Forecasting Air Passenger Demand between Settlements Worldwide Based on Socio-Economic Scenarios

Vom Promotionsausschuss der
Technischen Universität Hamburg-Harburg

zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

von
Dipl.-Ing. Ivan Terekhov

aus
Moskau

2017

Übersicht der Gutachter

1. Gutachter	Prof. Dr.-Ing. Volker Gollnick
2. Gutachter	Dr. Antony D. Evans
Vorsitzender des Prüfungsausschusses	Prof. Dr. rer. pol. Kathrin Fischer
Tag der mündlichen Prüfung:	18. Juli 2017

Acknowledgements

This thesis is the result of five years of research as a doctoral candidate at DLR Air Transportation Systems. I would like to take this opportunity to express my sincere gratitude to the many interesting and inspiring people I had the pleasure of meeting and working with during this time. All of you encouraged and supported me throughout my research work in one way or another and have helped to make this thesis what it is today.

I would first like to thank my first supervisor Prof. Volker Gollnick. Being supported by someone with such a wealth of expertise and knowledge was a rewarding experience I am extremely grateful for. Secondly, I would like to express my thanks to my second supervisor, Dr. Antony D. Evans, whose ongoing guidance taught me a great deal and whose encouragement and insights enabled me to stay on track and significantly improve my work. I also would like to thank Prof. Joachim Szodruch and Dr. Richard Degenhardt, who accepted me into the PhD program in the first place.

Many thanks to all of my colleagues at DLR Air Transportation Systems for their collaboration, support and friendship over the years, including Robin Ghosh, Malte Niklaß, Thomas Schilling, Kai Wicke, Marco Weiss, Katrin Kölker, Niclas Dzikus and Jan Berling. I really appreciate the time and effort they dedicated to me and my research.

My heartfelt thanks also go to my friends and family. Nicholas d'Apice and my friends in Moscow, Russia deserve a special mention. Thank you so much for always being there! To my parents, Valerii and Lida, my brother Misha, uncle Maksim and his family, to my grandmother Ira and all my relatives for their continuous support and belief in me. They have proven that encouragement knows no geographical distance. Lastly, I would like to express my endless gratitude to my brilliant home team: my wonderful wife, Emma, and precious son, Leo. Knowing that you were there, believing in me and keeping me motivated each day, was the greatest encouragement of all.

Hamburg, July 2017

Ivan Terekhov

Forecasting Air Passenger Demand between Settlements Worldwide Based on Socio-Economical Scenarios

Ivan Terekhov

Summary

This thesis presents an air passenger demand (APD) forecasting model which forecasts the future APD network between settlements worldwide and its corresponding passenger numbers based on socio-economic scenarios. This modeling approach has not yet been considered by existing studies. An APD forecast at settlement level is particularly important for rapidly developing countries such as China. By not modeling the APD at settlement level, future air passenger flows and, therefore, future air traffic volumes could be underestimated, resulting, for example, in increased detrimental environmental impacts deriving from CO_2 emission levels associated with aviation.

The presented APD forecasting model contains two main parts: forecasting the potential for demand between settlement pairs and the expected APD based on new and existing connections. For the first part, the topology forecasting model is developed to determine whether the potential for demand between a given settlement pair exists or not. For the second part, the passenger forecasting model, derived from the potential for the APD existence between settlements, seeks to forecast the realized APD between these settlements.

The APD forecasting model was validated on real data for one, five and ten-year time intervals to forecast the APD topology and passenger number from 2002, 2007 and 2011 to 2012 and includes more than 3,600 settlements worldwide. The modeled results were then compared to the actual real data from 2012. Since the APD generation process varies depending on the settlement, clustering methods were applied to allocate the settlements to nine groups according to their socio-economic indicators, where settlements in each group possess similar patterns. The APD model validation shows sufficient accuracy. After analyzing the validation results, it was

found that settlement clustering triggered APD topology forecasting model accuracy improvements from 68% correct APD connections to 78%. The highest overall model accuracy is 35% of APD connections which were correctly predicted to within ±20% of passengers. This nevertheless covers more than 70% of the actual passenger number in 2012. This accuracy was achieved using a one-year time interval from the base year. However, the error propagation analysis demonstrated that for the long-term forecast (more than ten years), the expected accuracy is higher using a ten-year time frame (0.3 on average for the APD connection forecast covering 47% of passengers for the ten-year interval versus an accuracy of 0.04 for the connection forecast covering 37% of passengers for the one-year time interval for ten years).

