium, high; which are defined by itinerary fares projections. The same simple relationship between airfares and oil fossil fuel price changes that has been considered for the O&D demand module has also been considered for this model. Note that scenario results from the O&D passenger demand module are matched to the same scenario at this stage (e.g. market passenger demand from low scenario will be used in the low scenario when computing itinerary shares). Figure 4.14 shows the average itinerary fare throughout projections years (i.e. 2008-2025) for the 3 scenarios.

To generate projections at this stage, itinerary fares are first projected by using Equation 3.17. Then, the model estimation results from the second stage of the itinerary choice model are applied as the set of coefficients for equation 3.16, so that the utility and the probability of choosing each itinerary can be calculated. Input dataset for this second stage will include the estimated itinerary fares from stage 1 as well as the residual value.

To evaluate the results a system-wide metric is computed: Revenue Passenger Kilometre (RPK). RPK is the basic measure of passenger traffic and reflects how many available seats were actually sold. RPK is defined as the product between number of passengers and kilometres they have flown. Figure 4.15 shows the comparison and projections for RPK, whereas Table 4.7 presents RPK for years 2008, 2018 and 2025.

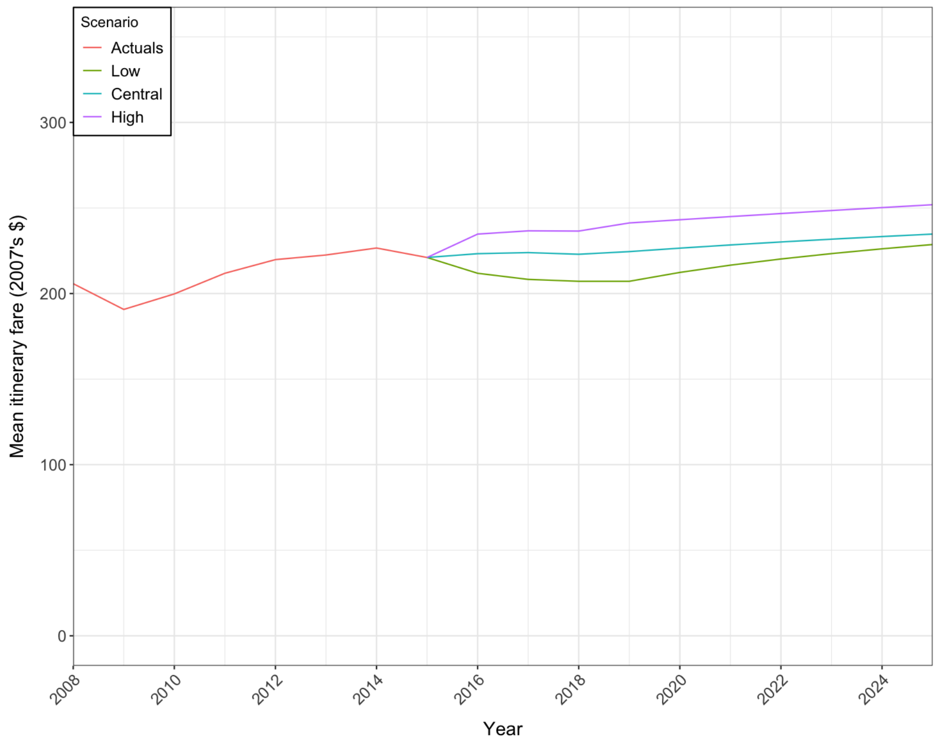

Figure 4.14: Itinerary fares actuals and projections.

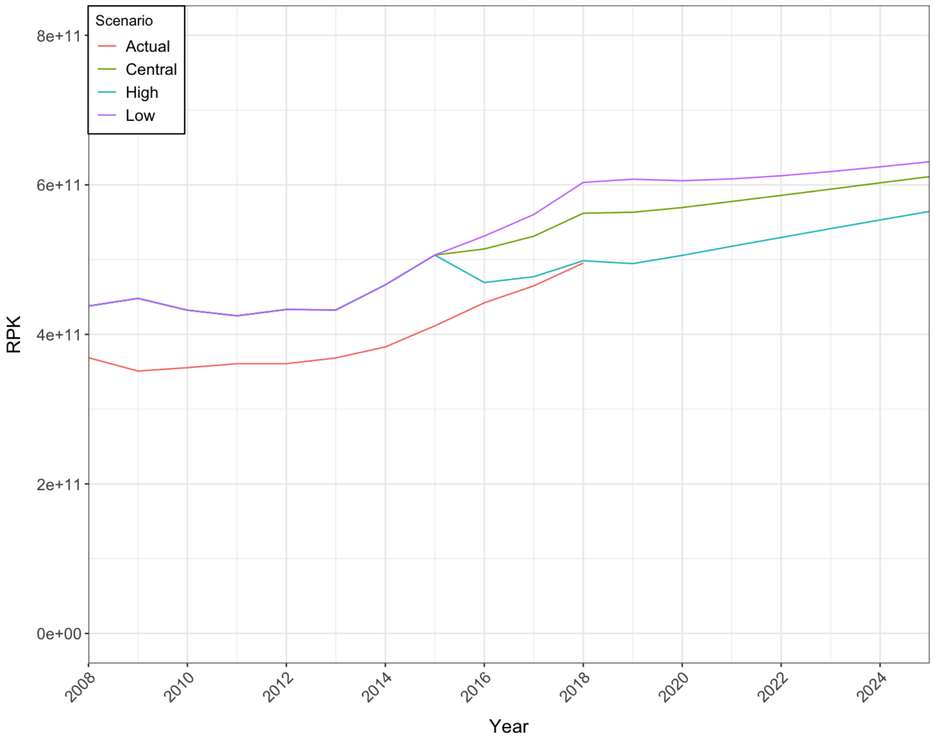

Figure 4.15: RPK actuals and projection values.

Scenario	2008	2018	2025	Change 2025 vs 2008
Actual	368,828	495,347		
Low	437,908	603,251	630,936	44.1%
Central	437,908	562,039	611,005	39.5%
High	437,908	498,498	564,428	28.9%

Table 4.7: RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using multinomial logit.

4.3.2 Case 1 - Itinerary choice model: neural network

When considering the itinerary choice model built using neural network, the input dataset is the same as the one used for the itinerary choice model estimated using multinomial logit model. To evaluate results the same metric is used, i.e. RPK. Figure 4.16 shows the comparison and projections for RPK, whereas Table 4.8 presents RPK for years 2008, 2018 and 2025.

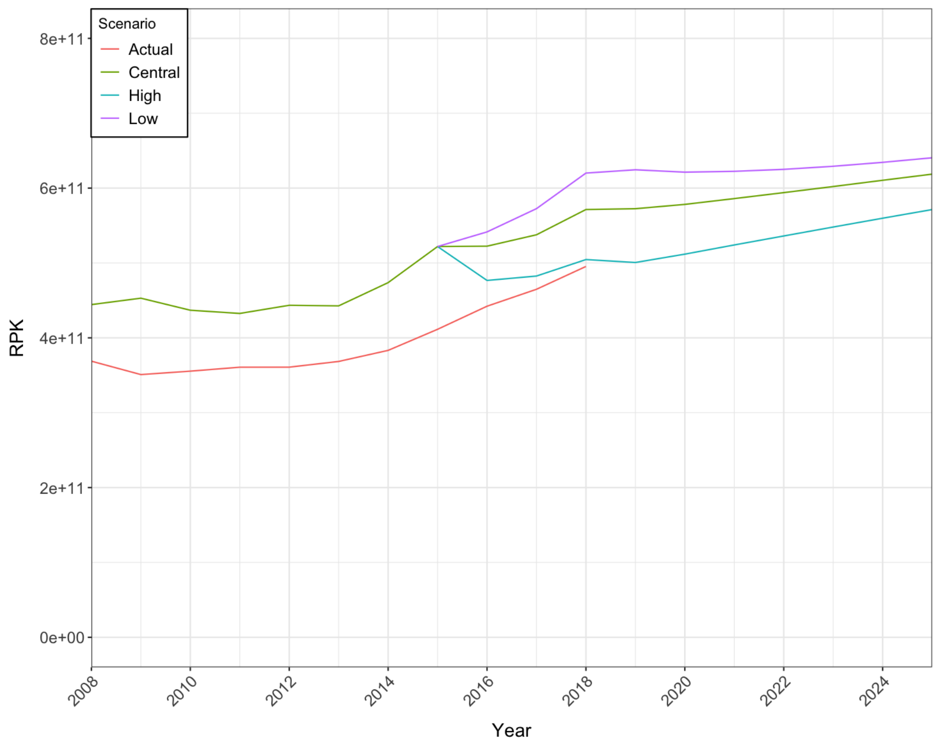

Figure 4.16: RPK actuals and projection values.

Scenario	2008	2018	2025	Change 2025 vs 2008
Actual	368,828	495,347		
Low	444,394	620,495	630,936	44.1%
Central	444,394	571,390	618,651	39.2%
High	444,394	504,563	571,324	28.6%

Table 4.8: RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using neural networks.

4.3.3 Case 1 - Itinerary choice model: comparison

To evaluate both models together, adjusted R^2 is computed for years in which actual data exist (i.e 2008-2018) for both sets of projections. Also, Figure 4.17 presents the comparison between RPK projected up to 2025 for the two models and actual RPK up to 2018. Overall, it is clear that results from both models are higher than those actuals. Two possible reasons could explain the over-prediction of RPK by the two models. First, itinerary shares are predicted based on city-pair passenger demand predicted during the first stage of the modelling framework and as presented in Section 4.1, models at that stage slightly over-predicted passenger demand, prompting error propagation through the itinerary share models. Secondly, a further check is done by computing the proportion of non-stop passengers that each of the models project (Figure 4.18), and it is compared to actuals' non-stop passenger proportions since a model predicting a larger number of one-stop passengers would tend to increase the system-wide RPK. From Figure 4.18 it is clear that both itinerary choice models tend to predict a greater proportion of one-stop passengers, especially the neural network model which shows a much more sharp fluctuation. This is likely because the demand over-predictions is from routes with small numbers of passengers (as these routes can be wrong upwards by more than they can be wrong downwards in absolute terms). These routes are more likely to be the one-stop ones. Therefore the RPK over-prediction is likely much stronger for one-stop routes. Also, it is worth to notice that neural network model projections seem to be more affected by airfares fluctuation, with a clear drop of non-stop passengers numbers in 2015 and a clear difference amongst the three scenarios, suggesting that other factors not considered in the model might be influencing the predictive power of this model. Overall, results for the multinomial logit model show a similar trend to those observed and a smoother curve; and adjusted R^2 values above 0.66 for the period 2008-2018, which considering the error propagation from the O&D demand model is relatively high.

	Year										
Model	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
MNL	0.707	0.705	0.731	0.726	0.712	0.713	0.711	0.687	0.676	0.663	0.661
NN	0.578	0.559	0.564	0.589	0.626	0.600	0.608	0.463	0.586	0.583	0.566

Table 4.9: Adjusted R^2 computed from the projections generated from both models up to 2018. Values referred to the Central scenario.

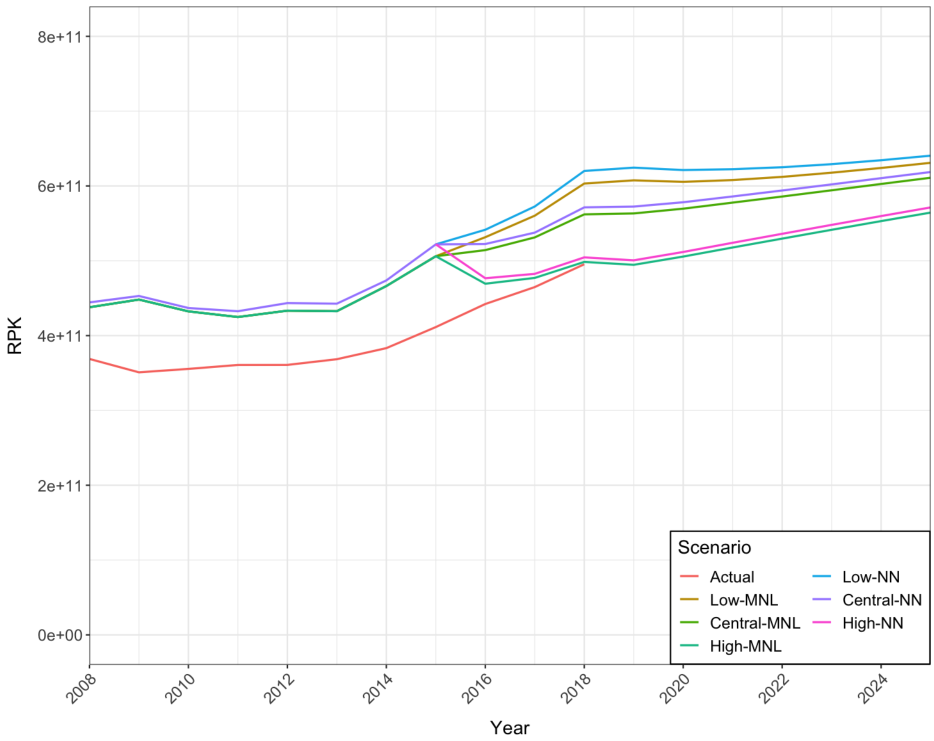

Figure 4.17: RPK actuals and projection values for both models: multinomial logit and NN.

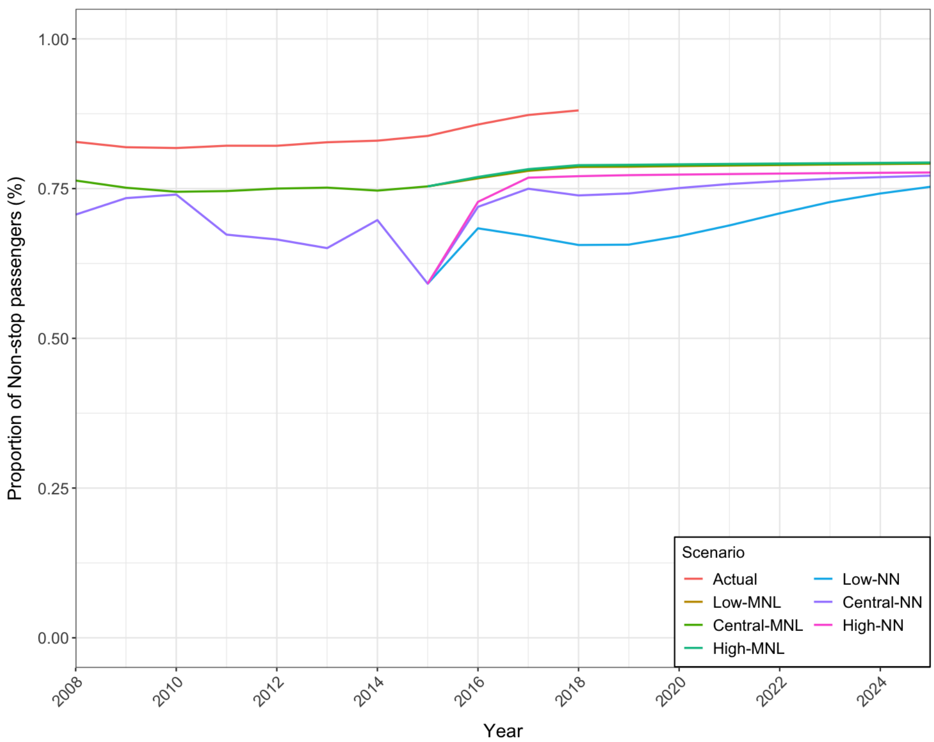

Figure 4.18: Comparison of percentage proportion of non-stop passengers obtained from the MNL and NN models up to 2025. And actual proportion of non-stop passengers up to 2018.

4.3.4 Case 1 - Air traffic levels model: Model 2-OLS

The third stage of the modelling framework projects air traffic levels per airport-pair (i.e. segment). In this case, the projected passenger demand by itinerary obtained during the previous stage is used as one of the explanatory variables of the model. In this sub-section projections generated using the Model 2-OLS without auto-regressive term are presented.

For projections, 3 different scenarios have been considered based on the passenger demand projections obtained for each scenario (i.e. low, medium, high). Load factor (LF) values are considered to be maintained as 2018's levels afterwards. Also, two sets of results have been computed: one considering the results obtained from the itinerary choice model using multinomial logit; and another one for the neural network results. Figures 4.19 and 4.20 show the comparison between predicted and actual values for years 2008, 2010, 2012, 2014, 2016 and 2018 for the set of MNL results and for the NN results respectively. Figure 4.21 shows traffic level projections up to 2025 for the two sets of results (i.e. MNL and NN) as well as actuals up to 2018. Table 4.10 presents the traffic levels from 2008, 2018 and 2025 for the different scenarios. As expected, results obtained from the application of the air traffic model to data obtained when the MNL model is used are closer to those observed air traffic levels; trends for this set of results are also smoother due to smoother itinerary share results (Figure 4.17).

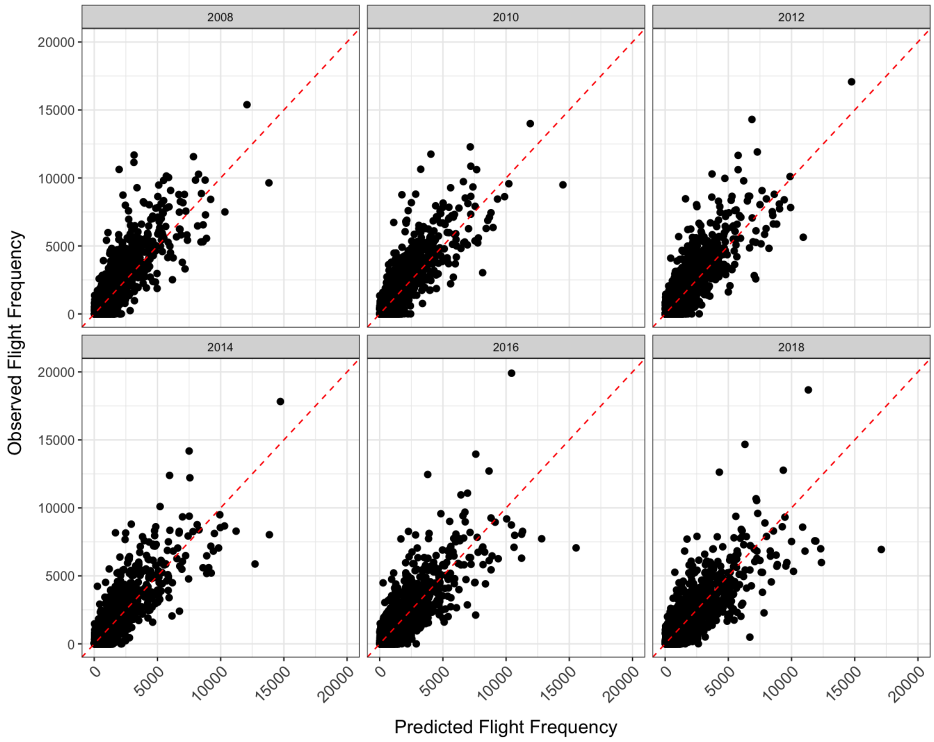

Figure 4.19: Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the MNL model. Model used: Model 2-OLS.

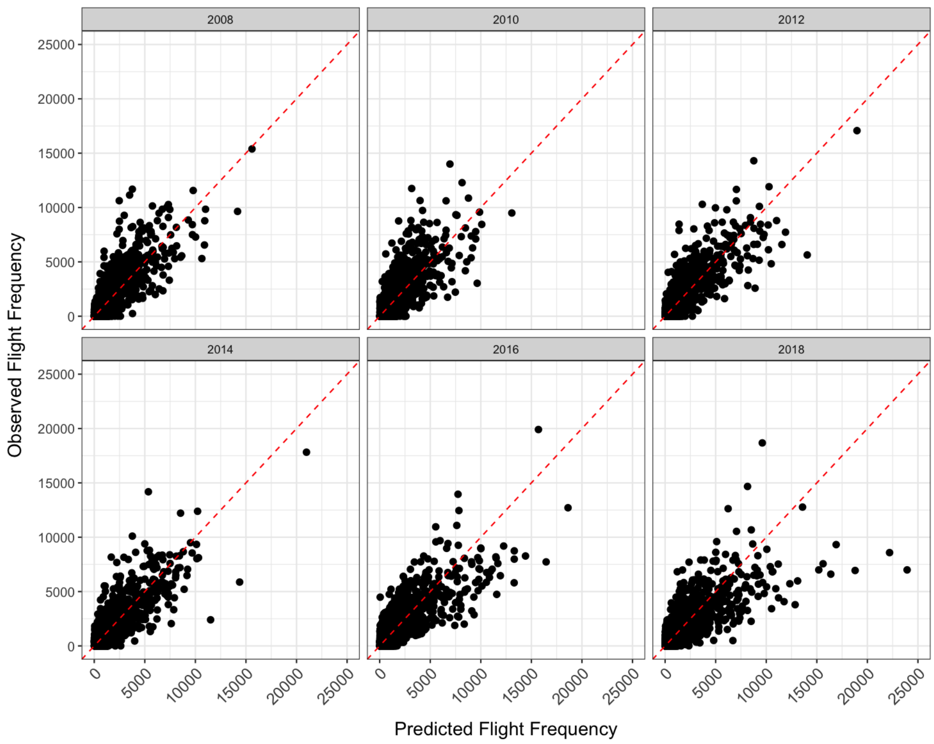

Figure 4.20: Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the NN model. Model used: Model 2-OLS.

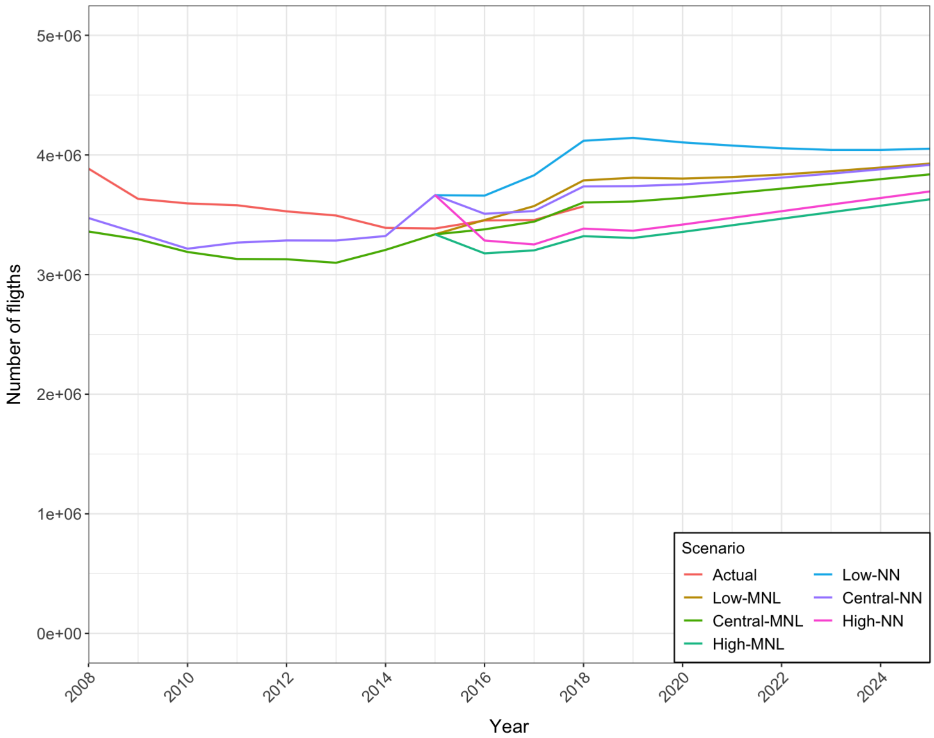

Figure 4.21: Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN. Actual traffic levels up to 2018. Model used: Model 2-OLS.