The validated APD forecasting model was applied to the four GEO-4 socio-economic scenarios developed by the UN from 2012 to 2042 using a ten-year time frame. Since the scenarios do not provide the average annual airfares between settlements, a simple airfare model was developed based on historical data. In order to verify the modeling results, they were compared to existing forecasts from Airbus, Boeing and ICAO FESG (International Civil Aviation Organization, Forecasting and Economic analysis Support Group) using a coefficient for transferring modeled demand passenger kilometers (DPK) to revenue passenger kilometers (RPK). The comparison showed that the existing forecasts are more optimistic in terms of future RPK growth in contrast to the results obtained from the APD forecasting model based on GEO-4 socio-economic scenarios. For instance, a comparison of the Boeing forecast and the APD forecasting model for the Asian region in 2034 shows 4287.7 and 3157.6 billion RPK respectively.

The next section of the thesis addresses APD evolution in China, which has experienced the largest growth rate. According to the model, China's global APD share for 2042 is estimated to be about 36% in comparison to around 15% for 2012. The APD in China grew from approximately 377 million in 2012 to an expected 2.97 billion in 2042. The analysis showed that some settlements in China, which did not have a high APD in 2012 and main internal journeys within China to only a few destinations, will generate a significant APD in 2042 for those traveling to various settlements both within China and globally.

Table of contents

List of figures

List of tables

Abbreviations

ADI	Airport Data Intelligence
AIC	Akaike Information Criterion
AIM	Aviation Integrated Modelling
AIRCAST	AIR travel forecast
ANSPs	Air Navigation Service Providers
APD	Air Passenger Demand
ATLab	Air Transport Laboratory
ATM	Air Traffic Management
ATS	Air Transportation System
AUC	Area Under the receiver operating Curve
BIC	Bayesian Information Criterion
DF	Directness Factor
DPK	Demand Passenger Kilometers
EC	Exclusive Clustering
GDP	Gross Domestic Product
GDS	Global Distribution System
GEO	Global Environment Outlook
GMF	Airbus Global Market Forecast
FESG	the Forecasting and Economic analysis Support Group
HC	Hierarchical Clustering
ICAO	International Civil Aviation Organization
ML	Maximum Likelihood
OD	Origin-Destination

PC	Probabilistic Clustering
PM	Probabilistic Models
QA	Quantitative Analogies
RPK	Revenue Passenger Kilometers
SAM	Simple Airfare Model
UK	United Kingdom
UNEP	United Nations Environmental Program
USA	United States of America
WAA	Weighted Adamic-Adar index
WRA	Weighted Resource Allocation index
WCN	Weighted Common Neighbors
WHDP	Weighted Hub Depressed index
WHPI	Weighted Hub Promoted Index
WLHN	Weighted Leicht-Holme-Newman index
WPA	Weighed Preferential Attachment index
WSA	Weighted Salton index
WSO	Weighted Sorensen index

1. Introduction

Since the beginning of the 20th century, air transport has been playing an increasingly significant role in passenger mobility worldwide. Air transport connects cities, providing opportunities to travel to almost anywhere in the world and, therefore, stimulating social and economic development (Lakshmanan, 2014) and globalization (Hummels, 2007). Air transport has achieved its high demand by offering swift transportation between global origins and destinations at a reasonable cost in comparison to other transport modes, especially over long distances. However, due to competition with other means of transportation and in order to enhance revenue and minimize costs, the air transportation system is constantly on the lookout for ways to improve. Developing new and improving existing technologies increases efficiency and air travel attractiveness for passengers. Therefore, rapid rates of technological development, high levels of air transport competition faced by companies and constantly increasing air travel demand have defined the structure of the air transportation system (ATS). This is a large, multi-disciplinary, complex system with various interactions between stakeholders (such as manufactures, operators, airports and air navigation service providers (ANSPs) and its environment (see Fig.1.1). The system's elements are highly integrated and have strong connections with external environments including politics, the economy, the environment, technologies, and society. Modifying any element of this interconnected structure may cause changes in the whole system. For example, the hub-and-spoke route network system, which was first implemented by Delta Air Lines in 1955 (Delta Air lines, 2016), demonstrated its efficiency to such an extent that it was optimized and consequently adopted by numerous air companies worldwide. The newly implemented system undoubtedly left a positive impression on the key

stakeholders and prompted adaptation to the new conditions. One example of the potential consequences of political decisions surrounding the ATS is the deregulation in the USA in 1978, where significant changes to the entire system were implemented (Morrison and Winston, 1995). The understanding of these changes and their relations between the stakeholders is essential to explain and assess the complex processes within the ATS as well as for future ATS development estimation.

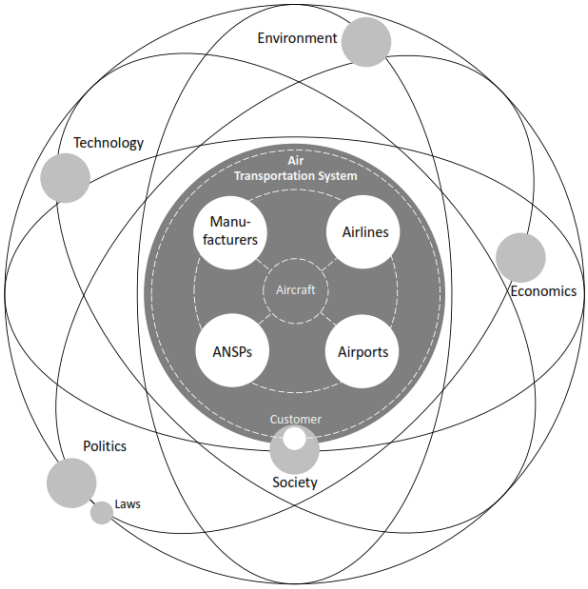


Fig.1.1. A basic representation of the ATS and its external environment (Ghosh et al, 2014)

The ATS has shown sustainable growth over the last decades. The number of settlements with at least one airport in 2012 was over 6,000 and more than 20,000 passenger aircraft were operated by airlines worldwide (Flightglobal, 2015). This growth has been stimulated by technological improvements; yet, more importantly, its dependence on general socio-economic development cannot be underestimated due to a clear correlation with world economic and social growth. In just one decade, i.e. from 2002 to 2012, the air passenger demand (APD) worldwide within the ATS increased from 1.5 to 3 billion passengers (Sabre Airline Solutions, 2014), while the combined world gross domestic product (GDP) and the world population grew at a comparatively lower rate, from 50 to 73 trillion US dollars (The World Bank, 2014) and from 6.28 to 7.1 million inhabitants (UN, 2014-2) in the same period (Fig.1.2). Such growth has undoubtedly had an increasingly negative effect on the environment, including impacts on air

quality, noise, and global climate. For example, carbon dioxide is a very long lived greenhouse gas which affects the climate system hundreds of years after being emitted into the atmosphere. Moreover, for current engines, carbon dioxide emission is unavoidable as the gas directly corresponds to fuel consumption (Svensson, 2005). Thus, in response, certain international organizations have set ambitious goals to reduce these effects. One of the major targets of the International Air Transport Association (IATA) to mitigate CO_2 emissions from air transport is to reduce net aviation CO_2 emissions by 50% by 2050, in comparison to levels recorded in 2005. Their "IATA Technology Roadmap Report" (IATA, 2013) analyses technologies for future aircraft that will reduce, neutralize and eventually eliminate the carbon footprint caused through aviation transport. The European Commission, together with key European aviation stakeholders, has established its vision on the future of air transport in "Flightpath 2050" (European Commission, 2011). The document focuses on two main challenges: meeting the needs of people and the market as well as maintaining global leadership. Among the key objectives are a 75% reduction in CO_2 emissions per passenger kilometer and a 90% reduction in NO_x emissions by 2050.