Model	Scenario	2008	2018	2025	Change 2025 vs 2008
Actual		3,885,429	3,570,238		
MNL	Low	3,359,747	3,787,203	3,927,787	16.9%
	Central	3,359,747	3,602,729	3,837,636	14.2%
	High	3,359,747	3,320,844	3,628,929	8%
NN	Low	3,472,379	4,118,369	4,051,668	16.7%
	Central	3,472,379	3,736,944	3,916,614	12.8%
	High	3,472,379	3,383,964	3,694,722	6.4%

Table 4.10: Air traffic levels projections for 2008, 2018 and 2025 for the two sets of results and scenarios. Actuals for 2008 and 2018.

4.3.5 Case 1 - Air traffic levels model: Model 1-2SLS

In this sub-section projections generated using the Model 1-2SLS which includes the auto-regressive term amongst the explanatory variable set are presented. To evaluate projections generated by this model, the total number of flights in the US ATS is plotted and compared against actuals for years up to 2018. Figures 4.22 and 4.23 show the comparison between predicted and actual values for years 2008, 2010, 2012, 2014, 2016 and 2018 for the set of MNL results and for the NN results respectively. Figure 4.24 shows traffic level projections up to 2025 for the two sets of results (i.e. MNL and NN) as well as actuals up to 2018. Table 4.11 presents the traffic levels for 2008, 2018 and 2025 for the different scenarios. From the results obtained few aspects can be highlighted: both sets of results generate less smooth trends than observed, with MNL set of results being slightly closer to those observed air traffic level; as expected NN set of results show a higher number of flights at the system-level since those results predicted higher proportion of one-stop passengers, this difference being much higher for years when the difference in non-stop passengers was much greater (i.e. 2015 as presented in Figure 4.18); and projected values for years 2018 onwards from both models do not show any clear difference across the different scenarios suggesting that this model has a high dependency on previous year's number of flights[6] omitting any other factor that could impact air traffic levels, and therefore making this model's applicability somehow limited. To illustrate the latter point better, Figure 4.25 shows the comparison between the projected total number of flights when the Model 2-OLS and the Model 1-2SLS are used, clearly showing the influence that other factors (e.g. such as airfares fluctuations) have towards air traffic levels for the former model.

[6]Note that for Case 1 results, the network is assumed to be static from 2018 onwards.

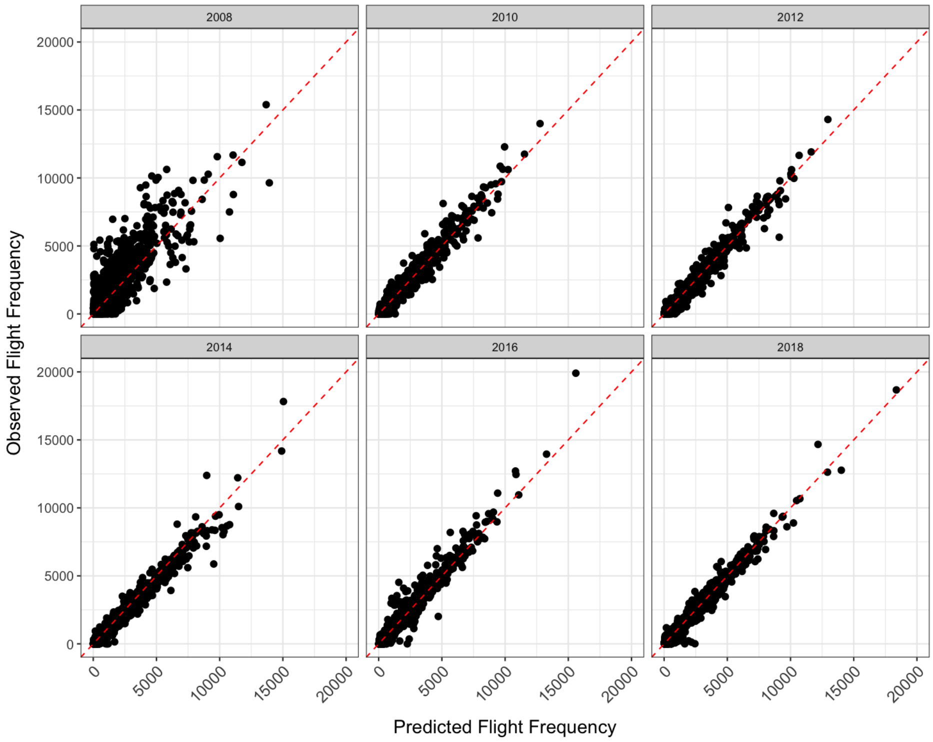

Figure 4.22: Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the MNL model. Model used: Model 1-2SLS with auto-regressive term.

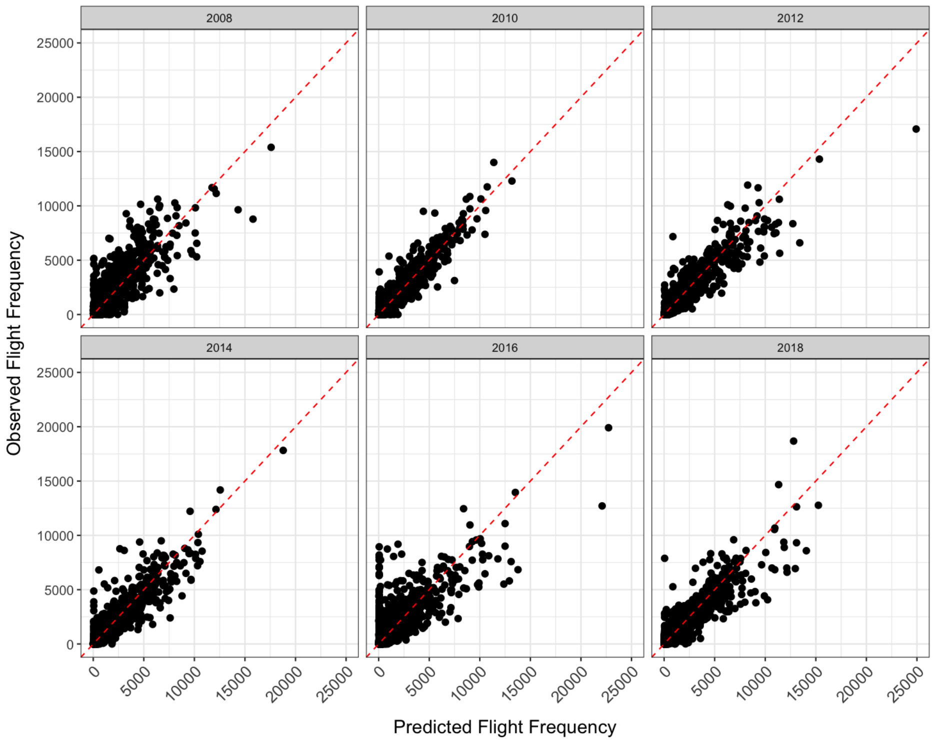

Figure 4.23: Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the NN model. Model used: Model 1-2SLS with auto-regressive term.

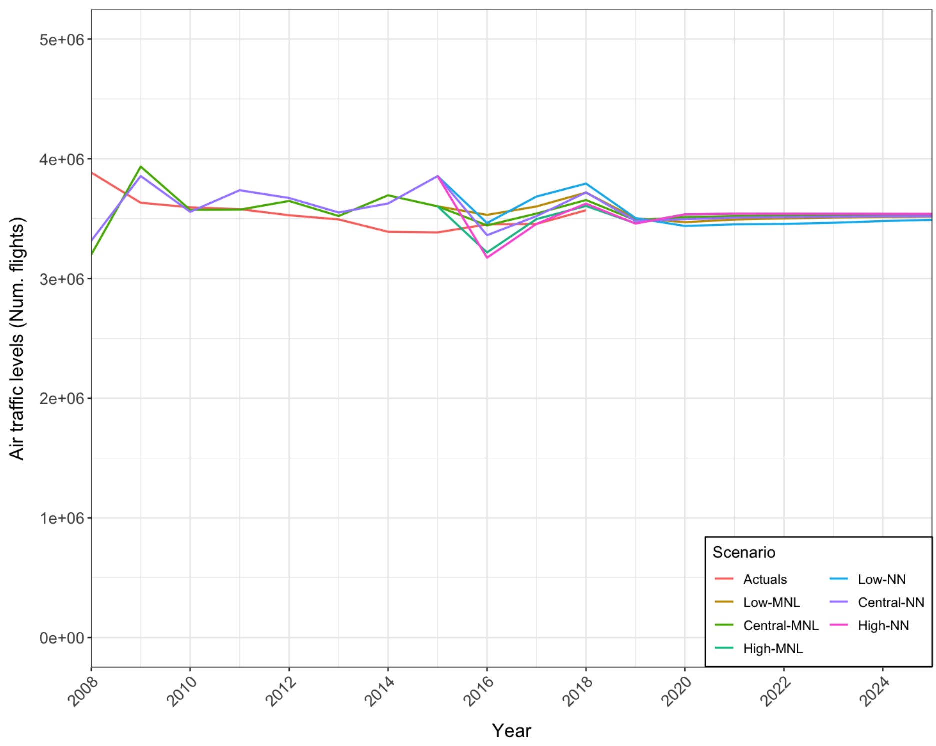

Figure 4.24: Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN; and using Model 1-2SLS with auto-regressive term. Actual traffic levels up to 2018.

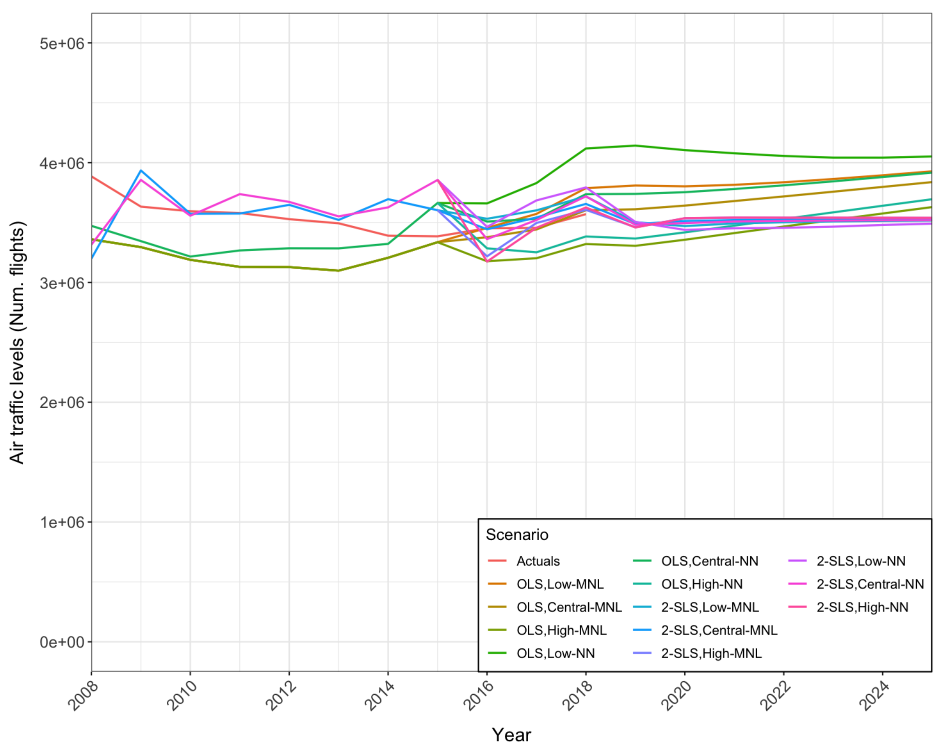

Figure 4.25: Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN; and using Model 1-2SLS with auto-regressive term. Actual traffic levels up to 2018.

Model	Scenario	2008	2018	2025	Change 2025 vs 2008
Actual		3,570,238	3,885,429		
MNL	Low	3,199,070	3,717,904	3,515,873	9.9%
	Central	3,199,070	3,655,594	3,523,311	10.1%
	High	3,199,070	3,605,952	3,538,469	10.6%
NN	Low	3,315,464	3,793,214	3,490,422	5.3%
	Central	3,315,464	3,719,219	3,520,000	6.2%
	High	3,315,464	3,625,661	3,539,673	6.8%

Table 4.11: Air traffic levels projections for 2008, 2018 and 2025 for the two sets of results and scenarios; when using Model 1-2SLS with auto-regressive term. Actuals for 2008 and 2018.

4.4 Case 2: Evolution of the US network considered

Another set of simulations are produced when considering the airport-pair connectivity results presented in Section 4.2. Similarly to when network evolution is not considered, two sets of results are generated based on the itinerary choice model used: multinomial logit or neural network. In this case, the Model 1-2SLS with auto-regressive term to predict air traffic levels is not used and only results obtained when using the Model 2-OLS are presented.

4.4.1 Case 2 - Itinerary choice model: multinomial logit model

When considering network evolution, to be able to generate projections of itinerary shares available itineraries need to be compiled since routes that were not possible may become available and routes that were once possible may become unavailable. In order to compile the different itineraries the following rules apply:

- A non-stop itinerary between city *o* and *d* would be possible if any airport-pair serving those two cities has been predicted as connected;
- A one-stop itinerary between city *o* and *d* would be possible only if both legs of a one-stop itinerary serving that city-pair (i.e. O&D) have been predicted as connected;

- For each O&D, new possible itineraries will be checked against existing ones. Only if a new itinerary has a shorter flight time than the maximum flight time of those existing itineraries, this will be added.
- For each O&D a maximum of 2 new itineraries can be added each year and the maximum number of available itineraries per O&D considered is 10. The most restrictive rule will apply.
- In the case when more than 2 new itineraries can be added, those would be ordered by journey time and only those top 2 will be added.

Journey times for those new itineraries are calculated by multiplying average speed with the market distance of the new itinerary. Average speeds are computed by hub airport used and group distance[7]. Itinerary fares for those new itineraries are computed in a similar way -i.e. product of average fare per mile times market distance-; and average fare per mile is also computed based on hub airport used and group distance.

To evaluate the results the same metric used when network evolution is not considered (Section 4.3.3) is computed (i.e. RPK). Projections are generated for the 3 scenarios considered: low, central and high. Figure 4.26 presents RPK projected values for the MNL set of results with Table 4.12 presenting the RPK values (in millions) for years 2008, 2018 and 2025. Results obtained (Figure 4.26) when applying the multinomial logit model show a similar trend to those obtained when the network evolution was not considered, suggesting that such model is not largely affected but network structure changes and tends to predict more towards non-stop passengers. A slight difference is noticeable in the fact that projected RPKs growth (change between 2008 and 2025) - i.e. from Table 4.12: 43.3% for the low scenario, 38.9% for the central scenario and 28.3% for the high scenario - are slightly softer than those obtained when the model was considered static - i.e. from Table 4.7: 44.1% for the low scenario, 39.5% for the central scenario and 28.9% for the high scenario. Overall, differences to those RPK levels observed are slightly higher than when considering a static network, which can be explained by the fact that new itineraries added to the network tend to be one-stop itineraries and therefore a further share of passengers would travel through such itineraries.

[7] Group distances are: 1 - 0 to 500 miles; 2- 500 to 750 miles; 3- 750 to 1000 miles; 4- 1000 to 1250 miles; 5- 1250 to 1500 miles; 6- 1500 to 1750; 7- 1750 to 2000; 8- more than 2000 miles.

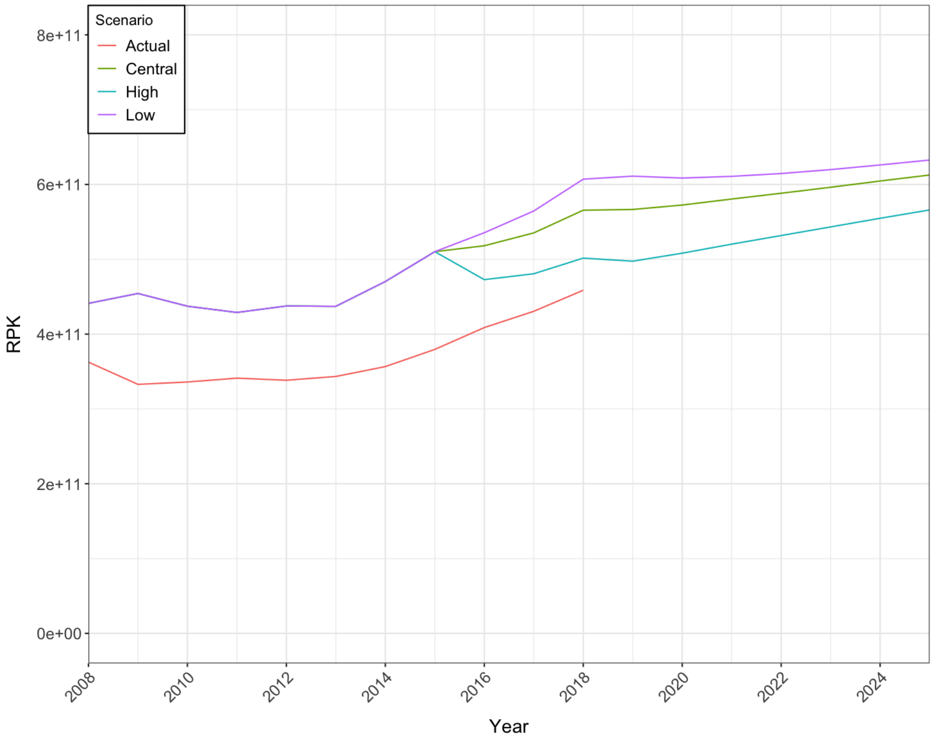

Figure 4.26: RPK actuals and projection values when applying the MNL model and considering evolution of the US ATS. Model used: MNL.

Scenario	2008	2018	2025	Change 2025 vs 2008
Actual	362,573	458,642		
Low	441,096	607,091	632,622	43.3%
Central	441,096	565,587	612,641	38.9%
High	441,096	501,595	565,913	28.3%

Table 4.12: RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using multinomial logit and considering evolution of the US ATS.

4.4.2 Case 2 - Itinerary choice model: neural network model

When considering the itinerary choice model built using neural network, the input dataset is the same as the one used for the itinerary choice model estimated using multinomial logit model and the same rules apply to compile the available itineraries. Also, to evaluate results the same metric is used, i.e. RPK. Figure 4.27 shows the comparison and projections for RPK, whereas Table 4.13 presents RPK for years 2008, 2018 and 2025. Similarly to MNL results, results obtained when applying the

NN model (Figure 4.27) present a similar trend to those obtained when network evolution was not considered. Also, RPK growth between 2008 and 2025 is softer when network evolution is considered - i.e. from Table 4.13: 42.7% for the low scenario, 38.1% for the central scenario and 27.5% for the high scenario -, than when it is not - i.e. from Table 4.8: 44.1% for the low scenario, 39.2% for the central scenario and 28.6% for the high scenario. Overall, the difference between predicted and observed RPK levels is larger when considering network evolution as well as when comparing it to the results obtained when using the MNL model.

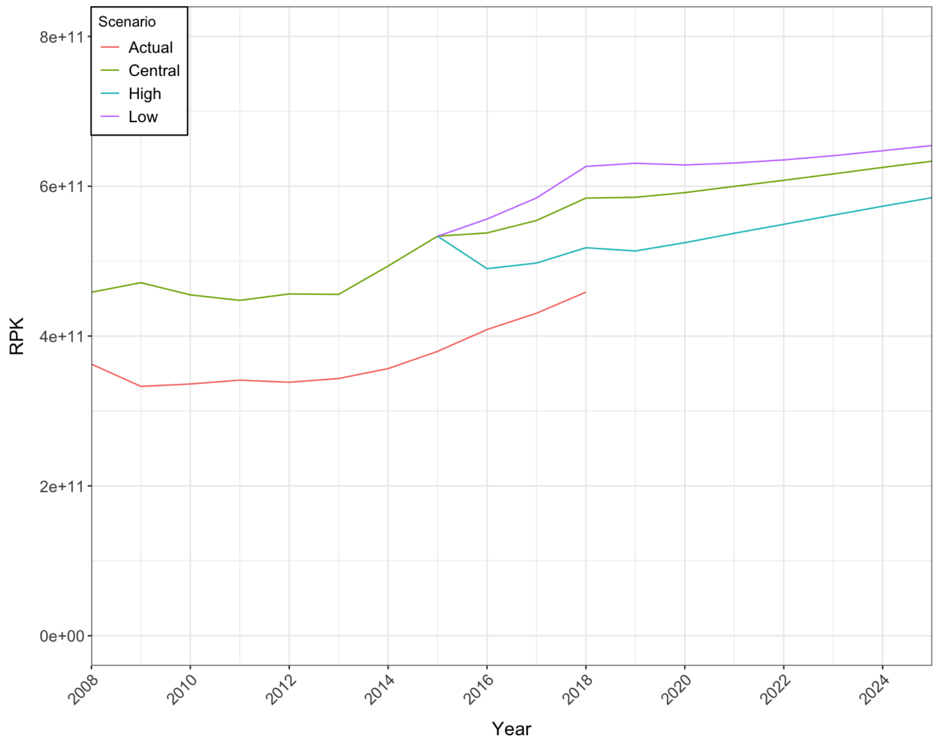

Figure 4.27: RPK actuals and projection values when applying the NN model and considering evolution of the US ATS.

Scenario	2008	2018	2025	Change 2025 vs 2008
Actual	362,573	458,642		
Low	458,469	626,452	654,277	42.7%
Central	458,469	584,171	633,439	38.1%
High	458,469	517,917	584,791	27.5%

Table 4.13: RPK actuals and projected (in millions) for years 2008, 2018 and 2025 for the itinerary choice model estimated using neural networks and considering evolution of the US ATS. Model used: NN.

4.4.3 Case 2 - Itinerary choice model: comparison

To evaluate both models together, adjusted R^2 is computed for years in which actual data exist (i.e 2008-2018) for both set of projections (Table 4.14). Also, Figure 4.28 presents the comparison between RPK projected up to 2025 for the two models and actual RPK up to 2018. A similar picture than the one obtained when considering a static network is obtained (Figure 4.17) when considering network evolution. NN results show higher values of RPK levels than those obtained through the MNL model; and overall predicted RPK tend to be on the over-prediction side, however, trends are smoother and quite in line with actuals up to 2018, specially for central and low scenario. Two main reasons are believed to be causing over-prediction: first, projected city-pair passenger demand obtained in stage 1 are slightly over-predicting prompting to error propagation; and both MNL and NN model tend to under predict the proportion of non-stop passengers in the network as shown in Figure 4.29, with differences to those actuals being larger than when the network was assumed to be static.

	Year										
Model	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
MNL	0.689	0.668	0.689	0.680	0.648	0.662	0.658	0.626	0.626	0.625	0.620
NN	0.417	0.419	0.406	0.365	0.415	0.415	0.346	0.329	0.485	0.516	0.506

Table 4.14: Adjusted R^2 computed from the projections generated from both models up to 2018 when considering network evolution (Case 2). Values referred to the Central scenario.

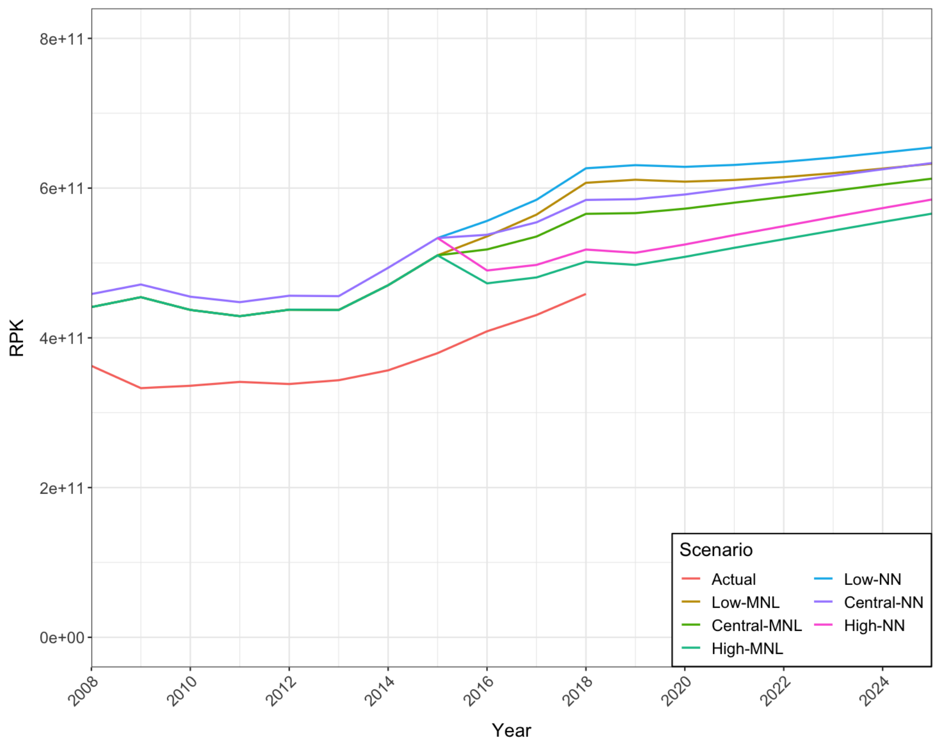

Figure 4.28: RPK actuals and projection values comparison for both models (MNL and NN) when considering evolution of the US ATS.

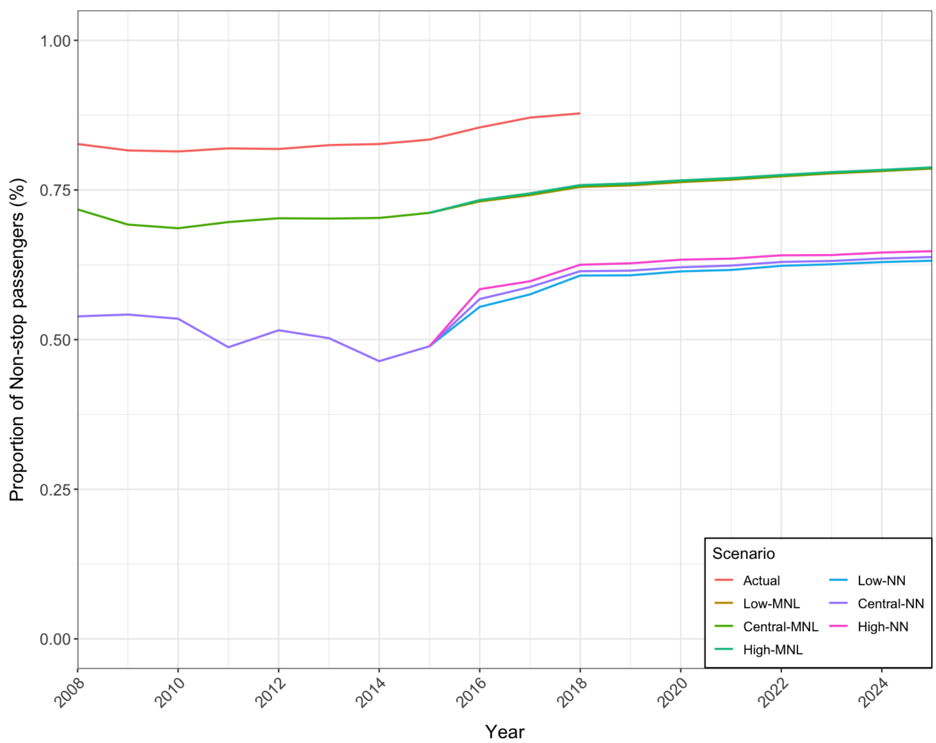

Figure 4.29: Comparison of percentage proportion of non-stop passengers obtained from the MNL and NN models up to 2025. And actual proportion of non-stop passengers up to 2018. Case when considering the evolution of the US ATS.

4.4.4 Case 2 - Air traffic levels: considering Airport-Pair connectivity

The results from the itinerary choice models are then used as input variables for the air traffic model. Two sets of results are produced: one regarding the MNL results; and one regarding the NN results. Note that projections are generated only using the Model 2-OLS. Amongst the set of explanatory variables, load factors have been considered to be constant from 2018 onwards; for new itineraries historical values have been assumed for the same itinerary; and for those new itineraries where historical information does not exist, load factors have been assumed to be equal to the average network load factor.

Figure 4.30 and 4.31 shows the comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018 when using segment passenger demand obtained from the MNL and NN model respectively. Whereas Figure 4.32 shows traffic level projections up to 2025 for the two sets of results (i.e. MNL and NN). Table 4.15 presents the traffic levels fro 2008, 2018 and 2025 for the different scenarios. Since MNL results are closer to those observed, better accuracy is expected to be obtained from applying the air traffic model to MNL results as seen when comparing Figures 4.30 and 4.31 specially in the long-term. Interesting to see, however, that total number of flights in the US ATS predicted when considering the set of NN results are quite close to those observed levels up to 2015, suggesting that the increase on one-stop passengers would tend to be mitigated by larger aircraft. Considering the results presented in Figure 4.32 a simple check on average number of passengers per aircraft can be done by simply dividing the the total number of passengers in the network at the segment level by the number of total predicted flights (Figure 4.33). Results show how the set of NN results assume a higher average number of passenger per aircraft which could suggest a shift towards the use of larger aircraft. In both set of results similar trends are however observed.

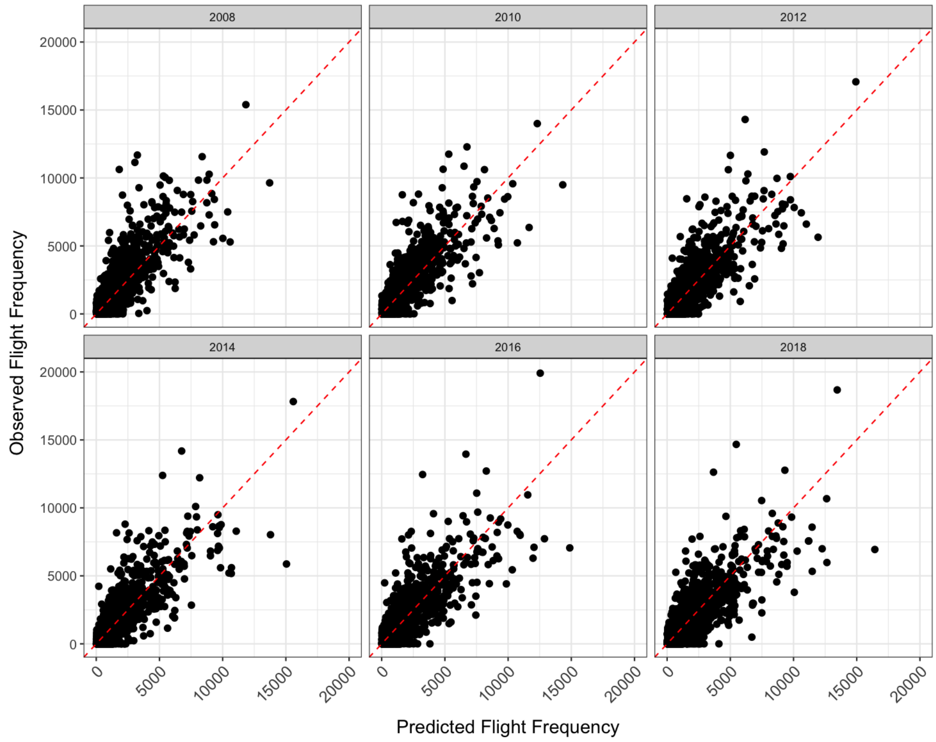

Figure 4.30: Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the MNL model. Case when considering the evolution of the US ATS.

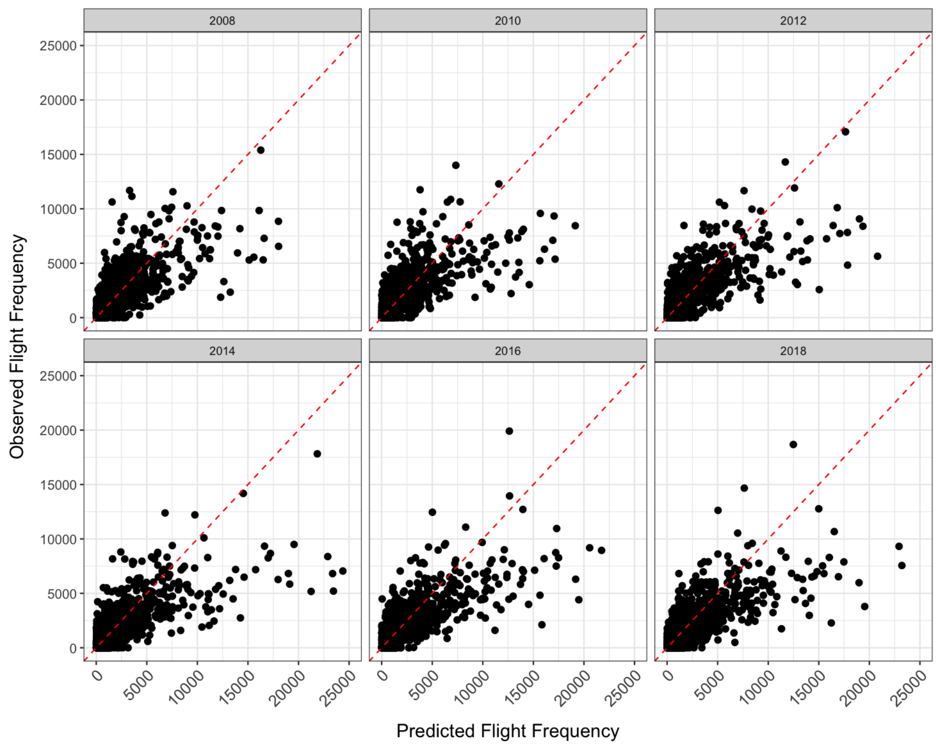

Figure 4.31: Comparison between predicted and observed traffic levels for years 2008, 2010, 2012, 2014, 2016 and 2018. Predicted values are obtained using segment passenger demand obtained from applying the NN model. Case when considering the evolution of the US ATS.

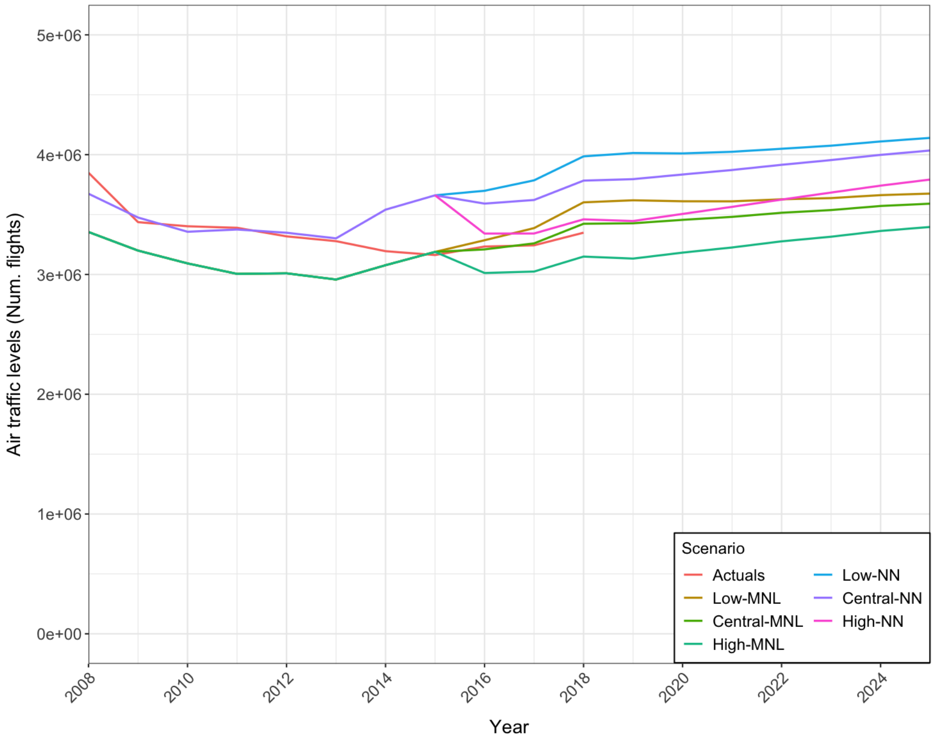

Figure 4.32: Total traffic level projections up to 2025 for the 2 sets of results: MNL and NN. Actual traffic levels up to 2018. Case when considering the evolution of the US ATS.

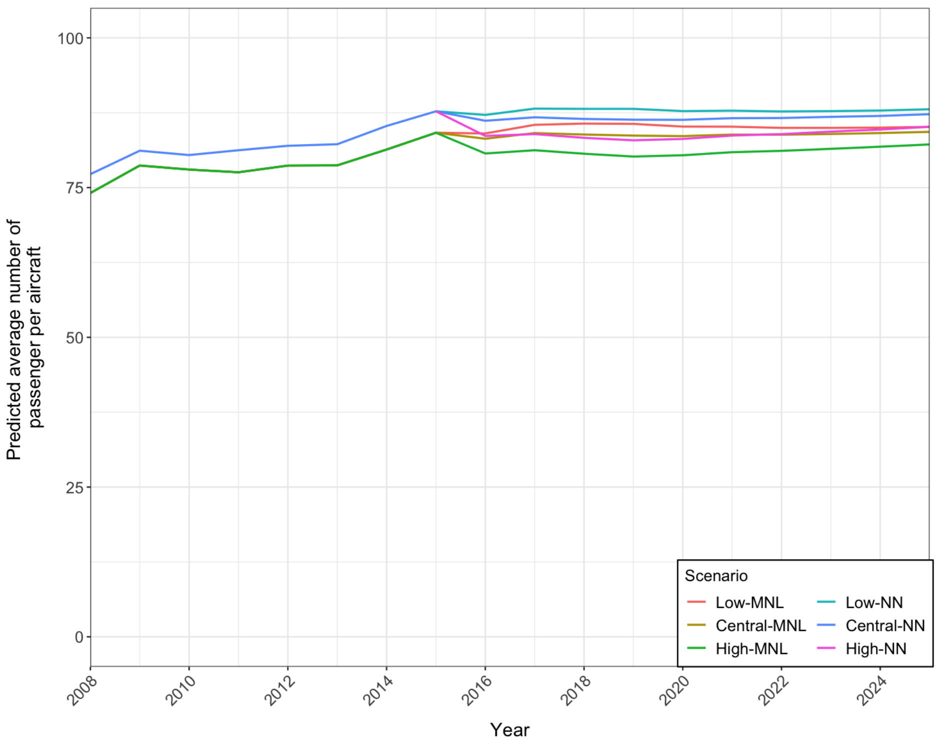

Figure 4.33: Average LF over time obtained up to 2025 for the 2 sets of results: MNL and NN. Actual traffic levels up to 2018. Case when considering the evolution of the US ATS.

Model	Scenario	2008	2018	2025	Change 2025 vs 2008
Actual		3,885,429	3,570,238		
MNL	Low	3,354,295	3,602,085	3,674,856	9.6%
	Central	3,354,295	3,422,943	3,590,924	7.05%
	High	3,354,295	3,149,278	3,396,427	1.36%
NN	Low	3,673,772	3,985,713	4,139,997	12.7%
	Central	3,673,772	3,783,158	4,034,515	9.8%
	High	3,673,772	3,459,970	3,792,052	3.2%

Table 4.15: Air traffic levels projections for 2008, 2018 and 2025 for the two sets of results and scenarios. Actuals for 2008 and 2018.

4.5 Comparison between Case 1 and Case 2

In order to assess the impact that network evolution has to the US ATS system results obtained from Case 2 are compared to those obtained in Case 1 (i.e. when the network was considered static). This is done considering results obtained only for the Central scenario, however results for the Low and High scenario yield a similar behaviour[8]. In order to performed the comparison, the following system-wide metrics for Case 1 and 2 are plotted: RPK; proportion of non-stop itineraries; and total number of flights. These metrics are shown in Figures 4.34, 4.35 and 4.36 respectively.

[8]Since the behaviour of the results from the low and high scenario are similar to those obtained in the central scenario, only the latter is plotted to help illustrate the differences.

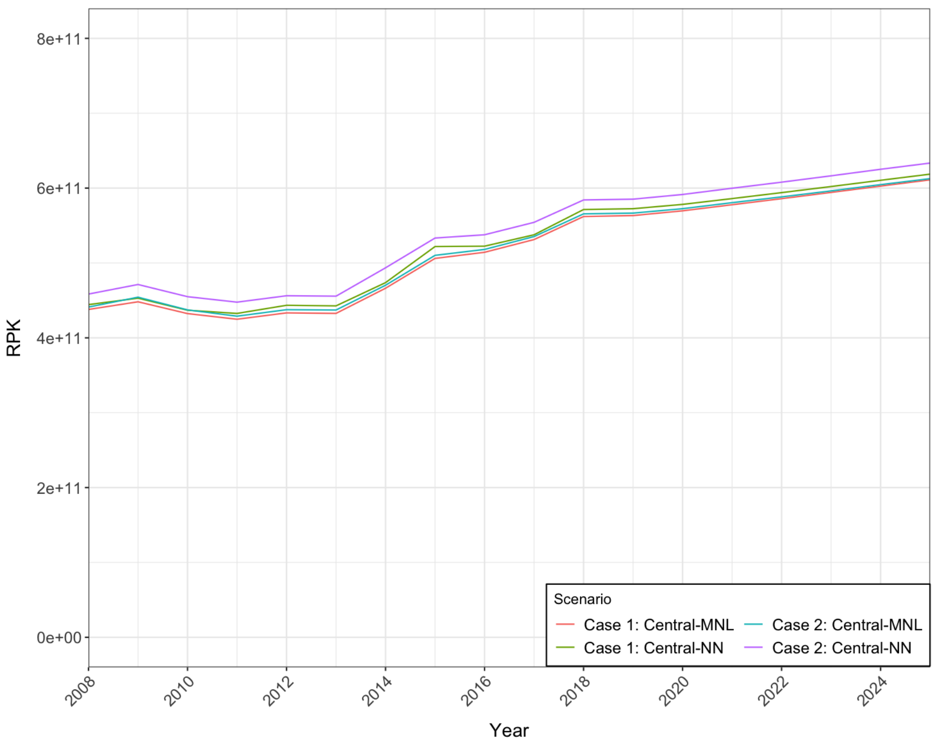

Figure 4.34: RPK projections for the Central scenario and considering no-evolution (Case 1) and evolution (Case 2) of the network.

From the comparison of RPK values obtained (Figure 4.34) it can be seen that in both cases (i.e. considering network evolution and without considering it) projected RPK values follow the same trend. As expected, results for the case when the NN model is used to estimate itinerary shares are higher than when using the MNL models since the former tends to predict a higher proportion of one-stop passengers than the latter, specially for the case when evolution is considered as it can be seen in Figure 4.35. From the comparison of proportion of non-stop passengers (Figure 4.35) it can be seen how NN models have a higher variability than the MNL models, suggesting that the former might be affected by external factors that have not been taken into account, such as the composition of the connections formed -i.e. the trend of proportion of non-stop passenger for the Case 2-NN model is smoother than those obtained for the Case 1-NN model -; or that the influence from factors such as fossil fuel prices might be larger than that experienced by the MNL models. Overall, results obtained by the MNL models are more in line with those observed currently in the network as seen in Sections 4.3.3 and 4.4.3. Considering that results obtained from the NN models tend to be larger than those obtained by the MNL models, results for the total number of flights are in line with expectations

(Figure 4.36). for example results obtained for Case 2 - NN show a peak in flight numbers in 2015 in line with the decrease of non-stop passengers predicted for this set of projections in 2015 (4.35); and a similar behaviour can be seen for the rest of years.

Results from the NN suggest that influencing factors towards itinerary choice are valued differently that in the MNL model, since the NN model favours stopping routes so much more than the MNL one - e.g. it might be that for the NN model journey time is less important. However, the characteristics of neural networks do not allow a direct interpretation of the influences of the input variables as it can be done with the MNL model (i.e. reason why neural networks are sometimes referred to as 'black box'). One way in which this could be done, and can be investigated in future work, is through LIME, which is a technique that attempts to understand the model by perturbing the input of data samples and understanding how the predictions change. Overall, results show that in a mature system such as the US ATS, air traffic flows are expected to not drastically change if network evolution occurs.

Finally, forecast trends obtained seem to be in line with those published by the FAA (2019), which forecasts that RPM will grow in the domestic market by 1.9% a year between 2019 and 2039. Considering the overall results for the Central scenario and for the entire period, yearly forecast RPK growth for Case 1 is 1.9% and 1.88% for the set of MNL and NN results respectively; whereas for Case 2, yearly forecast RPK growth is 1.87% and 1.84% for the set of MNL and NN projections respectively. Note that FAA's 2007 (2007) forecast expected a yearly average RPM growth of 2.5% for mainline carriers for the period between 2008 and 2020[9].