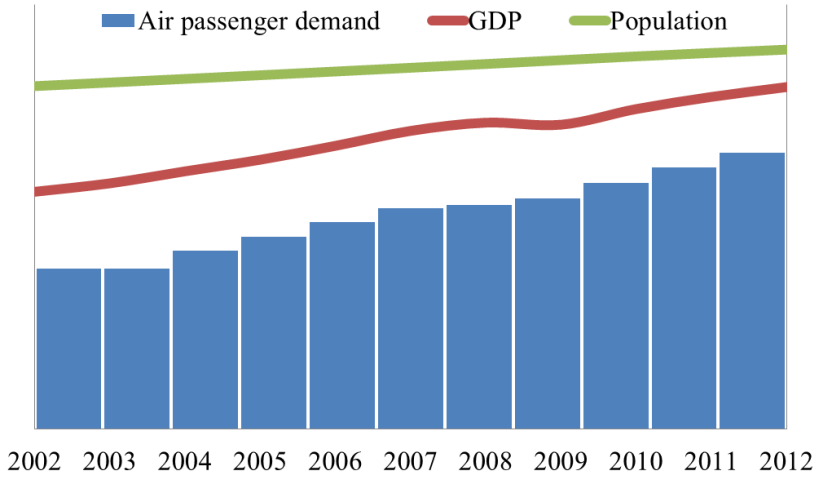


Fig.1.2. World GDP, population and APD from 2002 to 2013

As shown by Lee (Lee et al., 2009), air transport contributes 2-3% of global CO_2 emissions and 3.5-4.9% of global radiative forcing, if non-CO_2 effects are included. Today, the impacts of CO_2 emissions on the environment are closely being studied (Gössling and Upham, 2009; Apffelstaedt et al., 2009; Nolte et al., 2012; Pachauri et al., 2014). However, non-CO_2 effects

have not been subject to the same level of attention. Accordingly, a more robust scientific understanding of the effects of non-CO_2 emissions is still needed (Kollmuss and Crimmins, 2009). The non-CO_2 emissions have different impacts on the environment in different regions of the world (Grewe and Stenke, 2008; Koch et al., 2011). For example, NO_x has been shown to induce the short-lived greenhouse gas ozone. The gas produced at the equator has a higher radiative forcing than the same amount of emissions in northern regions. This implies that the geographical information of a flight route, such as the location of departure and destination airports, as well as the flight path, is essential for assessing the impact of non-CO_2 emissions on the environment. Accordingly, to assess the non-CO_2 impact, the number of flights and type of aircraft operated on routes must be known so as to quantify the amount of such emissions on a global scale. To obtain this information, the number of passengers on these routes must be estimated. Finally, to make such estimations, the APD between origin and destination settlements has to be determined. Thus, the forecast of the APD between settlements forms an important base for assessing the impact of future non-CO_2 emissions on the environment.

A number of integrated aviation-environmental models for assessing different ATS policies and their potential future impacts on the environment have been developed. An example of this model type could be the Aviation Integrated Modelling (AIM) Project (Dray et al., 2010; Dray et al., 2014). These models have various sub-models and inputs, but one of the important elements is an air passenger demand forecast model. This sub-model provides the simulated air passenger demand as an input to the next sub-models in the integrated environment. Using results from demand forecasting models, technology development scenarios and other inputs, integrated environmental models simulate air traffic growth and their environmental impacts. For instance, AIM predicts settlement-pair demand, but solely for a fixed settlement set defined by the largest cities with air passenger flights in the base year. Therefore, it doesn't capture new origins and destinations. Other demand sub-models and integrated models do not specifically concentrate on a settlement level of aggregation and instead primarily focus on regional level to assess the environmental impact. Along with the APD simulation within the aviation environmental integrated models, a number of APD models either exist independently or are included in other studies (e.g. the “General Market Forecast” from Airbus (Airbus, 2014)). These studies mainly aim to simulate the APD between regions or between particular origins and destinations in order