[9]Note that the comparison might not be like for like, since FAA forecasts includes a larger number of airports as well as regional services.

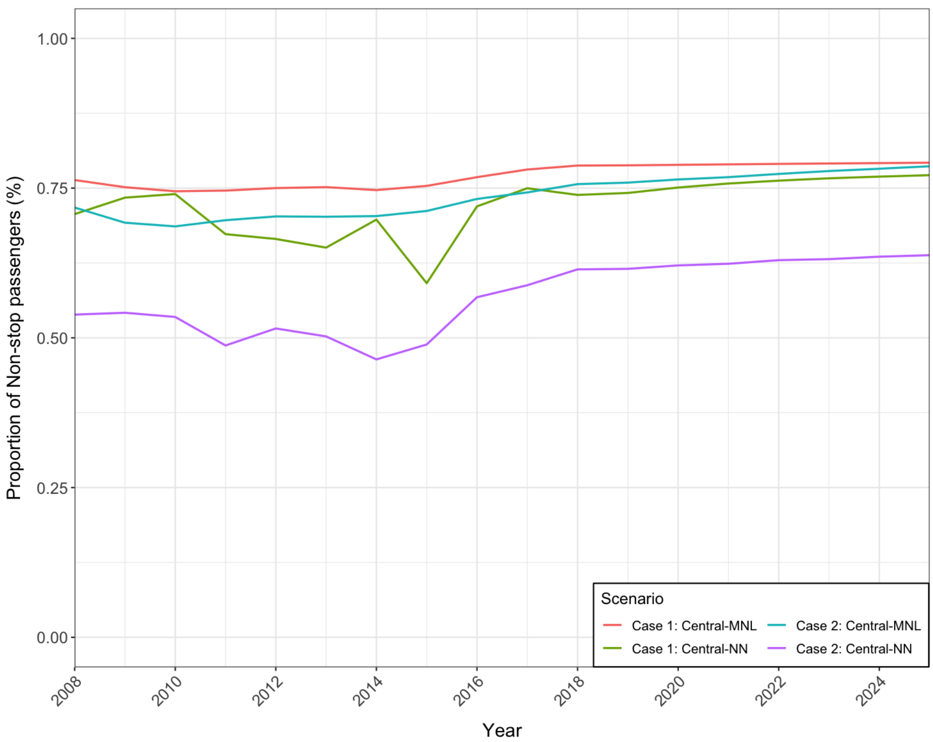

Figure 4.35: Proportion of non-stop passengers projections for the Central scenario and considering no-evolution (Case 1) and evolution (Case 2) of the network.

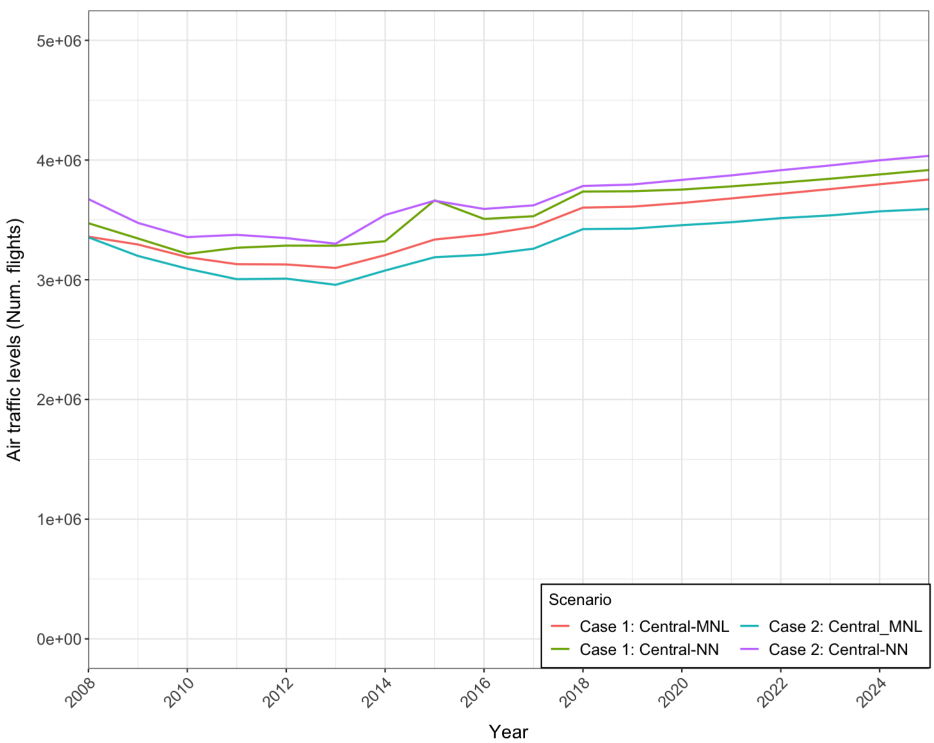

Figure 4.36: Air traffic levels projections for the Central scenario and considering no-evolution (Case 1) and evolution (Case 2) of the network.

Chapter 5

Conclusions and future work

5.1 Overall achievements compared to objectives

The work presented in this dissertation has studied the connectivity changes of an air transportation system and assessed the effect that those changes have on the rest of the network, such as route-level demand and air traffic levels. An airport connectivity model that predicts route addition and removal has been created and combined with an air itinerary choice model and an air traffic model to better reflect the impact of that evolution at the network level.

The review of forecasting models within aviation, discussed in Chapter 2, revealed that the latest forecasting methodologies do not consider the evolution of connectivity changes in the network, nor the impact that such changes have on the rest of the network. The current work has attempted to address this gap in modelling.

The development of an airport connectivity model has proved challenging, especially when the example application is a mature air transportation system (i.e. US) as discussed in Chapter 3. Model precision (percentage of correctly predicted connectivity changes over those that have been predicted as connectivity changes) achieved is low (i.e. 20%), which is believed to be due to the existence of factors that are impossible to model, specially for a mature network such as the US ATS, such as anticipating subsidies from airports to airlines to open new routes. However, some insights obtained from using network theory to understand network connectivity changes prove the potential of this area being applied to understand network evolution.

In the topic of air itinerary choice modelling, little research has been done at the network level on its own, with most of those existing using proprietary data and having their application limited to a subset of city-pairs mainly due to data availability and/or computational limitations during the estimation process, which limits the predictive capabilities of such models to a subset of cities; moreover, none of the existing research considers the dynamics of the network and therefore only itineraries available in the base year are considered as an option for future choices.

In the present work, a single itinerary choice model for an entire air transportation system has been created based on two alternative techniques, multinomial logit and neural network. This has allowed a comparative analysis of their predictive power when network evolution is considered.

A single model has been created using multinomial logit by using an approximation method to reduce the computational limitations; with accuracy levels comparable to those existing methodologies found in the literature, as discussed in Chapter 3. The single neural network model did not encounter the computational limitations, but the accuracy levels achieved were not as good as the multinomial logit model. It is expected that developing the neural network beyond the basic level explored in this thesis could improve the performance of the model; however further insight into why this model achieved poorer results than the MNL model would be required prior to such development.

For the air traffic model, the methodology used follows the widely used approaches found in the literature.

Overall, the objectives drawn at the start of this research have been met and, to the author's knowledge, this integrated research project, considering both network connectivity changes as well as the impact that these have at route- and segment-level is an original and valuable contribution to the body of knowledge.

5.2 Modelling framework

5.2.1 O&D demand model

An O&D passenger demand model has been developed using linear regression with logarithmic transformation. This research showed that O&D passenger demand for different distance groups is best modelled differently, as found in some cases in the literature. Splitting city-pairs by three different distance groups (i.e. short-, medium- and long-haul) and estimating a set of parameters for each of these distance groups improved the predicting power of the models, compared to a single model including all distance groups, from 0.5 to 0.8 as presented in Section 3.4.2; results that are in some cases higher than those found in the literature.

5.2.2 Airport connectivity model

Network evolution has been incorporated into the modelling framework through an airport connectivity model which predicts which airport-pairs would change their connectivity (i.e. either from connected to unconnected or vice-versa). The problem is split into two parts: a link addition model, which aims to predict those airport-pairs that are added to the network; and a link removal model, which aims to predict those airport-pairs that are removed from the network. The modelling approach explores the use of network theory metrics but, in contrast to previous work reported in the literature, additional input variables, such as passenger demand and demographic characteristics, are considered to improve the accuracy of the connectivity prediction. From these several possible input variables, the clustering coefficient is the main driver for the link removal model; whereas for link addition the important parameters are eigenvector centrality, distance between airport-pairs and whether neither or both airports are hubs. Other variables, such as O&D passenger demand, did not improve the models and therefore were disregarded.

The application of network theory metrics to study the connectivity of the US ATS produces similar results to those found in the literature: the US ATS follows a power-law degree distribution, by which most of the nodes have only few links while only few of the nodes have a high degree (consistent with a hub- and-spoke network); airport clustering coefficient is, in most cases, inversely proportional to node degree, which is a sign that some large airports might not be as robust to disruption (e.g. a link suddenly removed from the network or sudden closure of an airport) as other mid- and small- airports. The conclusion drawn is that the clustering coefficient might be capturing low-cost carriers' point-to-point strategy, rather than hub-and-spoke flag carrier's strategy, and therefore that the US ATS might be better modelled as a combination of distinct sub-networks.

The percentage of correctly predicted connectivity changes from both models is about 20%, which is quite low although consistent with results from the literature; however, the false positive rate (i.e. the percentage of actual negatives that are incorrectly identified as positive) for the link addition model is lower than those found in the literature, which shows the benefit of including a wider set of variables, beyond those associated with network theory, to improve this type of model. Moreover, the results showed that, although these models are not suitable for predicting airport-pair connectivity changes at the individual level, they seem to be able to capture the evolution of the network at the aggregate level (i.e. system-wide) in the short-term. This suggests that the incorrectly predicted individual connectivity changes predicted by the models nevertheless capture some of the characteristics observed in the wider system.

Overall, the airport connectivity work has revealed the much greater complexity needed to model network evolution, especially for cases when a mature system is considered. The comparison between observed and predicted system-wide network characteristics (discussed in Section 4.2) suggests that changes in fossil fuel prices may influence overall network characteristics, and consequently network evolution: for example, an increase of fossil fuel prices might lead to a decrease in connected links, due to the impact on airline's operational costs. Government expenditure in commercial aviation can also influence how the network develops, this being a reflection of the country's economy; and further consolidation in the industry would also lead to an increase in links removed, especially in small airports which are typically associated with less profitable routes.

5.2.3 Itinerary choice model

The impact of network evolution is first assessed by considering changes to air itinerary shares. Since the consideration of network dynamics in air itinerary choice modelling is a new area of research, this assessment is done by developing two itinerary choice models using two different modelling approaches to allow for a comparison: multinomial logit (MNL) and neural networks (NN). The work done in this topic appears to be a new contribution to this area, since a comparable research was not found in the literature.

A single MNL model applicable to the entire US ATS was successfully developed, overcoming computational limitations and endogeneity issues (associated with journey fare). Computational limitations were solved by applying the Berkson-Theil approximation method which transforms the maximum likelihood estimation process to a simple least-squares method. The endogeneity problem was solved by using a 2-stage control function method, by which predicted airfares are used as one of the input variables rather than actual fares; this is a common solution approach reported in the literature and described in Chapter 3. The combination of the two approaches is another significant contribution to the field of air itinerary choice modelling, since the high model accuracy obtained demonstrates the validity of using the Berkson-Theil approximation method with WLS, making this methodology comparable to the most common methods in the literature but with a higher degree of simplicity.

A single NN model was also successfully developed; however, results were not as good as those obtained from the MNL model and further work needs to be done to understand the reasons behind the respectable (but inferior to the MNL) accuracy achieved. The development and test of the NN model has shown the potential of this alternative methodology, which does not suffer from the computational limitations of MNL, and highlights that such models are have different sensitivities to the more commonly used MNL, for example sensitivity to journey time, fare and network changes. This suggests that other factors than those used in this study, such as fuel price, might yield a better fit when using neural networks.

5.2.4 Air traffic model

A linear regression model estimated using OLS estimator is used to model air traffic levels, as widely adopted in the literature. The set of input variables include the predicted passenger demand for a given airport-pair, longest runway length, distance between the airport-pair, load factor and hub information. Results are consistent with those found in the literature. The Model 2-OLS has also been compared with an earlier model using 2SLS as estimator with an auto-regressive term (i.e. Model 1-2SLS). In a mature system, such as the US, air traffic flows are expected not to change drastically, and therefore results from the Model 1-2SLS were unsurprisingly better. However, by using the auto-regressive term, the influence on air traffic levels from other variables, such as such as aircraft type limitation, is weakly captured, therefore limiting the applicability of this model; this was clear when projections of air traffic levels were generated for the US example and presented in Chapter 4.

5.3 Projections

Two sets of annual projections of passenger demand and air traffic levels at the US ATS level have been generated for the years 2008 to 2025: one considering a static network and the other considering network evolution. Overall projections show similar trends to actual observations; however, passenger demand is slightly over-predicted, leading to error propagation across the several models. Within the itinerary choice modelling stage, the MNL model has proved to generate projections more in line with observations; whereas the NN model shows greater sensitivity to input factors, since the variability of the results is greater.

Overall, the results have shown that changes in the structure of the network lead to a higher proportion of one-stop passengers being estimated, especially when the NN model is used. In reality, because the US has a relatively static structure, the proportion of non-stop passengers is not expected to change much over time (as the study of the US with the use of network theory metrics has shown). However, because the NN model seems to favour the hub-and-spoke network strategy, the predicted number of one-stop passengers is greater. Greater differences in results could be potentially seen if the model is applied to data from other regions of the world or if considering international flights, in which one-stop itineraries are needed to cover longer distances.

In addition, the comparison of results for a static network and for an evolving network have shown how certain characteristics of the network are more impacted by network evolution, such as the proportion of one-stop passengers; whereas the effect in others is softer, such as air traffic levels. However, all results showed that network changes have a relatively minor impact on overall system metrics in the case of the US ATS.

5.4 Suggestions for further work

Recommendations for further development of the research presented in this dissertation are presented below.

5.4.1 Airport connectivity

- Modelling airport connectivity of a mature system has proved challenging, since connectivity changes are relatively low. Better insight might be gained if network theory was applied to a less mature network, such as China, India or Brazil.

- For the specific case of the US, in which most of the airlines follow a hub-and-spoke strategy, the problem could be split based on the different sub-networks that exist since factors might be influencing those differently. Examples include between the point-to-point and the hub-and-spoke strategy; or considering the so-called community structure, in which groups of nodes have a high density of links amongst them but have a lower density of links between different groups.

- Comparison of the projected network characteristics (at the aggregated level) with those actually observed suggest that some other external factors, not captured by network theory metrics and not tested in this research, might influence the evolution of the network. While keeping the problem split into two models (link addition and link removal) the inclusion of additional factors could be explored, such as fossil fuel price changes or government expenditure. Also, the search space, especially for the link addition model (there being a large number of unconnected airport-pair candidates that could be added to the network) could be first narrowed down by using another classification technique such as support vector machine, random forest or even neural network, for example. This would reduce the level of data imbalance in the input dataset (i.e. a small percentage of unconnected airport-pairs changing to connected) and might help the logistic regression algorithm achieve better precision.

5.4.2 Itinerary choice modelling

- Although the results have shown the potential of using neural networks for modelling itinerary shares, it is clear that the neural network model is more susceptible to changes in airfare and/or time. In this study the neural network model specification was the same as the one used for the multinomial model, so it would be interesting to see whether considering a different set of input variables would improve model accuracy as well yielding insight into which factors have most influence on itinerary share within the neural network model.

- In this study one of the basic forms of neural network has been used; further investigation of regularization, optimization and loss could potentially lead to an improvement of performance accuracy of the neural network model. However, since the characteristics of the neural network do not allow a straightforward interpretation of the influences of the input variables, investigation could be first done through LIME, which is a technique that attempts to understand the model by perturbing the input of data samples and understanding how the predictions change. This would allow a better understanding of the strength and weaknesses of the current model and whether a better model refinement is needed.

5.4.3 Modelling framework

- The results obtained could be used to evaluate the impact of network dynamics at different levels of the network, such as airport congestion, flight delays and environmental impact through local and global emissions. For example, based on projected air traffic levels, the breakdown to aircraft type could be done, so that aircraft emissions could be calculated (e.g. the T-100 domestic segment dataset (2014) offers aircraft type information).

- Results presented in this dissertation are computed at the network level and referred to an annual timestamp. Further work could look at breaking down those aggregated projections to a smaller timestamp to see the impact of seasonality on those system-wide metrics, which is important for understanding flow patterns and which would help the resource planning of industry stakeholders. This could be done at the quarterly level, since information regarding airline tickets is quarterly provided by the US Bureau of Transportation Statistics (2014; 2014).

- In this research the impact of flight schedule changes due to network evolution on overall demand (i.e. passenger demand between city-pairs) has not been considered. Further work could address the level of passenger sensitivity to flight schedule and the potential effect that changes to overall passenger demand would have on itinerary shares. In this research, journey fares and flight times for those new links have been simply calculated by multiplying average fare or flight time by the flight distance; further enhancement could be achieved by including a fare model into the framework which would allow a better assessment of the impact of network evolution on the overall passenger demand.

Appendix A

Socio-economic data sources

As mentioned in Chapter 3, socio-economic information for years between 2008 and 2016 is taken from the US Census Bureau (2014a) and the Bureau of Economic Analysis (2014) websites. For population projections corresponding to further years, the different state governmental websites have been consulted since the above mentioned source only provides a country growth rate for years after 2016. Since socio-economic information has been compiled for each of the cities considered in this research, in an attempt to capture a more accurate evolution of this information, a large set of sources have been used. The following list gathers the several state websites that have been consulted. For those states in which no population information by city existed a flat overall growth/decline rate for the overall state has been applied. Economic growth information by city is available only for the first set of years (i.e. 2008-2016). From 2017 onwards, state annual growth/decline rate has been assumed. This was extracted from the economic tables produced by the FAA, in order to be line with their forecasts (FAA, 2018).

- Alabama: https://cber.cba.ua.edu/edata/est_prj.html
- Alaska: http://live.laborstats.alaska.gov/pop/projections/pub/popproj.pdf
- Arizona: https://population.az.gov; https://population.az.gov/sites/default/files/documents/files/pop-estimates2014-04pla.pdf
- California: http://www.dof.ca.gov/Forecasting/Demographics/Projections/
- Colorado: https://demography.dola.colorado.gov/population/population-

totals-colorado-substate/#population-totals-for-colorado-and-sub-state-regions

- Connecticut: https://ctsdc.uconn.edu/2015_2025_projections/
- Delaware: https://stateplanning.delaware.gov/information/dpc_projections.shtml
- District of Columbia: https://planning.dc.gov/sites/default/files/dc/sites/op/publication/attachments/District%20of%20Columbia%20QuickFacts_2016.pdf; https://planning.dc.gov/node/1212966
- Florida: http://edr.state.fl.us/Content/population-demographics/data/index.cfm; http://edr.state.fl.us/Content/population-demographics/data/CountyPopulation_2016.pdf
- Georgia: State annual population growth rate has been applied as follows: 1.06% between 2016 and 2010 and 1.05% between 2020 and 2025. Sources: http://www.georgialibraries.org/lib/stategrants_accounting/pop-projections.php; http://www.georgialibraries.org/lib/construction/georgia_population_projections_march_2010.pdf
- Guam: Several sites have been consulted as follows: http://www.investguam.com/public-finance-opportunities/economic-indicators/; http://population.city/guam/; https://www.livepopulation.com/population-projections/guam-2020.html
- Hawaii: https://dbedt.hawaii.gov/economic/databook/2012-individual/_01/; http://files.hawaii.gov/dbedt/economic/data_reports/2040-long-range-forecast/2040-long-range-forecast.pdf; http://www.hiloagent.com/images/Hawaii%20County%20Population%20by%20District.pdf
- Iowa: http://www.iowadatacenter.org/browse/projections.html
- Idaho: State annual population growth rate has been applied as follows: 1.09% between 2016 and 2010 and 1.08% between 2020 and 2025.
- Illinois: https://www2.illinois.gov/sites/hfsrb/InventoriesData/Pages/Population-Projections.aspx

- Indiana: http://www.stats.indiana.edu/pop_proj/
- Kansas: https://www.ipsr.ku.edu/ksdata/ksah/population/
- Kentucky: http://www.ksdc.louisville.edu/wp-content/uploads/2016/10/projection-report-v16.pdf; http://www.ksdc.louisville.edu/data-downloads/projections/
- Louisiana: http://louisiana.gov/Explore/Population_Projections/
- Maine: https://www.maine.gov/economist/projections/index.shtml; https://www.maine.gov/economist/projections/pub/MaineCityTownPopulationProjections2034.pdf
- Maryland: http://www.mdp.state.md.us/msdc/S3_Projection.shtml
- Massachusetts: https://pep.donahue-institute.org
- Michigan: http://www.michigan.gov/documents/8515_26106_7.pdf
- Minnesota: https://mn.gov/admin/demography/data-by-topic/population-data/our-estimates/; https://mn.gov/admin/demography/data-by-topic/population-data/our-projections/
- Mississipi: http://www.mississippi.edu/urc/downloads/PopProjections/PopulationProjections.pdf
- Missouri: https://oa.mo.gov/budget-planning/demographic-information/population-projections/2000-2030-projections
- Montana: http://ceic.mt.gov/Population/PopProjections_StateTotalsPage.aspx
- Nebraska: https://www.osbm.nc.gov/demog/county-projections
- Nevada: http://nsla.nv.gov/Library/StateDataCenter/NVProjections_Links/; http://nvdemography.org/wp-content/uploads/2014/06/2013-Nevada-Summary-Workbook-ASRHO-Estimates-and-Projections-REV-051614-B.pdf; https://tax.nv.gov/uploadedFiles/taxnvgov/Content/TaxLibrary/2017-20-Year-Total-Population-Projections-Report.pdf
- New Jersey: https://www.commerce.nd.gov/census/Demographics/

- New Mexico: https://gonm.biz/site-selection/census-data/
- New York: http://pad.human.cornell.edu/che/BLCC/pad/data/projections.cfm; http://pad.human.cornell.edu/counties/projections.cfmf
- North Carolina: https://www.osbm.nc.gov/demog/county-projections
- North Dakota: https://www.commerce.nd.gov/census/Demographics/
- Ohio: https://development.ohio.gov/reports/reports_pop_proj_map.htm
- Oklahoma: https://okcommerce.gov/data/demographics/
- Oregon: http://www.oregon.gov/das/OEA/Pages/forecastdemographic.aspx
- Pennsylvania: https://pasdc.hbg.psu.edu/Data/Projections/tabid/1013/Default.aspx
- Puerto Rico: http://ddec.pr.gov/es/blog/wp-content/uploads/2016/12/Estudio-economico-2016-2030.pdf; https://www.politico.com/f/?id=00000152-57f8-dad1-a977-77fd856b0000
- Rhode Island: http://www.planning.ri.gov/planning-areas/demographics/data/population-projections.php; http://www.planning.ri.gov/documents/census/tp162.pdf; http://www.dlt.ri.gov/lmi/census/pop/townest.htm
- South Carolina: http://sccommunityprofiles.org/census/proj_c2010.html
- South Dakota: https://www.sdstate.edu/sociology-rural-studies/census-data-center/age-and-sex-structure
- Tennesse: http://tndata.utk.edu/sdcpopulationprojections.htm
- Texas: http://txsdc.utsa.edu/data/TPEPP/Projections/Index
- Utah: https://gomb.utah.gov/budget-policy/demographic-economic-analysis/
- Vermont: http://accd.vermont.gov/sites/accdnew/files/documents/CD/CPR/ACCD-DED-VTPopulationProjections-2010-2030.pdf
- Virginia: https://demographics.coopercenter.org/virginia-population-estimates; http://demographics.coopercenter.org/virginia-population-projections/

- Washington: https://ofm.wa.gov/washington-data-research/population-demographics/population-forecasts-and-projections/growth-management-act-county-projections/growth-management-act-population-projections-counties-2010-2040-0
- West Virginia: http://busecon.wvu.edu/bber/pdfs/BBER-2014-04.pdf
- Wisconsin: https://doa.wi.gov/DIR/FinalProjs2040_Publication.pdf; https://doa.wi.gov/Pages/LocalGovtsGrants/Population_Estimates.aspx; https://www.wisconsin-demographics.com/counties_by_population

Appendix B

Value of time and CPI calculation

B.1 Value of Travel Time

Value of travel time is an important factor to consider when evaluating the benefits of new transport infrastructure and a tool used to aid the decision making process on transportation investments and rule making initiatives. Travel time is seen as having a negative impact on demand, and directly related to traveller's willingness to pay to reduce it. There is not a standard value of travel time as it depends on different factors such as the purpose of the trip, the traveller, the circumstances of the trip, etc. However, research shows that a large part of trips share similar purposes and characteristics - e.g. commuting to work- (U.S. DoT, 2014).

For the purpose of this dissertation a generic and broadly representative value of time has been used, since predictions in this study are evaluated at an aggregate level and because of the infeasibility of calculating an individual's value of time. For these reasons, in order to calculate the value of time in this dissertation, the US Department of Transportation methodology (2014) has been used. The US DoT methodology consists of a weighted average between the value of travel time for business purposes and the value of time for personal/leisure purposes. In Table B.1 business and personal trip weights used to calculate the value of time for years 2000, 2009 and 2013 are shown (U.S. DoT, 2014). Weights for the in-between years have been calculated by linearly extrapolating.

Trip Purposes	2000	2009	2013
Business	31.7%	40.4%	44.4%
Personal	68.7%	59.6%	55.6%

Table B.1: Business and personal/leisure travel weights used to calculate value of time.

Following the US DoT methodology for air travel (2014), business' value of time ($VOT_{business}$) is calculated by multiplying a factor of 2.5 times the median gross wage for all occupations - Eq. B.1-. The factor 2.5 is derived from using distinct wage information and calculated through dividing household income of air travellers by median household income as shown in Equation B.2. Values used are extracted from the BTS National Household Travel Survey (NHTS) of 2001, since no other survey of such type has been conducted since 2001-. This $factor_{business}$ is assumed to be constant across the entire period considered in this dissertation.

$$\begin{aligned} VOT_{business} &= factor_{business} * Median\ Gross\ Wage \\ &= 2.5 * Median\ Gross\ Wage \end{aligned} \tag{B.1}$$

$$factor_{business} = \frac{Household\ income\ of\ air\ travellers}{median\ household\ income} = \frac{\$105,000}{\$42,228} = 2.5 \tag{B.2}$$

Median gross wage for all occupations is calculated using Equation B.3. Data is obtained from from the Bureau of Labour Statistics (BLS). The median hourly wage is obtained from the BLS Occupational Employment and Wages Estimates (2014c) and corresponds to civilian workers including private industry and state and local government workers. The estimated hourly benefit is approximated by multiplying the median hourly wage by the ratio of mean gross compensation to mean money wages as shown in Equation B.4 and is obtained from BLS Employer Costs for Employee Compensation (2014b).

$$Median\ Gross\ Wage = Median\ hourly\ wage + estimate\ hourly\ benefits \tag{B.3}$$

$$Estimate\ hourly\ benefits = (\frac{Avg.\ benefits}{Avg.\ salary\ and\ wages}) * Median\ Hourly\ Wage \quad (B.4)$$

Personal air travel time value ($VOT_{personal}$), shown in Equation B.5, is estimated by multiplying a factor of 1.9 times 70% of the hourly median household income which is calculated by dividing the median household income by 2,080 hours -i.e. annual working hours -. $factor_{personal}$ is defined by the NTHS as the ratio of the 2001 median household income of air travellers on personal business to the nation-wide median household income of 2001. Again, the value of this factor is assumed to be constant across the entire period considered in this dissertation. Information regarding the mean household income is obtained from the US Census Bureau website (2014b).

$$\begin{aligned} VOT_{personal} &= factor_{personal} * (0.7 * \frac{Median\ household\ income}{2,080}) \\ &= 1.5 * (0.7 * Hourly\ Median\ household\ income) \end{aligned} \quad (B.5)$$

Once the value of air travel time for business and personal purposes is calculated for each year and in terms of \$/hours, value of air travel time (VOT) is calculated applying the weights presented in Table B.1. Value of delayed travel time (VOT_{Delay}) is assumed to be 3 times the value of travel time following the assumption of other research (Evans, 2010). Value of travel time is computed using the sources mentioned above for the years between 2003 and 2014. For further years in which future projections are generated, since data was not available an annual growth rate of 0.5% has been assumed. This corresponds to the average annual growth rate between 2009 and 2014. Table B.2 presents the values of travel time and delayed travel time for the entire period (in 2007's dollars).

Year	VOT	VOT_{Delay}	Year	VOT	VOT_{Delay}
2003	$38.9	$116.7	2015	$41.4	$124.1
2004	$39.2	$117.5	2016	$41.6	$124.7
2005	$39.5	$118.5	2017	$41.8	$125.3
2006	$40.0	$119.9	2018	$42.0	$126.0
2007	$40.5	$121.6	2019	$42.2	$126.6
2008	$40.0	$119.9	2020	$42.4	$127.2
2009	$40.7	$122.0	2021	$42.6	$127.9
2010	$40.4	$121.2	2022	$42.8	$128.5
2011	$40.3	$120.9	2023	$43.0	$129.1
2012	$40.3	$120.9	2024	$43.3	$129.8
2013	$40.8	$122.4	2025	$43.5	$130.4
2014	$41.2	$123.5			

Table B.2: Value of business and personal travel time for years between 2003 and 2025.

B.2 Consumer Price Index

Across the modelling approach presented in this dissertation, several variables are related to economic information, such as the mean household income and airfare values. Because 2007 is the threshold year: with years up to 2007 used to estimate the model, and with years afterwards being those in which projections have been generated, all monetary data have been converted to 2007's dollars. In order to do that the Consumer Price Index (CPI) is defined by the BLS as *a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services* (Bureau of Labour Statistics, US Department of Labour, 2018).

The BLS methodology is followed to convert monetary value to 2007 US$ (Bureau of Labour Statistics, US Department of Labour, 2014a). CPI values for all years are obtained from the BLS Consumer Price Index Detailed reports (2014a) with 2007's value used as the reference index. The difference between both indexes is computed so that the index point change is obtained (i.e. Equation B.6). In order to generate the factor value to which monetary data needs to be multiplied to obtain the value of money in 2007 dollars, the ratio between the index point change and 2007's CPI value is calculated. Since this is expressed as percentage change to 2007's value, it can be converted to a multiplying factor by the sum of this change with 1 as shown in Equation B.7. Table B.3 presents the CPI values for years between 2003 and 2014 as well as the factor used to convert monetary data to 2007's dollars.

$$IndexPointChange = CPI_{2007} - CPI_{year} \tag{B.6}$$

$$factor_{CPI} = 1 + \frac{\textit{Index Point Change}}{CPI_{2007}} = 1 + \frac{\textit{Index Point Change}}{CPI_{2007}} \tag{B.7}$$

Year	CPI_{2007}	CPI_{year}	Index Point Change	$factor_{CPI}$
2002	207.3	179.9	27.4	1.15
2003	207.3	184.0	23.4	1.13
2004	207.3	188.9	18.4	1.10
2005	207.3	195.3	12.0	1.06
2006	207.3	201.6	5.7	1.03
2007	207.3	207.3	-	1.00
2008	207.3	215.3	-8.0	0.96
2009	207.3	214.5	-7.2	0.97
2010	207.3	218.1	-10.7	0.95
2011	207.3	224.9	-17.6	0.92
2012	207.3	229.6	-22.3	0.90
2013	207.3	233.0	-25.6	0.89
2014	207.3	236.7	-29.4	0.88

Table B.3: $Factor_{CPI}$. calculation in order to convert monetary data to 2007's dollars.

Appendix C

Previous attempts - O&D demand model

This appendix presents a brief summary of previous attempts to model O&D passenger demand. Differences between the modelling approaches presented here and the one in Chapter 3 are based on the model specification and and model applications that those have; the methodology has not changed so that previous attempts also used linear regression with logarithmic transformation as explained in Chapter 3. This section is split into two: first previous modelling attempts are presented; and then, the error investigation performed to understand the large error obtained from those early attempts is also presented.

C.1 Previous modelling specifications

Initially, the approach to estimate the O&D passenger demand was based on a single model for the entire US air transportation system. Data used to train the model referred to 2007 and input variables included population, mean household income per capita, generalised cost, a dummy variable indicating whether the city-pair is also connected by other transport modes such as train (i.e. R) and a set of two dummy variables indicating whether none or one of the cities in the pair is a major tourism or business destination (i.e. *S1* and *S2*). Two sets of coefficients were estimated based on the estimation process: using ordinary least squares (OLS); using weighted least squares (WLS) with number of passengers as weight. Estimation results are presented in Table C.1 for both OLS and WLS models.

All estimated coefficients obtained are statistically significant at the 95% confidence level. Also all coefficients are of the expected sign with exception of the dummy variable indicating whether the city-pair is also connected by other transport modes such as train (i.e. R) for the WLS model. This suggests that the model does not capture competition from other transport modes well when used in this type of model formulation, and therefore it should be dropped.

Variables	OLS	WLS
Intercept	10.54^{***}	7.81^{***}
Population	1.015^{***}	1.20^{***}
Income	1.32^{***}	1.47^{***}
S1	0.58^{***}	1.14^{***}
S2	-1.34^{***}	-1.13^{***}
R	-1.56^{***}	2.33^{***}
Generalised cost	-0.87^{***}	-1.29^{***}
Adjusted R^2	0.614	0.795
Error in validation	-36%	-34%

$^{***}p < 01$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table C.1: 1st attempts: estimated coefficients obtained when using OLS and WLS as estimation process. Error refers to the average difference of predicted total network passenger demand against observed during the validation years (2008-2013).

C.2 Error Investigation

Large error (i.e. about -35%) is obtained due to over-predictions as shown in Figure C.1, which shows the error distribution clearly left-skewed (i.e. a sign of over-prediction). Further checks are performed to investigate the source of such high over-predictions, which could be due to the following reasons:

- Most of the over-predictions occur on high demand O&Ds - solution would be adding extra dummy variables such as variables related to high speed rail network.
- Most of the over-predictions occur on low demand O&Ds - the solution would be excluding those O&Ds with low demand during the training and validation phase.

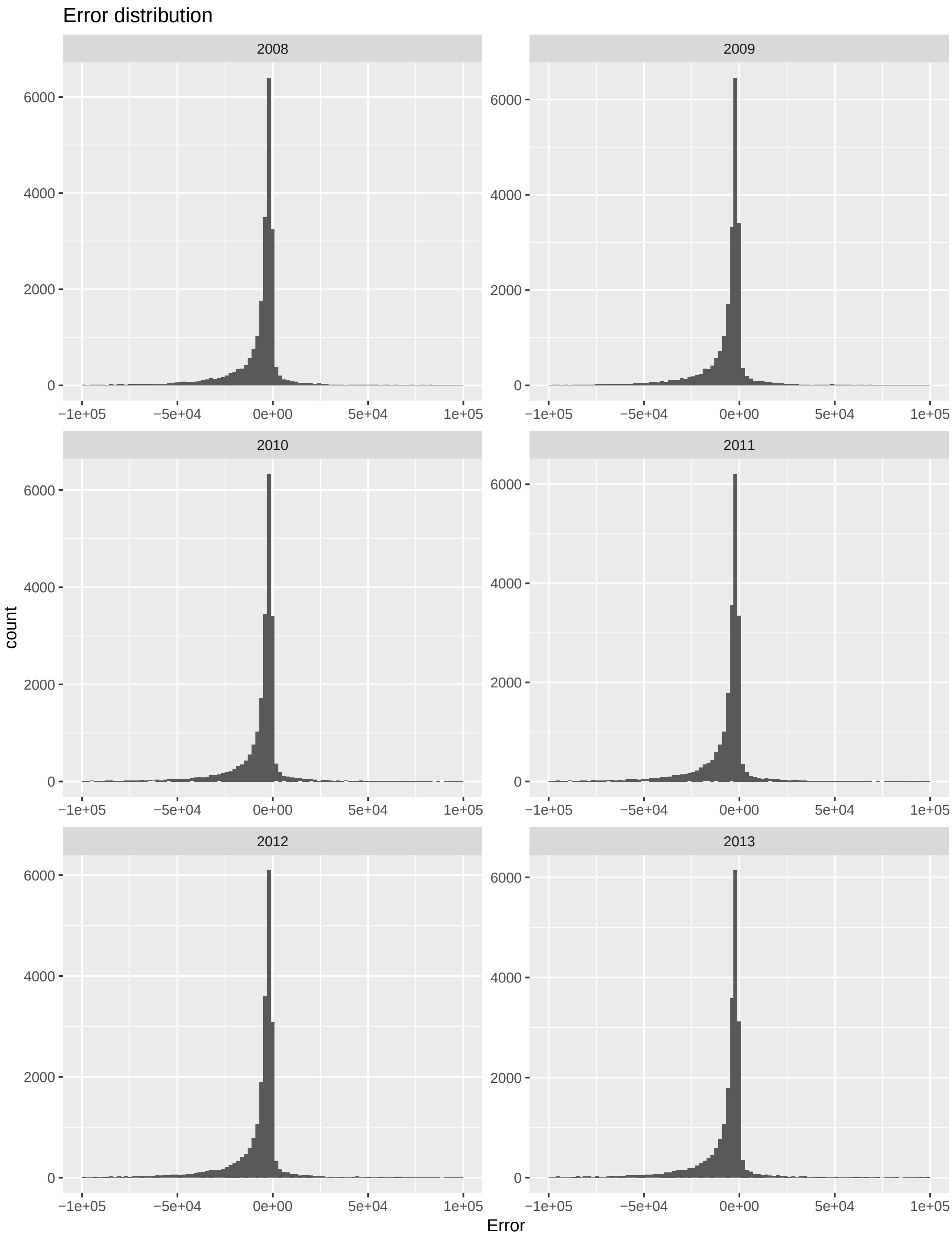

Figure C.1: Error distribution throughout the validation years, closed up to error values between -100,000 and +100,000.

Few case examples for which a large error is obtained are checked against the source data (BTS-RITA, Bureau of Transportation Statistics - Research and Innovative Technology Administration, 2014), revealing that those O&D have a much lower passenger demand than expected based on the city-pair attributes. The set of examples checked are the following:

- NYC&IAD - both pair of cities have high income and population and are a major tourism and/or business destination, but demand is shown to be between 600k and 700k passengers in 2014. Projected values are about 3 million for the same year.
- NYC&BOS - both cities with associated high income and population with passenger demand to be about 700k passengers in 2014. Projected levels of demand are about 2 million passengers for the same year.
- MIA&TPA - both cities are a major tourism destination located quite close to each other. Observed demand is just below 200k in 2014; however, projected demand is about 1 million for the same year.
- DFW&IAD - similarly to the cases above observed demand is about 600k a year in 2014; however projected demand is about 1.8 million passengers for the same year.

Case examples such as the above are exceptional cases of city-pairs that theoretically should have a higher demand but they do not - i.e. from literature, air travel is believed to be income elastic, and with all the city-pairs above having associated a high income variable as well as most of them being a major tourism and/or business destination, their related passenger demand would be expected to be much higher than the observed. Note that the above examples are likely to have a lower demand than expected due to competition with mainly road transport.

A further check includes looking at the distribution of demand across all O&Ds in bins of 10,000 passengers and calculate how much error these bins have associated. Figure C.2 shows total error per bin against mean passenger demand per bin. The color of the points shows the number of O&Ds in each bin. Figure C.2 shows that bins with lower mean passenger demand make most of the error; and therefore would confirm that the source of the error is in line with hypothesis two stated above - i.e.

most of the over-predictions occur in low demand O&Ds. The bin with the most error is the first one, which contains O&Ds with less than 10,000 passenger, and has associated passenger demand from about 18,819 O&Ds with a mean average passenger demand of 1,785; the total error associated to this first bin is 1.16e+08. The second bin, which contains O&Ds with a passenger demand between 10k and 20k, has a mean passenger demand of 14,120 per O&D and sum up a total of 2.96e+07 error.

Because O&Ds with less than 20,000 passenger demand account for around 42% of the error, further exploration of those O&Ds is done. Using bins of 500 passengers, Figure C.3 shows total error per bin against mean passenger demand per bin. Again, most of the error is gather within the first two bins (O&Ds with less than 1,000 passengers per year), which accounts for about 22.5% of total error. Mean passenger demand for first bin is 194 passengers, while for the second bin is 729 passengers. The over-prediction amongst this low demand O&Ds is expected, because given the way the model is formulated demand cannot go below 0, i.e. a city-pair with yearly demand of 200 can only under-predict by 200, but can over-predict by an unlimited amount.

The model presented above was re-trained only using O&Ds with more than 20k, 25k, 30k and 40k passengers and although error obtained for the model when considering O&Ds with more than 25k passengers had an improved performance with respect the first attempts (i.e. error decreases just below 30%), results were not accurate enough and an alternative solution, such as the one presented in Chapter 3 was then pursued.

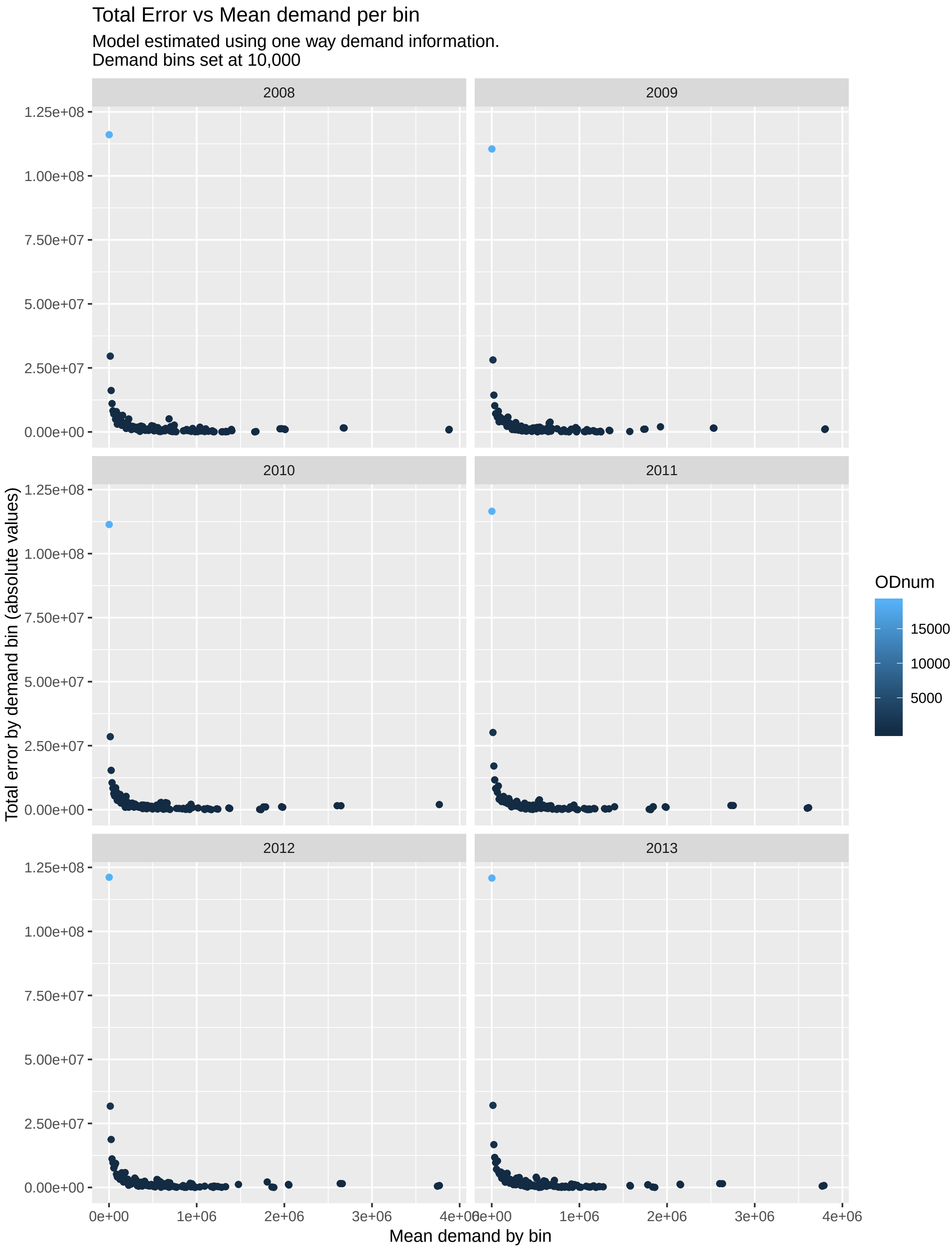

Figure C.2: Total error against mean passenger demand per bin. Bin size 10,000. Range years 2008-2013.