to predict particular figures (e.g. the future world fleet or airport capacities). Models developed within these studies after modification could also be used for the global forecast of environmental impacts at regional level or on particular routes.

The APD models play an important role in modeling ATS growth. Most of the existing approaches for APD forecasting worldwide on their level of aggregation assume that all elements (e.g. regions) are interconnected (i.e. AIM's main approach). In other words, all considered elements in the demand model are connected to each other and this connectivity does not change within the forecast period. This seems reasonable at regional level, but the same can hardly be said at settlement aggregation level. In the future, it is likely that there will be a number of settlements with significant air traffic connections that have no air traffic connections today. This is particularly true for rapidly developing countries such as China. Not including air traffic to these growing cities would result in an under-estimation of global growth in air traffic; which, in turn, would mean that emission levels associated with aviation would also be underestimated. Thus, it is particularly important to have an accurate forecast of APD growth so as to assist in developing a more accurate forecast of the environmental impacts (non-CO_2, for instance) of the growing ATS. On the one hand, accurately forecasting growth is important for setting realistic targets, and understanding what needs to be done to meet those targets. Carefully forecasting environmental impacts is important for setting targets that are effective at actually reducing these impacts, and using effective metrics for these targets. Thus, without reliable predictions of the demand for air passenger travel, it is impossible to estimate air traffic flows, aircraft emissions, and ultimately the environmental impact of aviation. On the other hand, in order to achieve the set environmental targets, the decision or policy maker should develop a strategy containing a number of decisions which need to be made in a certain time frame. Thus, by using different input socio-economic scenarios of future development, the decision or policy maker is able to assess the influence of his future decisions on the entire future APD worldwide and, therefore, develop an optimal strategy to achieve the final aim. In addition, such an APD modeling approach allows the creation of a dynamic model whose outputs at settlement level could be an important basis for further modeling in the ATS: such as routes and hubs modeling, aircraft size and their frequencies on routes and, finally, the future environmental impacts.

Therefore, at DLR Air Transportation Systems, a special set of models called AIRCAST (AIR travel foreCAST) is being developed within a modular environment in order to forecast future development of the global ATS at settlement level based on socio-economic scenarios. Uniting models through AIRCAST allows a range of possible outcomes for the future ATS to be simulated and, for example, for the impact of new technologies on the APD, the size and the aircraft number on particular routes, emission amount on particular routes to be assessed. The set of models is distributed between four interconnected layers as follows (Fig.1.3):

1) The origin-destination demand layer – defines passenger number on settlement pairs
2) The route layer – defines which settlement pair will be operated by air carriers
3) The aircraft movement layer – defines the aircraft number on routes, their types and capacities
4) The trajectory layer: defines trajectories for aircraft on routes, amount and type of emitted emissions on routes