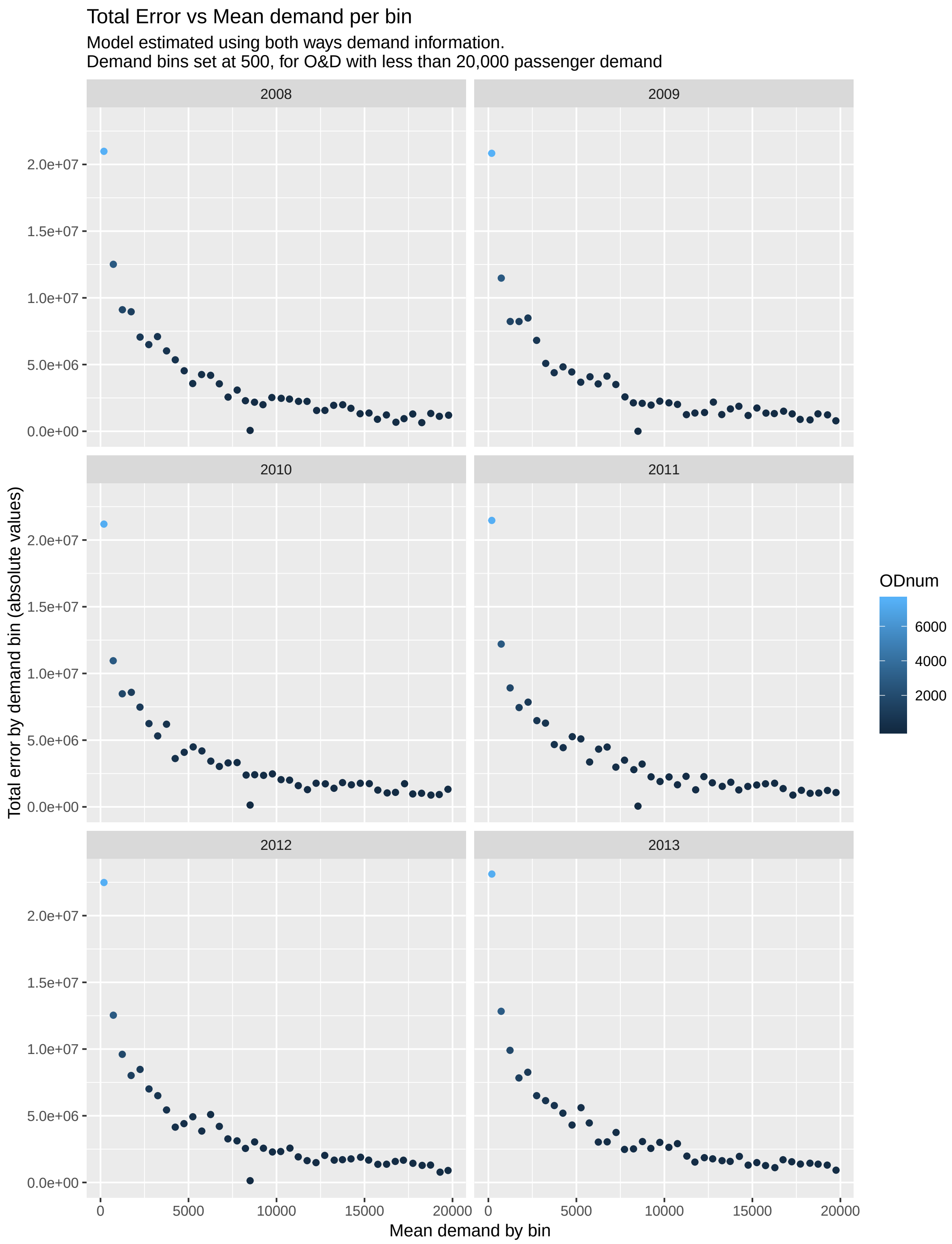

Figure C.3: Total error against mean passenger demand per bin for O&Ds with less than 20,000 passenger demand. Bin size 500. Range years 2008-2013.

Appendix D

Previous attempts - Airport connectivity

This appendix presents a brief summary of previous attempts to model airport connectivity which involved a modelling approach with a single model aimed to predict the connectivity of airport-pairs (i.e. '1' if connected, and hence, a flight service exists between them; '0' if disconnected, and hence, a flight service does not exist between them) rather than connectivity changes in airport-pairs (i.e. an airport-pair going from connected to disconnected or vice versa). Initially, O&D passenger demand[1] was included in the model specification as the early attempts below show; however given that the problem involved the identification of a small number of connectivity changes, errors obtained were relatively large and another approach was eventually used as presented in Chapter 3.

Section D.1 below presents some of the specifications used for when O&D passenger demand was included amongst the input dataset. Section D.2 presents some illustrative figures related to links added and removed in 2011 and 2012, which helps to visualise the fact that there is not a clear pattern on where the connectivity changes occur.

[1]Note O&D passenger demand is associated to the city-pair which the airport-pair under study connects.

D.1 Airport connectivity models including O&D passenger demand

Early attempts to model airport connectivity are similar to the work presented in Chapter 3 as per using logistic regression and data from 2007 as training data and from 2008-2013 for validation. However set of explanatory variables differ and initial attempts did not split the problem in two (i.e. link addition and link removal). In early attempts, model specification included a larger set of explanatory variables which are briefly described below.

- Population - Combined variable calculated through Equation 3.2 using population values associated for each of the airports being part of the link under study.
- City attractiveness - set of two dummy variables indicating whether none or one of the cities in the pair are a major tourism or business destination.
- Hub information - set of two dummy variables indicating whether none or both of the airports forming the airport-pair are considered hub airports.
- Fuel price - average annual fuel price.
- O&D passenger demand - passenger demand between two cities. Given the case of multi-airport cities, the same O&D passenger demand will be allocated from airports from same cities. Referred as N_{od} in Table D.1.
- Previous year existing link between a given O&D - binary variable that captures whether a given O&D was connected by any available flight service the previous year (i.e. referred as $Link_{t-1}$ in Table D.1):
  - '0' when none of the possible airport-pairs connecting a given O&D had a flight service operating the previous year.
  - '1' when at least one of the possible airport-pairs connecting a given O&D had a flight service operating the previous year.
- Count of connections within a given O&D - set of two variables that look at the degree[2] of the origin and destination airport within a given O&D. For

[2] Airport degree is the number of airports that the airport is connected to (i.e. a flight service exists)

example, given the following O&D: Atlanta and Tri-cities/Greeneville; there are two possible airport-pairs that can connect those two cities: ATL-TRI and/or ATL-GCY; however, in 2006 only ATL-TRI was connected by flight, making the degree of ATL and TRI within this O&D equal to 1, while the degree of GCY is 0. This is referred as $O\&D_{degree,i}$ for origin airport i and $O\&D_{degree,j}$ for destination airport j in Table D.1

- Network theory variables [3]- node degree (k), node weight (s), node eigenvector centrality (EVC) and clustering coefficient (CC). Note that these metrics are included in the model either as a combined variable (Equation 3.9) or individually for each of the airports forming the airport-pair, such as k_i would be referred to node degree of airport i and k_j would be referred to node degree of airport j.

The first attempt, considered the following input variables: O&D passenger demand, population, city attractiveness, hub information, fuel price and the set of all network theory metrics computed as a combined variable. This early attempt resulted on a true positive rate - i.e. which measures the percentage of actual positives (connected) which are correctly identified -of 66.58% across all validation years; whereas the true negative rate - i.e. which measures the percentage of actual negatives (disconnected) that are correctly identified as such - obtained was 97.65%. These results showed how the model had a good accuracy at predicting the majority class (i.e. unconnected airport-pairs) while failed to identify the minority class (i.e. connected airport-pairs).

After this first attempt another set of model specifications were considered. Those included newly created input variables such as the count of connections within a given O&D or the existence of a direct connection the previous year. A set of model specifications were considered as shown in Table D.1.

During training the algorithm aims to minimise the overall error, since the training dataset is imbalanced - i.e. the number of cases of one class (i.e. in this case the number of unconnected airport-pairs), outnumbers the number of cases for the other class - the minority class will contribute little to this minimisation process; and therefore estimates would tend to be biased towards reducing the error

[3]Please refer to Chapter 3 for explanation on network theory metrics.

of the majority class. In order to avoid classifier bias towards the majority class, 5 different techniques are used to balance the dataset. Note that during the evaluation phase when the problem was split into two (i.e. link addition and link removal), the scenario when no technique is used to artificially balance the dataset yielded the best results and therefore the actual data has been used to develop the models presented in Chapter 3. The techniques used to artificially balance the dataset when a single model was used are briefly described below.

- Over-sampling - this techniques looks at replicating the observations from minority class to balance the data.
- Under-sampling - this techniques looks at sampling the majority class so that the number of observations within the majority class is similar to the minority class size.
- Combination of over&under sampling - this technique looks at applying over-sampling at under-sampling at the same time and finding a mid-point size for both classes.
- Random Over-Sampling Examples (ROSE) (Lunardon et al., 2014) - this technique uses smoothed bootstrapping to generate artificial data from the feature space neighbourhood around the minority class.
- Synthetic Minority Oversampling Technique (SMOTE) (Ganganwar, 2012)- this technique aims to generate artificial data of the minority class. SMOTE algorithm creates artificial data based on feature space similarities from minority samples. Increasing the number of minority class observations is an attempt to shift the classifier learning bias towards the minority class. This technique can only be applied with continuous data, therefore SMOTE has not been used for model number 5.

	Model number				
Variable	1	2	3	4	5
N_{od}	X	X	X	X	
$Link_{t-1}$					X
$O\&D_{degree,i}$					X
$O\&D_{degree,j}$				X	
k_i	X		X		
k_j	X		X		
s_i				X	
s_j				X	
EVC_i	X	X	X	X	
EVC_j		X	X	X	X
CC_i	X	X	X	X	
CC_j	X	X	X	X	

Table D.1: Model specifications used during early attempts to model airport connectivity.

The above models were validated using data from years 2008-2013 and the performance metrics and error rates calculated to evaluate model performance are presented in Table D.3 and described as per below. Note that because in this case a single model is applied to the entire air transport system, actual positives refers to connected airport-pairs and actual negatives refers to unconnected airport-pairs:

- Specificity (True Negative Rate or TNR): percentage of actual negatives (unconnected airport-pairs) that are correctly identified as not being connected.
- Sensitivity (True Positive Rate or TPR): percentage of actual positives (connected airport-pairs) that are correctly identified as being connected.
- Precision: percentage of airport-pairs predicted connected that are actually positive over the total number of predicted airport-pairs being connected.

Considering the fact that the model has a relatively large error when predicting the minority class (i.e. which airport-pairs will be connected) as presented in Tables D.2 and D.3, an alternative methodology was used and it was decided that the problem would be split into two: link addition and link removal.

Predictions (0 = Un-connected; 1 = connected)	Actuals (0 = Un-connected; 1 = connected)									
	over-sampling		under-sampling		OU-sampling		ROSE		SMOTE	
Model 1										
	0	1	0	1	0	1	0	1	0	1
0	18088	763	18016	740	18094	762	17793	662	18841	1208
1	857	1272	929	1295	851	1273	1152	1373	104	828
Model 2										
0	17717	588	17636	569	17722	591	17555	550	18570	920
1	1228	1446	1308	1466	1222	1444	1390	1485	374	1116
Model 3										
0	17997	702	17930	688	17982	692	17796	622	18742	1063
1	948	1333	1015	1347	962	1343	1149	1412	203	972
Model 4										
0	17857	657	17812	655	17854	658	17796	642	18672	1004
1	1088	1378	1133	1380	1091	1377	1149	1393	273	1031
Model 5										
0	18451	257	18491	264	18453	250	18470	240	NA	NA
1	494	1778	454	1771	492	1785	475	1795	NA	NA

Table D.2: Confusion matrix obtained during validation. Average values across all validation years: 2008-2013.

	sampling technique	Sensitivity	Specificity	Precision
Model 1	over-sampling	0.62	0.95	0.60
	under-sampling	0.64	0.95	0.58
	OU-sampling	0.62	0.96	0.60
	ROSE	0.67	0.94	0.55
	SMOTE	0.41	0.99	0.89
Model 2	over-sampling	0.71	0.94	0.54
	under-sampling	0.72	0.93	0.53
	OU-sampling	0.71	0.94	0.54
	ROSE	0.73	0.93	0.52 1
	SMOTE	0.55	0.98	0.75
Model 3	over-sampling	0.65	0.95	0.59
	under-sampling	0.66	0.95	0.57
	OU-sampling	0.66	0.95	0.58
	ROSE	0.69	0.94	0.55
	SMOTE	0.48	0.99	0.83
Model 4	over-sampling	0.68	0.94	0.56
	under-sampling	0.68	0.94	0.55
	OU-sampling	0.68	0.94	0.56
	ROSE	0.68	0.94	0.55
	SMOTE	0.51	0.99	0.79
Model 5	over-sampling	0.87	0.97	0.78
	under-sampling	0.87	0.98	0.80
	OU-sampling	0.88	0.97	0.78
	ROSE	0.88	0.97	0.79

Table D.3: Performance metric for all 5 estimated models when validating the models over data from 2008 to 2013. Values shown are average across the 6 validation years.

D.2 Link additions and removals in 2011 and 2012

Figures D.1, D.2, D.3 and D.4 show a graphic representation of the links added and removed in 2011 and 2012 respectively. Figures below clearly show there is not an obvious trend on link addition and removal but rather a quite even distribution of those links changing connectivity across the entire network.

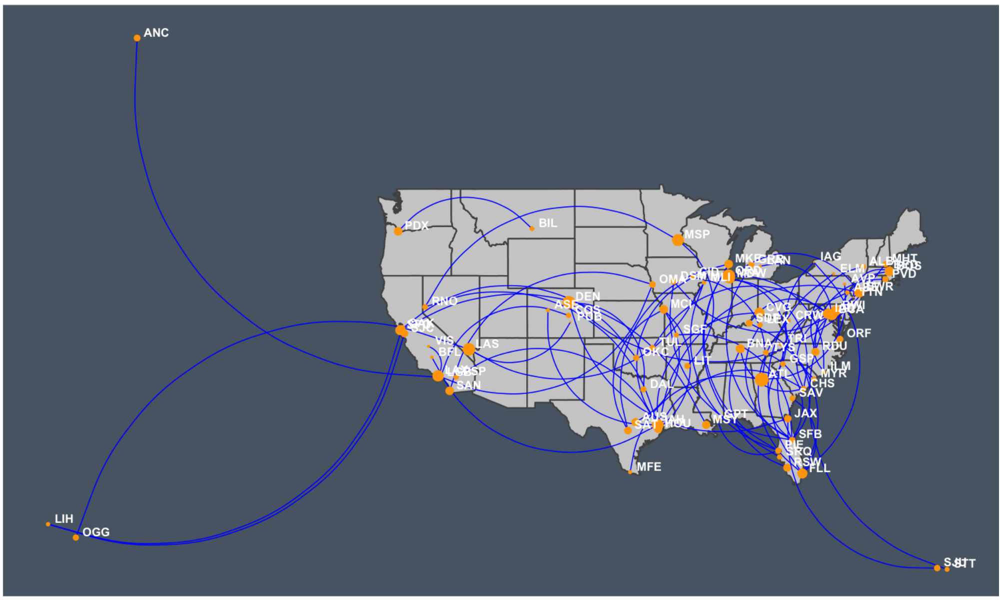

Figure D.1: Links that were added in 2011 (airport-pairs that were unconnected in 2010).

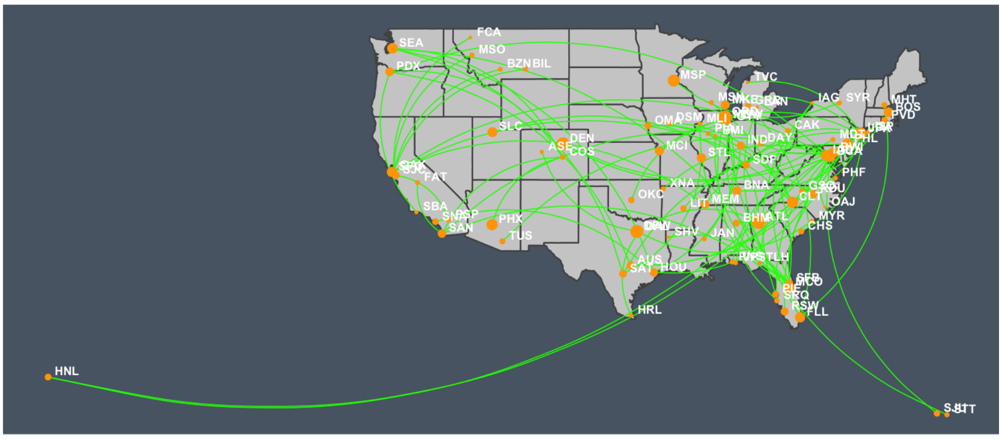

Figure D.2: Links that were added in 2012 (airport-pairs that were unconnected in 2011).

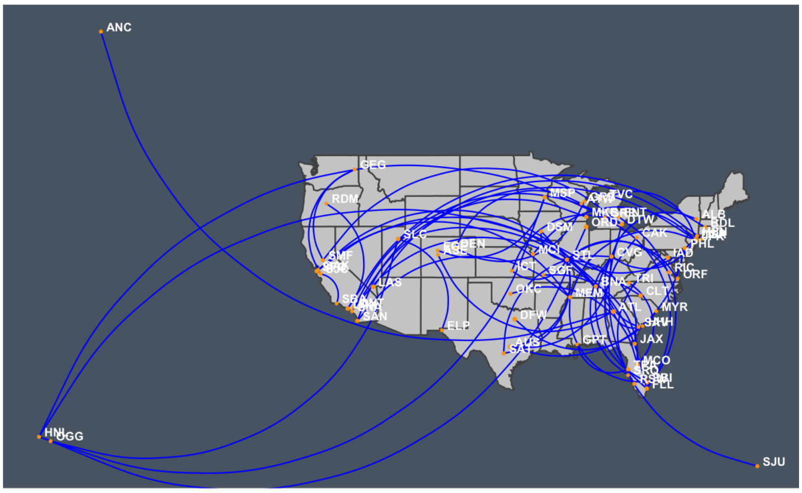

Figure D.3: Links that were removed in 2011 (airport-pairs that were connected in 2010).

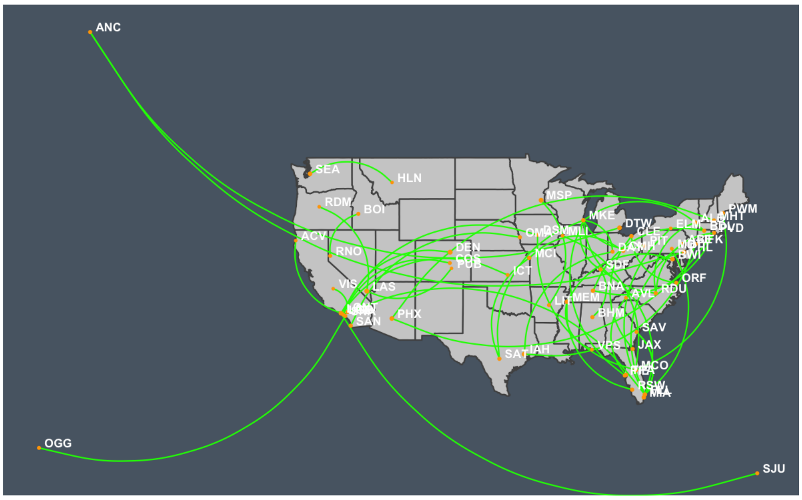

Figure D.4: Links that were removed in 2012 (airport-pairs that were connected in 2011).

Appendix E

Previous attempts - Itinerary choice model

This appendix presents a brief summary of previous attempts to model itinerary shares. This appendix is split in two sections, the first one looks at early modelling approaches investigated using multinomial logit (section E.1); the second presents some early investigation done to decide which was the best neural network architecture to model itinerary shares (section E.2).

E.1 Multinomial logit model

As discussed in Chapter 3 one of the concerns within the topic of itinerary choice modelling is journey fare, which has proven to be the most problematic of the explanatory variables and when considering aggregate values might lead to erroneous estimates. Earlier attempts to model itinerary shares did not use the 2-stage control function to correct this issue and therefore values of time obtained were too high and in some occasions negative. Below, some of the earlier approaches which also used Berkson-Theil approximation to estimate the model coefficients are presented.

Explanatory variables considered in initial attempts included journey time, journey fare and number of airlines operating on a given itinerary as continuous variables as well as the set of alternative specific constants (ASCs) included in the model presented in Chapter 3 -i.e. 26 ASCs referred to the itinerary level of service, including one for non-stop itineraries and 25 for one-top itineraries through one of the connecting airports included in this study. The three models considered at this stage are characterised by the following:

- Model 1: accounts for interactions between fare and the boolean variable indicating the level of service of the itinerary (i.e. non-stop or one-stop). Journey fare variable is included in the model as a log formulation and divided by 100;
- Model 2: similar to model 1 but without any data transformation;
- Model 3: in this model journey fare is not split between non-stop and one-stop itineraries, and there is added variables to account for whether true-origin and true-destination journey airports are hub or not.

Note that in Model 3, variables regarding hub information are only present for those city-pairs in which at least one of the available option is a non-stop itinerary. In all the models, number of passengers for a given itinerary has been used as weight during the estimation process with the exception of model 3 in which the ratio between the itinerary's number of passenger and the number of passengers from the reference alternative has been used. Table E.1 shows the model specification for each of these 3 models; whereas Table E.1 presents the estimation model results obtained.

	Coefficients	Model	Explanatory variables
Constant		ASC_{NS}	1 x $\mathrm{HUB_i}$=="Non-stop"
		ASC_{ATL}	1 x $\mathrm{HUB_i}$=="ATL"
		$\vdots$	$\vdots$
		ASC_k	1 x $\mathrm{HUB_i}$=="k"
Journey fare	Model 1	β_{fare}^{NS}	$ln(\frac{fare_i}{100})$ x non-stop$_i$
		β_{fare}^{OS}	$ln(\frac{fare_i}{100})$ x one-stop$_i$
	Model 2	β_{fare}^{NS}	fare$_i$ x non-stop$_i$
		β_{fare}^{OS}	fare$_i$ x one-stop$_i$
	Model 3	β_{fare}	$\frac{fare_i}{100}$
Journey time	All models	β_{time}	time$_i$
Num. of airlines	All models	$\beta_{airlines}$	num_airlines$_i$
Hub information	Model 3	β_{hub1}	NS_Market x hub1$_i$
		β_{nohub}	NS_Market x nohub$_i$
		$\beta_{hub2hub}$	NS_Market x hub2hub$_i$

Table E.1: Specification table of the utility function for models estimated using Berkson-Theil method.

Validation process is as explained in Chapter 3, input data between 2008 and 2013 is used to generate predictions of number of passengers across all available itineraries. To evaluate the models predicting power average adjusted R^2 square across all the validation years is calculated for each of the three models (Table E.3). Figures E.1, E.2 and E.3 below show the comparison between observed and predicted number of passengers across the validation years for all 3 models.