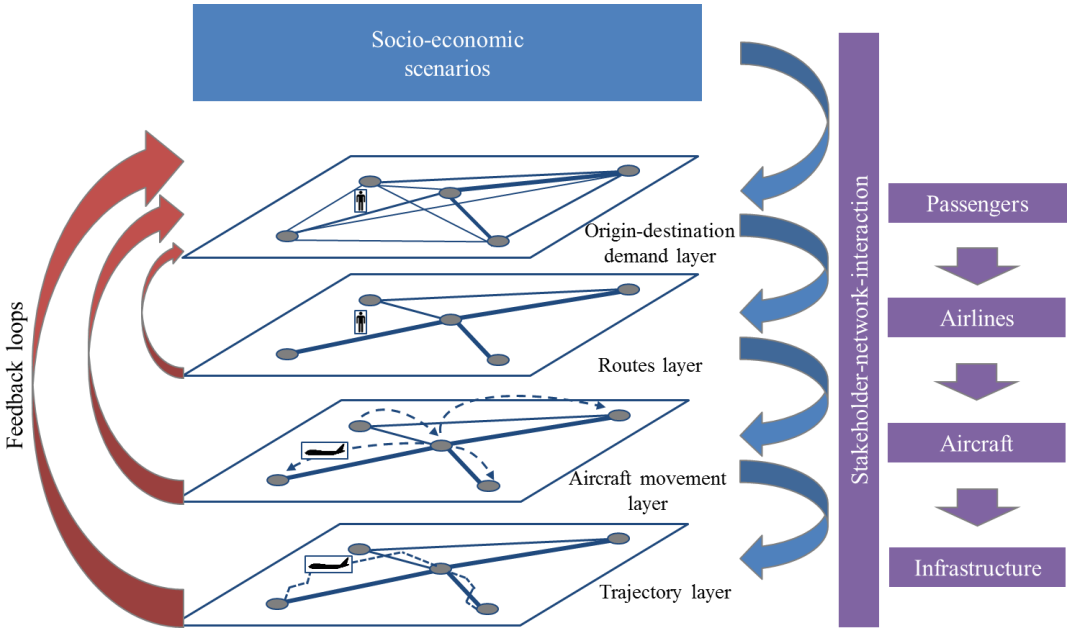


Fig.1.3. AIRCAST four-layer approach

Layers are connected by back loops which allow models to be calibrated. For instance, the origin-destination demand network layer receives socio-economic indicators as input and transfers a modeled passenger number on modeled settlement pairs to the next layers as output.

Back loops from other layers (red arrows in Fig.1.3) transmit, for example, travel time between settlement pairs, aircraft flight frequencies between settlements and aircraft capacities to the demand layer. Using this information, models at origin-destination demand network level are calibrated by generating updated outputs for the next models. Thus, the modeling process itself is iterative and terminates when a selected criterion or criteria are met.

This thesis describes the APD forecasting model from the first AIRCAST layer in order to assess the impact of different socio-economic scenarios on the world APD and its topological structure. Since the model chain is under development, the APD forecasting model in this thesis disregards the back loops from other AIRCAST layers. Based on socio-economic scenarios, the APD forecasting model calculates future passenger numbers between settlements, taking into account the possibility of changes in the number of APD connections between settlements within the ATS over time. The thesis is organized as follows. Chapter 2 presents a literary review of the different APD forecast studies, taking industry and academic forecasts into account. Analyzing the current studies, a gap is defined in the area of worldwide settlement level APD modeling as well as a lack of APD forecasting based on different socio-economic scenarios at this level. Based on the identified gaps, Chapter 3 describes the research objectives for the thesis to develop, evaluate and apply the APD forecasting model. The research methodology is expanded on there to fulfill the introduced research objectives. Chapter 4 presents the APD forecasting model which has been created to fulfill the developed research objectives. The section describes the overall model framework, required inputs into the APD forecasting model as well as all the sub-model descriptions: clustering settlements by their socio-economic indicators, the topology forecast model and the passenger forecast model. Chapter 5 describes the APD forecasting model development and validation. Firstly, the model validation framework is shown. This is followed by a study to define the appropriate clustering method for the settlements separated into groups. Thereafter, the appropriate method for APD connection forecast is defined for the topology forecasting model. This is based on the analysis of various link prediction approaches in the network science area. The obtained algorithm is validated on real data by forecasting the APD network from 2002, 2007 and 2011 all connecting to 2012 and comparing them to the actual 2012 values. The validated topology forecasting model is then adapted to all cluster pairs by defining the boundaries from the historical data, after which time the passenger forecasting model

itself is described. The model contains two sub-modes: quantitative analogies for passenger forecast on the newly appeared connections in the APD network and the correlation between passenger and GDP growth for APD connections remaining in the network. The passenger forecast model is validated on real data by forecasting passenger numbers from 2002, 2007 and 2011 all connecting to 2012 and then comparing them to the actual 2012 values. At the end of the chapter, the overall APD forecasting accuracy is calculated and the expected error propagation is analyzed.

Once validated, the APD forecasting model is then applied to four GEO-4 socio-economic scenarios from the UN from 2012 to 2042 in Chapter 6. Firstly, in order to forecast the APD, a simple airfare model is developed and validated. This model allows the average annual airfare between settlements to be modeled based on the distance between them and the average annual crude oil price. Then, the description of four GEO-4 scenarios is presented, showing the scenario storylines. Using these scenarios, the APD forecasting results are subsequently presented. The results are verified in relation to existing forecasts such as Airbus, Boeing and ICAO FESG APD forecasts. In order to demonstrate the level of detail of the APD forecasting model, the APD forecast for China in 2042 is analyzed for the Sustainability First GEO-4 scenario.

Chapter 7 offers conclusions for this thesis. The conclusions show that the APD forecasting model is a valuable tool for forecasting APD at settlement level which is able to provide relevant results for decision makers. Chapter 8 ends the thesis with recommendations for future research. These recommendations are mainly given in order to improve the APD forecasting model's performance by carrying out further study on the assumptions accepted in this study.

2. Literature review

As shown in Chapter 1, the APD forecast is an important basis for planning in the constantly changing ATS. This chapter presents different existing APD modeling approaches. The aircraft industry and researchers study APD and develop forecast models using various techniques and levels of aggregation. Section 2.1 demonstrates current APD modeling approaches of the main players in the air transportation industry such as Airbus and Boeing in the context of their market forecasts. In Section 2.2, academic studies on APD are presented. A summary of this chapter and conclusion is given in Section 2.3.

2.1. Industry forecasts

In this section, demand models of the aviation industry are considered. Aviation companies generally make their own market forecasts to demonstrate their visions on the future development of air transportation to potential customers. For example, Airbus positions their Global Market Forecast (GMF) as a "view of the demand for civil passenger and freighter aircraft that will serve as a reference for airlines, airports, investors, government and non-government agencies, air transport and economic planners world-wide" (Airbus, 2013). The purpose of Boeing's Current Market Outlook is "to shape [the] Boeing product strategy and guide long-term business planning …share our outlook with the public to inform airlines, suppliers, and the financial community of trends we see in the industry" (Boeing, 2015). The aim of these forecasts is to show customers that the demand for these companies' productions in different regions is growing and in helping

customers to make the *right* decisions about which items to acquire (Anker, 2000). Since these market forecasts have been made by industry companies, the objectiveness of such calculations have to be critically scrutinized in the context of overall air transportation system development. In addition, aviation industry companies do not publish details about their forecast procedures which makes it difficult to understand the underlying principles and assumptions of their methodologies, models and sub-models. However, approaches to forecast air passenger demand are the basis of these market forecasts and will be considered in this section.

As discussed in Doucet et al., 2014, an origin destination air passenger demand model is an important part of the Airbus Global Market Forecast (GMF) (Airbus, 2014) methodology. The GMF methodology for forecasting future ATS contains three basic steps: traffic forecast over the next 20 years, a network forecast[1] and a required aircraft number forecast. Air passenger demand forecast in the GMF forms part of the second step. For the network forecast, a traffic forecast between countries was initially disaggregated to a set of city pairs. Then, flight segments were modeled between any two settlements in the set. The obtained flight segments network included existing routes as well as future possible routes. Applying a market share model, a percentage of air passengers was assigned to each flight segment. Finally, the number of passengers was defined as the percentage of passengers on each flight, multiplied by the origin-destination demand between settlements. The origin-destination air passenger demand model used a modified gravity model to forecast the number of passengers between 279 settlements around the world. The modified gravity model took into account a spatial dependence between origin and destination. In other words, it considered the impact on air passenger flow between settlements by employing characteristics at proximal settlements.