Parameter	Model 1	Model 2	Model 3
β_{fare}			-0.112***
β_{fare}^{NS}	-0.239***	- 0.001***	
β_{fare}^{OS}	-0.305***	-0.0014***	
β_{time}	-0.595***	-0.598 ***	-0.328***
$\beta_{airlines}$	0.150 ***	0.149***	1.3645 ***
β_{hub1}			-0.050***
β_{nohub}			-0.0775***
$\beta_{hub2hub}$			-0.0563*
Adj R-squared	0.7	0.699	0.57

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table E.2: Table results for Berkson-Theil.

	Model 1	Model 2	Model 3
Adjusted R^2	0.858	0.858	0.818

Table E.3: Comparison of predictive powers between the 6 models estimated using Berkson-Theil method.

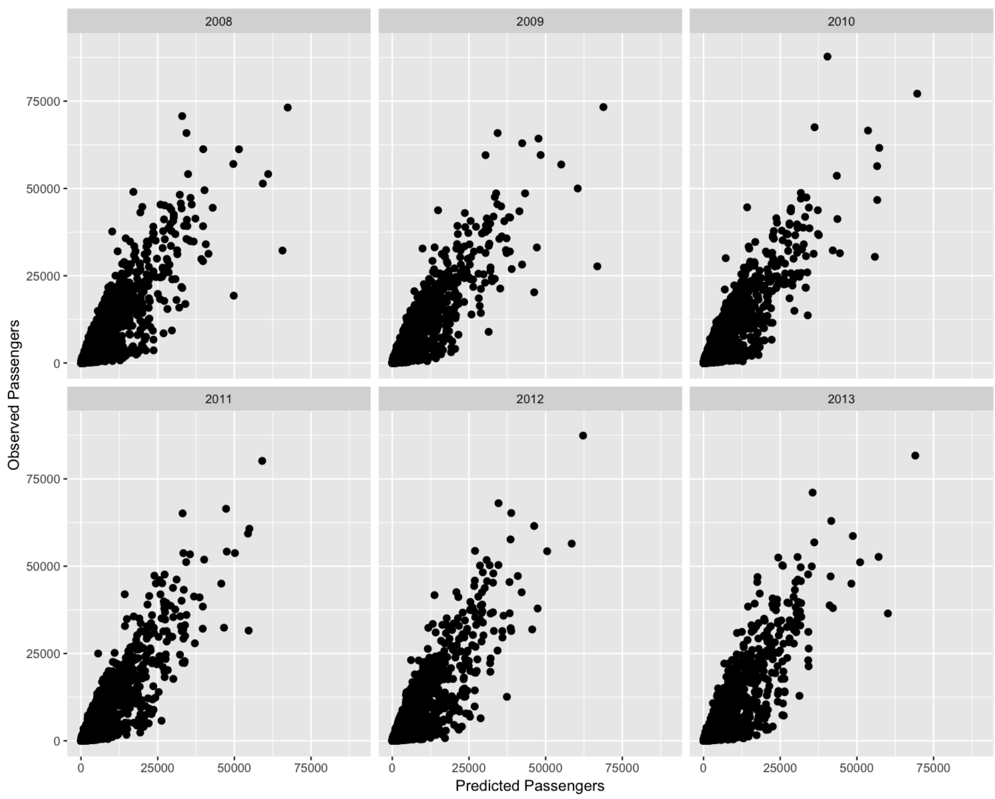

Figure E.1: Observed against predicted number of passengers using model 1 throughout the validation years.

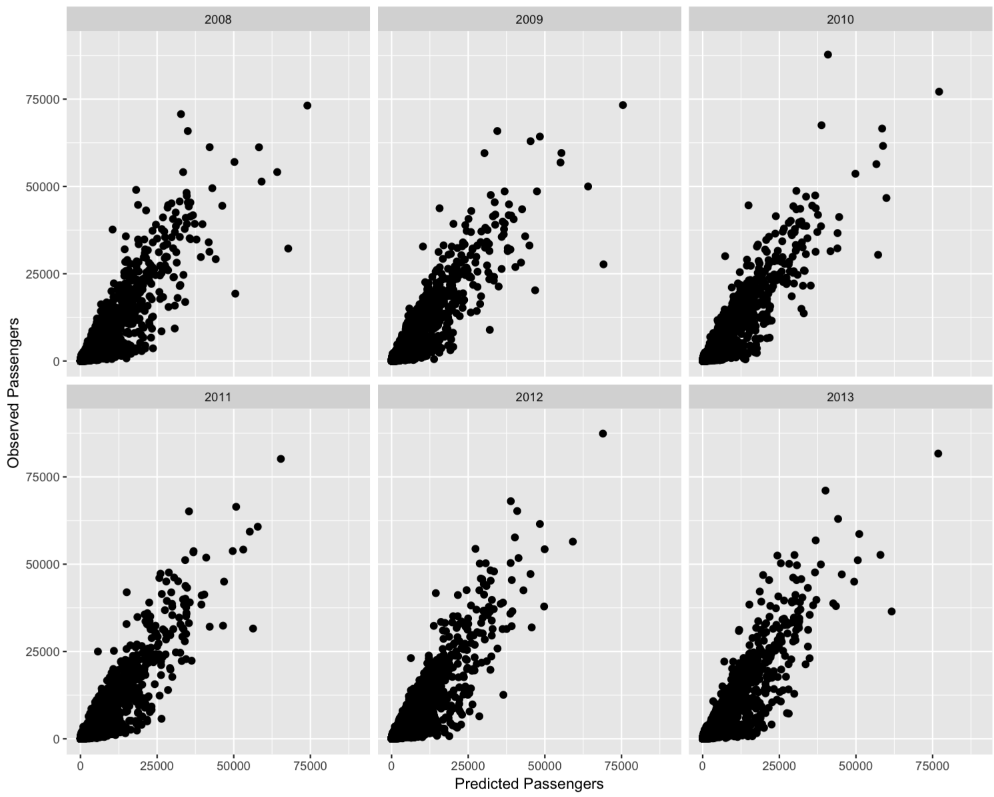

Figure E.2: Observed against predicted number of passengers using model 2 throughout the validation years.

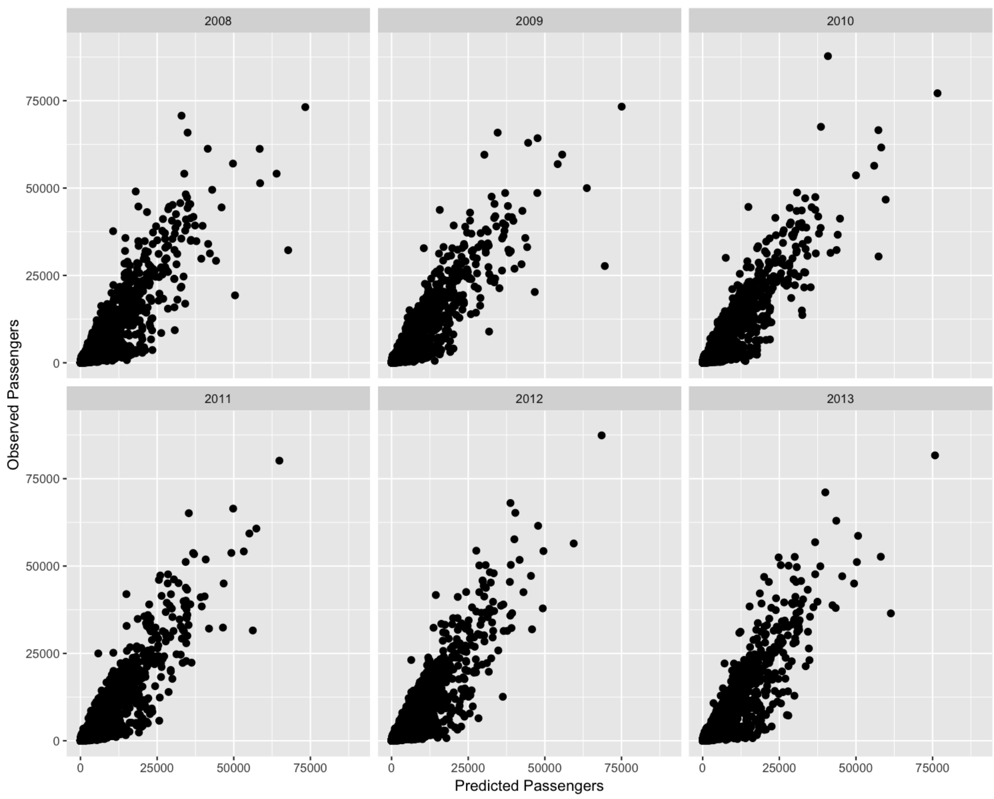

Figure E.3: Observed against predicted number of passengers using model 3 throughout the validation years.

Values of time for all three models are computed and presented in Table E.4[1]. Those result in much higher values of time compared to those in the literature (Hsiao and Hansen, 2011; Atasoy and Bierlaire, 2012), suggesting that fare endogeneity might be an issue. For illustration purposes, Figure E.4 presents the distribution of value of time at the itinerary level (i.e. non-stop and one-stop) for model 2 as an example; which clearly shows those high values of time across all level of service and no clear difference across the connecting hubs with some exceptions (e.g. DTW, IAH, BWO).

	Model 1	Model 2	Model 3
VOT_{NS}	\$425/h	\$569/h	
VOT_{OS}	\$388/h	\$407/h	
VOT			\$296/h

Table E.4: Comparison of Value of Time. Note the average VOT values for model 1 are presented in this table.

[1]Note that for models 1, VOT is calculated by multiplying the itinerary fare, hence VOT values vary across different itineraries. In this case the average fare for non-stop and one-stop itineraries has been used to calculate average values of time.

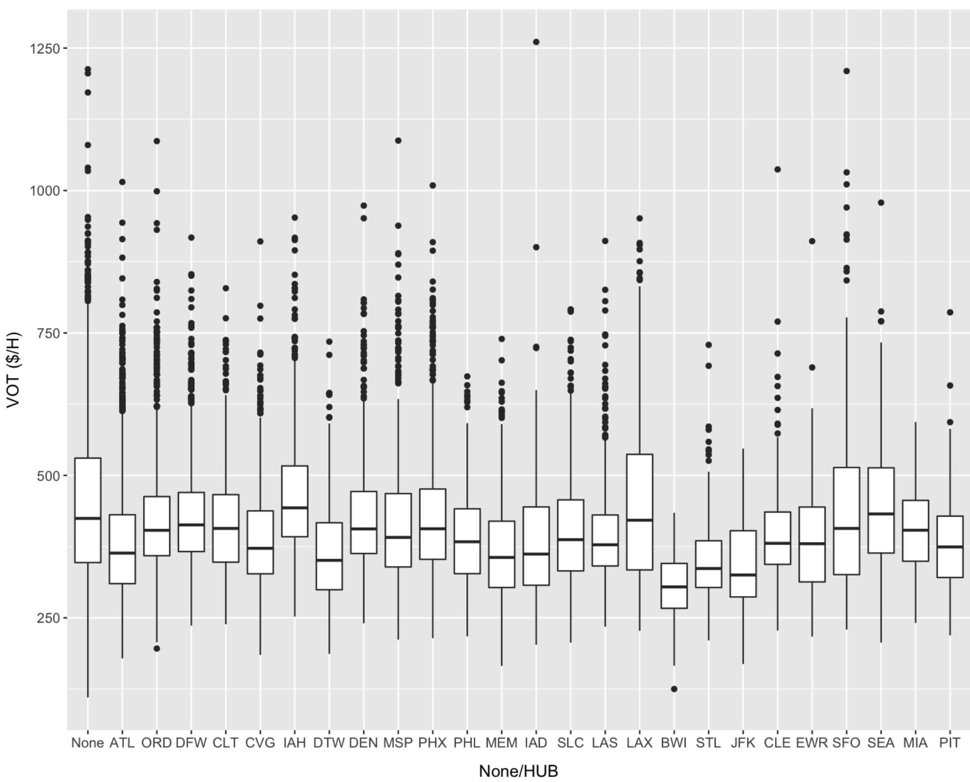

Figure E.4: VOT calculation from model 1.

E.2 Neural network model

The itinerary choice model using neural network presented in Chapter 3 is estimated using the Levenberg-Marquart algorithm (LMA) during training; however previous exploratory analysis also included the use of backpropagation and backpropagation with momentum as alternative training algorithm. Eventually, Levenberg-Marquart algorithm was chosen to train the model due to better performance in terms of lower training time and number of epochs during training as well as lower mean square error obtained when applying the model to a validation set. Figures E.5, E.6 and E.7 show the three metrics used to evaluate the three training algorithms across the range of neural network (NN) architectures considered in this study. The three training algorithms seem to perform fairly similarly, with MSE values under 0.03 for most of the NN topologies. However, Levenberg-Marquart algorithm seems to perform consistently better than the other two algorithms, specially in terms of training time, with few exceptions - e.g. such as the NN architectures 10/5, 25/5 or 30/5.

[2]Values from applying Levenberg-Marquardt are not visible due to being much smaller in comparison to the rest of algorithms shown on the chart.

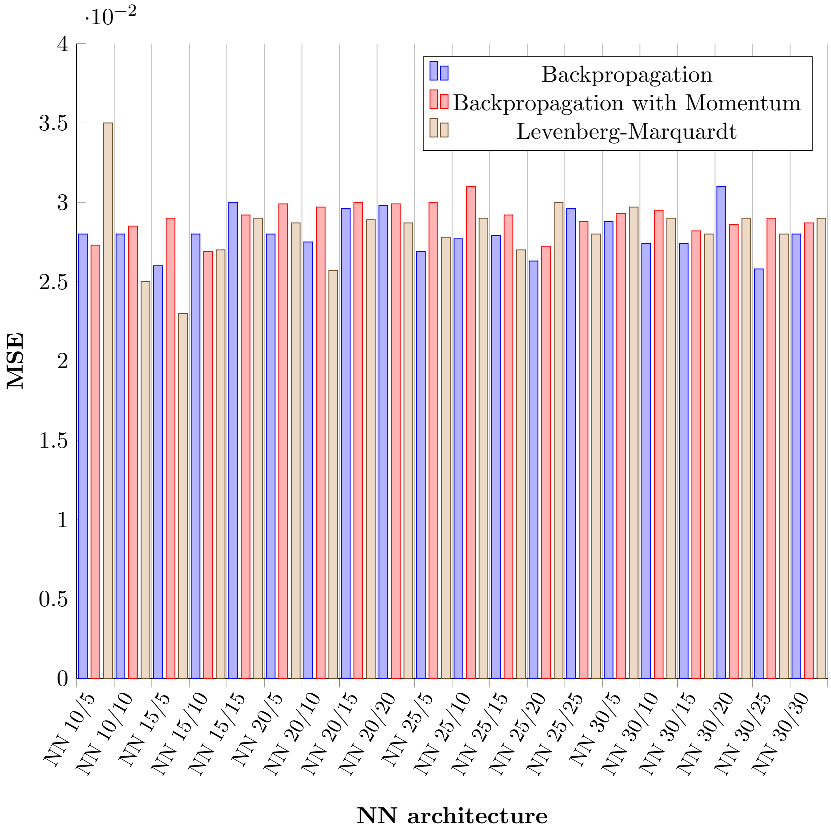

Figure E.5: Validation performance comparison (MSE) between the best performance results for several NN architectures with 3 different training algorithms.

Considering the choice of Levenberg-Marquart algorithm to use during training, an evaluation was also performed to check which activation function to use for each of the layers of the neural network architecture . The several combinations of activation functions chosen to test are presented in Table E.5. The same metrics (i.e. mean square error, number of epochs and training time) were used to evaluate their performance presented in Figures E.8, E.9 and E.10 respectively.

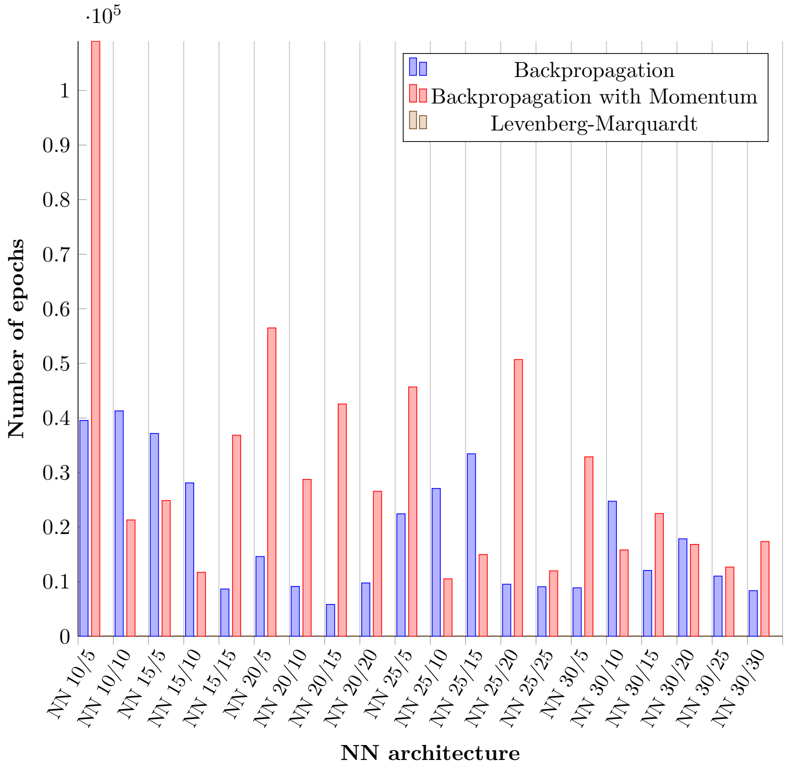

Figure E.6: Number of epochs comparison between the best performance results for several NN architectures with 3 different training algorithms.[2]

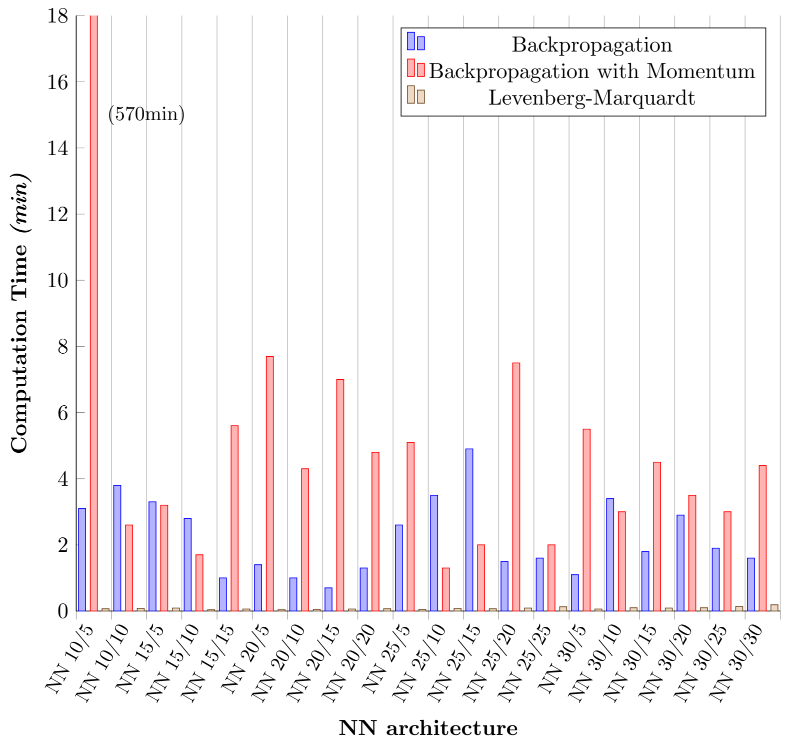

Figure E.7: Time comparison between the best performance results for several NN architectures with 3 different training algorithms.

Version	Act. Fcn 1st layer	Act. Fcn 2nd layer	Act. Fcn Output layer
v1	Hyperbolic Tangent	Hyperbolic Tangent	Linear
v2	Sigmoid	Sigmoid	Linear
v3	Sigmoid	Sigmoid	Sigmoid
v4	Hyperbolic Tangent	Hyperbolic Tangent	Sigmoid

Table E.5: Combination of activation functions considered in this research.

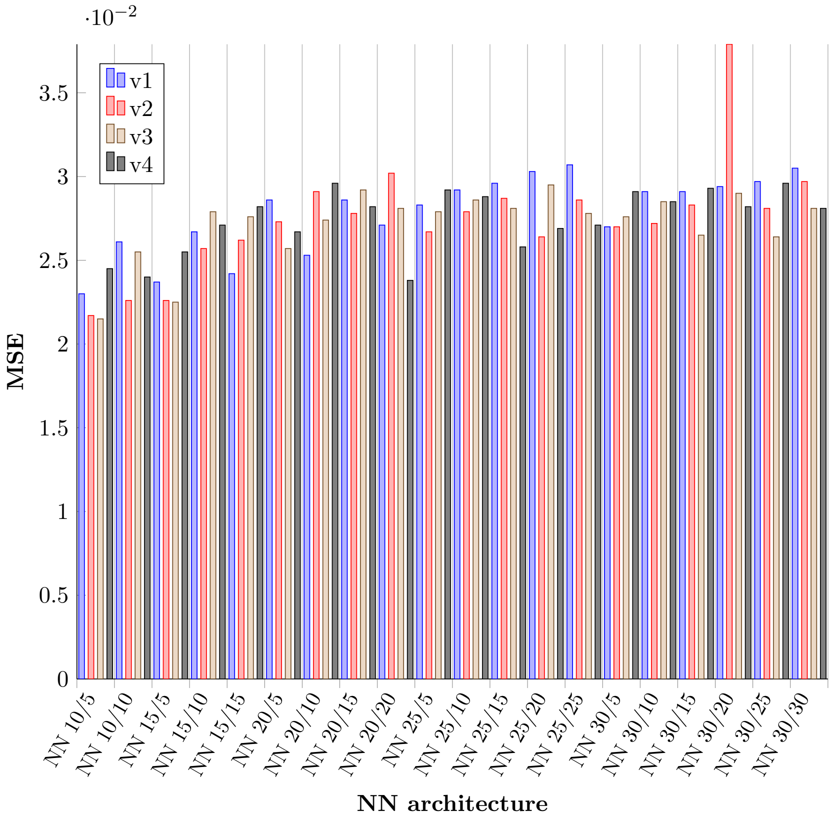

Figure E.8: Validation performance comparison (MSE) between the best performance run considering four activation function combinations from table E.5.

'As the performance of all versions is consistently similar, *v4* -i.e. hyperbolic tangent function for hidden layers and sigmoid function for the output layer- is taken to build the air itinerary choice model. With output values referring to the market share of an itinerary serving a given city-pair, and hence ranging between 0 and 1, the choice of sigmoid activation function in the output layer implies that no data transformation is needed. Comparing *v4* against *v3* -i.e. sigmoid function for hidden and output layer- shows that *v4* performs slightly better when looking at number of epochs and computing time.

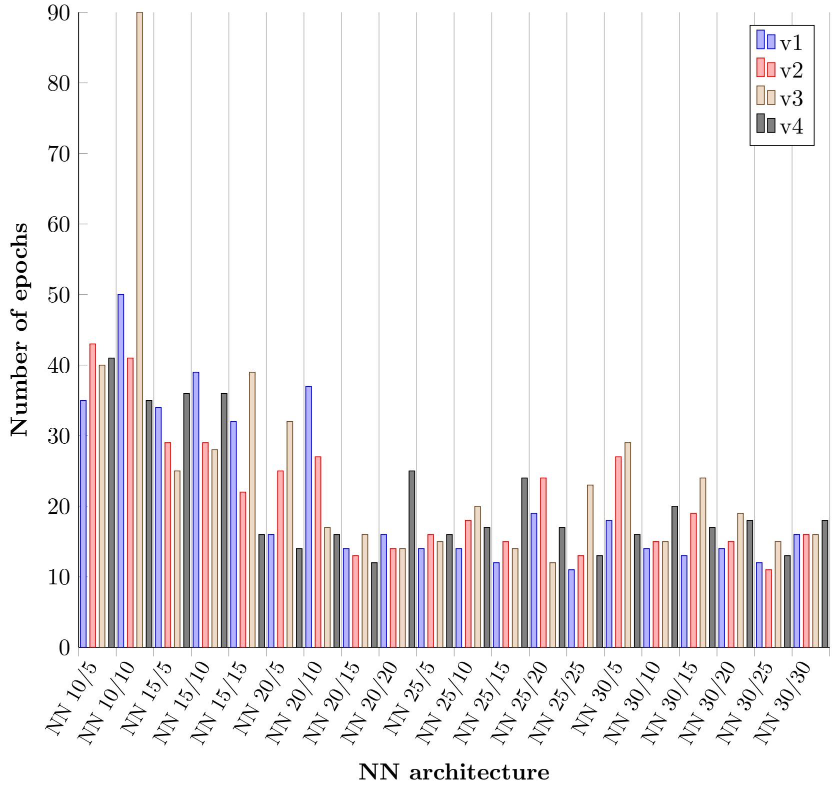

Figure E.9: Number of eqpochs comparison between the best performance run considering four activation function combinations from table E.5.

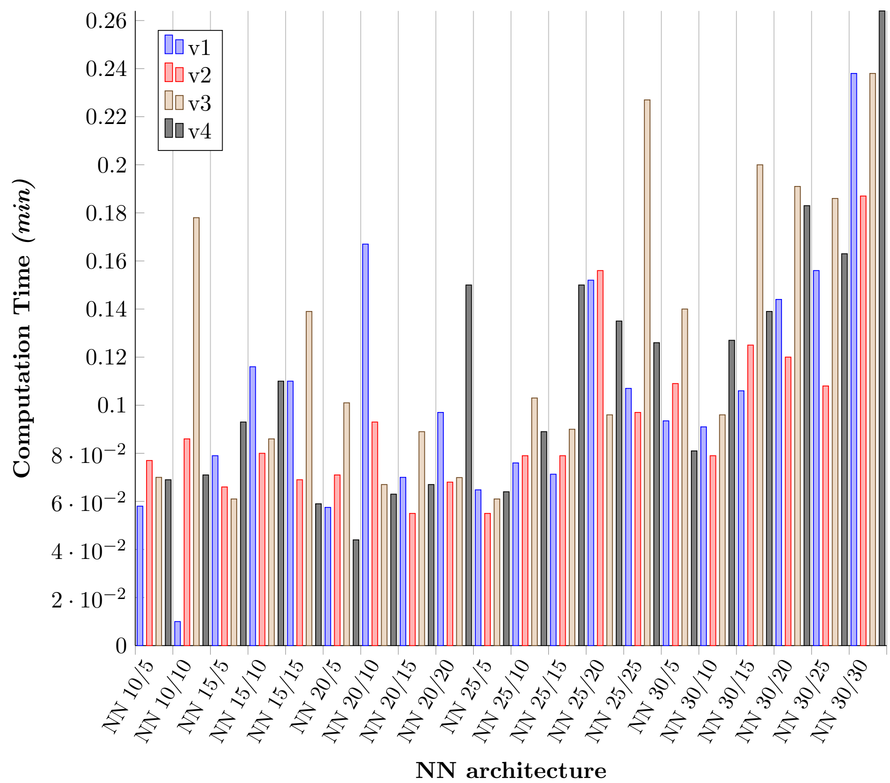

Figure E.10: Time comparison between the best performance run considering four activation function combinations from table E.5.

Bibliography

2017, I. C. I. I. (2007). Estimating air travel demand elasticities. final report.

Adler, N. (2001). Competition in a deregulated air transportation market. *European Journal of Operational Research*, 129(2):337–345.

Airbus (2017). Global Market Forecast. Growing Horizons. 2017-2036.

Airports Commission, G. U. (2013). Airports Comision: Interim Report. http://www.gov.uk/government/uploads/system/uploads/attachment_data/file/271231/airports-commission-interim-report.pdf. Accessed: 2014-08-03.

Alexander, F. J. (2013). Machine learning. *Computing in Science & Engineering*, 15(5):9–11.

Amemiya, T. (1978). The estimation of a simultaneous equation generalized probit model. *Econometrica: Journal of the Econometric Society*, pages 1193–1205.

Arvis, J.-F. and Shepherd, B. (2011). The air connectivity index: measuring integration in the global air transport network. Technical report, The World Bank, Poverty Reduction and Economic Management Network. International Trade Department.

ATAG (2014). When the system stops working. Air Transport Action Group.

Atasoy, B. and Bierlaire, M. (2012). An air itinerary choice model based on a mixed rp/sp dataset. Technical report, Transport and Mobility Laboratory. Ecole Polytechnique Fédérale de Lausanne.

BEA, B. (2014). Ca1-3 personal income summary: 2003 to 2007. online database, http://bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=2#reqid=70&step=25&isuri=1&7022=20&7023=7&7024=non-industry&7001=720&7029=20&7090=70.

BEA, B. (2019). National income and product accounts: Table 1.1.6 real gross domestic product, adjusted for inflation using 2012 dollars. online database, https://apps.bea.gov/itable/index.cfm.

Ben-Akiva, M. E., Lerman, S. R., and Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press.

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics, London*, 22(1):23–36.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, 48(263):565–599.

Boeing (2017). Current Market Outlook. 2017-2036. http://www.boeing.com/resources/boeingdotcom/commercial/market/current-market-outlook-2017/assets/downloads/2017-cmo-6-19.pdf.

Bonnefoy, P. A. and Hansman, R. J. (2007). Potential impacts of very light jets on the national airspace system. *Journal of Aircraft*, 44(4):1318–1326.

BTS (2017). 2016 Annual and December U.S. Airline Traffic Data. https://www.bts.gov/newsroom/2016-annual-and-december-us-airline-traffic-data. Accessed: 2017-10-28.

BTS-RITA, Bureau of Transportation Statistics - Research and Innovative Technology Administration (2014). Origin and destination survey: Db1bmarket for 2003 to 2007,2010. online database, http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=247&DB_Short_Name=Origin%20and%20Destination%20Survey.

BTS-RITA, Bureau of Transportation Statistics. Research and Innovative Technology Administration (2014). T-100 domestic segment (u.s. carriers): 2003 to 2007,2010. online database, http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=259&DB_Short_Name=Air%20Carriers.

BTS-RITA, Bureau of Transportation Statistics. Research and Innovative Technology Administration (2019). Airline fuel cost and consumption (u.s, carriers - scheduled) (january 2000 - december 2018). online database, https://www.transtats.bts.gov/fuel.asp.

Bureau of Labour Statistics, US Department of Labour (2014a). Annual average indexes 2007 (tables 1a-23a). online database, http://www.bls.gov/cpi/cpi_dr.htm#2007.

Bureau of Labour Statistics, US Department of Labour (2014b). Employer costs for employee compensation. online database, https://www.bls.gov/bls/newsrels.htm.

Bureau of Labour Statistics, US Department of Labour (2014c). National occupational employment and wage estimates. online database, https://www.bls.gov/oes/current/oes_nat.htm#00-0000.

Bureau of Labour Statistics, US Department of Labour (2018). Consumer price index (cpi): Definition. online database, https://www.bls.gov/cpi/.

Busquets, J. G., Alonso, E., and Evans, A. D. (2018). Air itinerary shares estimation using multinomial logit models. *Transportation Planning and Technology*, 41(1):3–16.

Carrier, E. (2008). *Modeling the choice of an airline itinerary and fare product using booking and seat availability data*. PhD thesis, Massachusetts Institute of Technology, Department of Civil and Environmental Engineering.

Carrier, E. and Weatherford, L. (2014). Estimating a model of airline passenger choice using grouped booking data and least squares regression. *Journal of Revenue and Pricing Management*, 13(5):347–353.

Cheng, T., Cui, D., and Cheng, P. (2003). Data mining for air traffic flow forecasting: a hybrid model of neural network and statistical analysis. In *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*, volume 1, pages 211–215. IEEE.

Cheung, D. P. and Gunes, M. H. (2012). A complex network analysis of the united states air transportation. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 699–701. IEEE Computer Society.

Chi, L. and Cai, X. (2004). Structural changes caused by error and attack tolerance in us airport network. *International Journal of Modern Physics B*, 18(17n19):2394–2400.

Coldren, G. M. and Koppelman, F. S. (2005). Modeling the competition among air-travel itinerary shares: Gev model development. *Transportation Research Part A: Policy and Practice*, 39(4):345–365.

Coldren, G. M., Koppelman, F. S., Kasturirangan, K., and Mukherjee, A. (2003). Modeling aggregate air-travel itinerary shares: logit model development at a major us airline. *Journal of Air Transport Management*, 9(6):361–369.

Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020.

Das, S., Matthews, B. L., Srivastava, A. N., and Oza, N. C. (2010). Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 47–56. ACM.

d'Avila Garcez, A. (2014). Lecture notes in Neural and Evolutionary Computing. artificial neural networls and hybrid systems. Computer Science Department. City, University of London.

DECC (2015). Fossil fuel price projections: 2015. online database, https://www.gov.uk/government/publications/fossil-fuel-price-projections-2015.

DeLaurentis, D., Han, E.-P., and Kotegawa, T. (2008). Network-theoretic approach for analyzing connectivity in air transportation networks. *Journal of Aircraft*, 45(5):1669.

Dennis, N. P. (2002). Long-term route traffic forecasts and flight schedule pattern for a medium-sized european airport. *Jouranl of Air Transport Management*, 8(5):313–324.

DfT (2013). Department for Transport. UK Aviation Forecasts. https://www.gov.uk/government/publications/uk-aviation-forecasts-2013. Accessed: 2014-08-03.

DfT (2017). UK Aviation forecasts. Moving Britain Aheah. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/653821/uk-aviation-forecasts-2017.pdf. Accessed: 2017-11-21.

Dray, L., Evans, A., Reynolds, T., Rogers, H., Schäfer, A., and Vera-Morales, M. (2009). Air transport within an emissions trading regime: a network-based analysis of the united states and india. In *TRB 88th Annual Meeting, Washington DC*, pages 11–15.

Dray, L., Evans, A., Reynolds, T., Schäfer, A. W., Vera-Morales, M., and Bosbach, W. (2014). Airline fleet replacement funded by a carbon tax: an integrated assessment. *Transport Policy*, 34:75–84.

Dray, L. M., Krammer, P., Doyme, K., Wang, B., Al Zayat, K., O'Sullivan, A., and Schäfer, A. W. (2019). Aim2015: Validation and initial results from an open-source aviation systems model. *Transport Policy*, 79:93–102.

Eurocontrol (2013). Challenges of growth 2013: Task 4: European air traffic in 2035.

Eurocontrol (2017). Eurocontrol Seven-year forecast September 2017. Flight Movements and Service Units 2017-2023.

Evans, A. (2010). *Simulating airline operational responses to environmental constraints*. PhD thesis, University of Cambridge.

Evans, A. D. and Schäfer, A. W. (2014). Simulating airline operational responses to airport capacity constrains. *Transport policy*, 34:5–13.

FAA (2007). Terminal area forecast summary. fiscal years: 2008-2025. online source, http://libraryonline.erau.edu/online-full-text/books-online/TAF/TAF2008-2025.pdf.

FAA (2014). Aviation system performance metrics (aspm) manuals. online database, http://aspmhelp.faa.gov/index.php/ASPM_Manuals#User_Manuals.

FAA (2016). Terminal area forecast summary. fiscal years 2016-2045. https://taf.faa.gov/Downloads/TAFSummaryFY2016-2045.pdf. Accessed: 2017-11-21.

FAA (2018). Faa aerospace forecasts. economic tables (tables 1-4). online source, https://www.faa.gov/data_research/aviation/aerospace_forecasts/.

FAA (2019). Faa aerospace forecast. fiscal years: 2019-2039. online source, https://www.faa.gov/data_research/aviation/aerospace_forecasts/media/FY2019-39_FAA_Aerospace_Forecast.pdf.

Fleurquin, P., Ramasco, J. J., and Eguiluz, V. M. (2013). Systemic delay propagation in the us airport network. *Scientific reports*, 3:1159.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47.

Ghobrial, A. and Soliman, S. (1992). An assessment of some factors influencing the competitive strategies of airlines in domestic markets. *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, pages 247–258.

Grammig, J., Hujer, R., and Scheidler, M. (2005). Discrete choice modelling in airline network management. *Journal of Applied Econometrics*, 20(4):467–486.

Grosche, T. and Rothlauf, F. (2007). Air travel itinerary market share estimation. Technical report, University of Mannheim.

Guevara, C. A. and Ben-Akiva, M. (2006). Endogeneity in residential location choice models. *Transportation Research Record*, 1977(1):60–66.

Guevara, C. A. and Polanco, D. (2013). Correcting for endogeneity without instruments in discrete choice models: the multiple indicator solution. In *Proceedings of the Third International Choice Modelling Conference*.

Guimera, R., Mossa, S., Turtschi, A., and Amaral, L. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799.

Guimera Busquets, J. (2015). Using log-log models to improve on existing air traffic forecasting methodologies. In *University Transport Study Group*.

Guimera Busquets, J., Alonso, E., and Evans, A. (2015). Application of data mining in air traffic forecasting. In *AIAA Aviation Technology, Integration and Operations Conference*.

Hansen, M. (1995). Positive feedback model of multiple-airport systems. *Journal of transportation engineering*, 121(6):453–460.

Hansen, M. and Liu, Y. (2015). Airline competition and market frequency: A comparison of the s-curve and schedule delay models. *Transportation Research Part B: Methodological*, 78:301–317.

Heaton, J. (2015). Artificial intelligence for humans, volume 3: Deep learning and neural networks, heaton research. *Inc.: St. Louis, MO, USA*.

Hess, S. and Polak, J. W. (2005). Mixed logit modelling of airport choice in multi-airport regions. *Journal of Air Transport Management*, 11(2):59–68.

Hollnagel, E. (2013). *Resilience engineering in practice: A guidebook.* Ashgate Publishing, Ltd.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Hsiao, C.-Y. and Hansen, M. (2011). A passenger demand model for air transportation in a hub-and-spoke network. *Transportation Research Part E: Logistics and Transportation Review*, 47(6):1112–1125.

ICAO (2016). ICAO Long-Term Traffic Forecasts. Passenger and Cargo. https://www.icao.int/safety/ngap/NGAP8%20Presentations/ICAO-Long-Term-Traffic-Forecasts-July-2016.pdf#search=long%2Dterm. Accessed: 2017-11-12.

Iverson, D. L. (2004). Inductive system health monitoring. In *The 2004 International Conferenc on Artificial Intelligence.*

Kamiyama, N., Iijima, N., Taguchi, A., Mitsui, H., Yoshida, Y., and Sone, M. (1992). Tuning of learning rate and momentum on back-propagation. In *Singapore ICCS/ISITA'92.'Communications on the Move'*, pages 528–532. IEEE.

Kotegawa, T. (2012). *Analyzing the evolutionary mechanisms of the air transportation system-of-systems using Network Theory and machine learning algorithms.* PhD thesis, Purdue University.

Kotegawa, T., DeLaurentis, D., and Harden, G. (2010). Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. In *27th International Congress of the Aeronautical Sciences.* ICAS.

Lacasa, L., Cea, M., and Zanin, M. (2009). Jamming transition in air transportation networks. *Physica A: Statistical Mechanics and its Applications*, 388(18):3948–3954.

Lawson, D. and Castillo, W. (2012). Predicting flight delays. Technical report, Technical report, Computer Science Department, CS 229, Stanford University, Stanford, CA.

Lewis, T. G. (2009). *Network science: theory and practice.* John Wiley & Sons, Hoboken, New Jersey.

Li, W., Kamargianni, M., Krammer, P., Dray, L. M., and Schäfer, A. (2017). Assessing the performance of berkson-theil method on multiple choice sets and aggregated choice data. Semantic Scholar. (Online Source).

Lunardon, N., Menardi, G., and Torelli, N. (2014). Rose: A package for binary imbalanced learning. *R journal*, 6(1).

Lurkin, V., Garrow, L. A., Higgins, M. J., Newman, J. P., and Schyns, M. (2017). Accounting for price endogeneity in airline itinerary choice models: An application to continental us markets. *Transportation Research Part A: Policy and Practice*, 100:228–246.

Nam, K. and Schaefer, T. (1995). Forecasting international airline passenger traffic using neural networks. *Logistics and Transportation Review*, 31(3):239.

Ng, A. Y. (2013). Machine learning course. Online course MOOC (Massive Open Online Courses); Coursera by Standford University. Online lecture.

Ortuzar, J. d. and Willumsen, L. G. (2001). *Modelling transport.* Wiley, Chichester, 3 edition.

Pathomsiri, S. and Haghani, A. (2005). Taste variations in airport choice models. *Transportation research record*, 1915(1):27–35.

Pearce, B. (2007). Investing in air transport connectivity to boost national productivity and economic growth. In *World Economic Forum, The Travel & Tourism Competitiveness Report.*

Pindyck, R. S. and Rubinfeld, D. L. (1998). *Econometric models and economic forecasts*, volume 4. Irwin/McGraw-Hill Boston.

Proussaloglou, K. and Koppelman, F. (1995). Air carrier demand: an analysis of market share determinants. *Transportation*, 22(4):371–388.

Reynolds-Feighan, A. and McLay, P. (2006). Accessibility and attractiveness of european airports: A simple small community perspective. *Journal of Air Transport Management*, 12(6):313–323.

SABRE database (2017). Global scheduled seats.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415.

Schäfer, A. (2007). Long-term trends in global passenger mobility. In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2006 Symposium*, page 85. National Academies Press.

Schäfer, A. and Dray, L. (2015). Aviation integrated modelling (AIM) project. http://www.aimproject.aero/. Accessed: 2014-08-04.

Schmidheiny, K. (2018). Lecture notes in Instrumental variables. Short Guide to Microeconometrics. Universität Basel.

Smalley, E. (2012). TNASA Applies Text Analytics to Airline Safety. Data Informed. http://data-informed.com/nasa-applies-text-analytics-to-airline-safety/. Accessed: 2013-10-11.

Srivastava, A. (2011). NASA Chat: Data Mining Digs Up Clues to Safer Flights. NASA Ames Research Center. NASA chat. http://www.nasa.gov/connect/chat/data_mining_chat.html#.U-owc_ldWSo. Accessed: 2013-10-23.

Staiger, D. O. and Stock, J. H. (1994). Instrumental variables regression with weak instruments.

Swan, W. (2008). Forecasting air travel with open skies. In *joint EWCKOTI Conference*.

Taneja, N. K. (1971). A model for forecasting future air travel demand on the north atlantic. Technical report, Cambridge, Mass. Massachusetts Institute of Technology, Flight Transportation Laboratory,[1971].

Teyssier, N. (2010). Aviation data. *Proceedings of the 37th Session of the ICAO Assembly.*

Teyssier, N. (2012). *How the consumer confidence index could increase air travel demand forecast accuracy?* PhD thesis, Air Transport Group, School of engineering. Cranfield University.

Theil, H. (1969). A multinomial extension of the linear logit model. *International economic review*, 10(3):251–259.

US Census Bureau (2014a). Annual estimates of the population of metropolitan and micropolitan statistical areas: April 1 2000 to july 1 2009. vintage 2009: Metropolitan and micropolitan statistical areas tables (cbsa-est2009-01). data retrieved from the US Census Bureau website, http://www.census.gov/popest/data/historical/2000s/vintage_2009/metro.html.

US Census Bureau (2014b). Households by total money income, race, and hisponic origin of householder: 1967 to 2014 (table a-1). data retrieved from the US Census Bureau website, https://www.census.gov/data/tables/2015/demo/income-poverty/p60-252.html.

U.S. DoT (2014). Revised departmental guidance on valuation of travel time in economic analysis. u.s. department of transportation. online source, https://www.transportation.gov/office-policy/transportation-policy/revised-departmental-guidance-valuation-travel-time-economic.

Viken, J., Dollyhigh, S., Smith, J., Trani, A., Baik, H., Hinze, N., and Ashiabor, S. (2006). Utilizing traveler demand modeling to predict future commercial flight schedules in the nas. In *Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Converence*, Portsmouth, Virginia. American Institute of Aeronautics and Astronautics.

Warburg, V., Bhat, C., and Adler, T. (2006). Modeling demographic and unobserved heterogeneity in air passengers' sensitivity to service attributes in itinerary choice. *Transportation Research Record: Journal of the Transportation Research Board*, 1951:7–16.

Weidner, T. J. (1996). Hubbing in us air transportation system: Economic approach. *Transportation research record*, 1562(1):28–37.

Wickham, R. R. (1995). *Evaluation of forecasting techniques for short-term demand of air transportation*. PhD thesis, Massachusetts Institute of Technology.

Wilkinson, S. M., Dunn, S., and Ma, S. (2012). The vulnerability of the european air traffic network to spatial hazards. *Natural hazards*, 60(3):1027–1036.

Windle, R. and Dresner, M. (1995). Airport choice in multiple-airport regions. *Journal of Transportation Engineering*, 121(4):332–337.

Wittman, M. D. and Swelbar, W. S. (2013a). Modeling changes in connectivity at us airports: A small community perspective. Technical report, MIT International Center for Air Transportation.

Wittman, M. D. and Swelbar, W. S. (2013b). Trends and market forces shaping small community air service in the united states. Technical report, MIT International Center for Air Transportation.

Zanin, M. and Lillo, F. (2013). Modelling the air transport with complex networks: A short review. *The European Physical Journal Special Topics*, 215(1):5–21.

Zhao, J., Agusdinata, D., and DeLaurentis, D. A. (2009). System dynamic fleet and network forecasting with technology and emission goals. In *9th AIAA Aviation Technology, Integration, and Operations Conference (ATIO) and AIAA/AAAF Aircraft Noise and Emissions Reduction Symposium (ANERS).*