Other industry forecasts mainly predict Revenue Passenger Kilometers (RPK). These forecasts do not present a separate air passenger demand model. The Boeing Current Market Outlook 2013-2032 (Boeing, 2013) forecast used an empirical equation where RPK growth between regions was equal to the sum of GDP growth and a "time-varying function" (Eq.2.1). The function was not directly associated with GDP growth. This component of growth derived from the value travelers place on the speed and convenience that only air travel can offer.

[1] Here "network forecast" implies a forecast of routes between cities

$$RPK_{growth} = GDP_{growth} + f(t) \qquad \text{(Eq.2.1)}$$

Where $f(t)$ is a time-varying function that typically centers around 2 percent.

The Rolls-Royce (2012) Market Outlook developed an RPK forecast between regions and did not indicate the forecasting methodology. Market Outlook 2015-2034 from Embraer (2015) provided an RPK forecast between world regions. The forecast was based on historical traffic data, GDP development, trade, tourism, fuel price, population dynamics, airline competition and traffic of other methods of transportation.

Japan Aircraft Development Corporation (2010) presented their "Worldwide market forecast for commercial air transport 2010-2029" at region level. Eleven regions were considered in the forecast and passenger traffic was analyzed from the past to obtain regression equations for each region. These equations were based on the relationships between GDP, airfare and RPK, yet the equations were not printed in their publication.

The United Kingdom Department for Transport's 2013 UK Aviation Forecasts (UK Department for Transport, 2013) included the National Air Passenger Demand Model. This model used a combination of a set of time series econometric models of past UK air passenger demand including projections of key driving variables and assumptions about how the relationship between UK air travel and its key drivers would change in the future. The model provided forecasts for domestic destinations within the UK, international regions of origin for flights to the UK and international passengers connecting through UK airports. It also accounted for airport capacity constraints, redistributing demand to other airports when demand exceeded capacity.

2.2. Academic studies

The topic of the air passenger demand modeling has also received a great deal of attention in the academic field. As input for various models, air passenger demand models are developed at different levels of aggregation with different scopes and aims.

The Aviation Integrated Modelling (AIM) (Dray et al., 2010; Dray et al., 2014) project was initiated by the University of Cambridge, UK. The aim of this project was to develop a tool to assess different current and future policies in aviation (Reynolds et al., 2007). The AIM project contained a set of connected modules that were created to fulfill the policy assessment goals of the task. These modules were: Aircraft Technology & Cost Module, Air Transport Demand Module, Airport Activity Module, Aircraft Movement Module, Global Climate Module, Local Air Quality & Noise Module, and Regional Economic Module. The Air Transport Demand Module deals with true origin-destination (OD) air passenger and freight demand. Currently, this module contains a simple model at city level which considers realized undirected air passenger demand. The model was represented as a gravity model with OD connections between 700 settlements around the world (Eq.2.2).

$$D_{ij} = K(I_i I_j)^{\alpha}(P_i P_j)^{\gamma} e^{\delta A_{ij}} e^{\varepsilon B_{ij}} e^{\varphi S_{ij}} e^{\omega DF_{ij}} e^{\mu R_{ij}} C_{ij}^{\tau} \qquad \text{(Eq.2.2)}$$

Where I_i represents the average local per capita income of city i, P_i is the greater metropolitan area or equivalent population of city i and C_{ij} is the generalized cost to a passenger of air travel between settlements including delays, A_{ij} and B_{ij} are dummy variables indicating whether one or both cities in the pair are major tourism or business destinations, S_{ij} and R_{ij} show whether a road or high-speed rail link exists between i and j, DF_{ij} indicates whether the route is a domestic one, $K, \alpha, \gamma, \delta, \varepsilon, \varphi, \omega, \mu$ and τ are parameters to be estimated. The gravity equation was adapted to short-haul, medium-haul and long-haul as well as for different regions. The equation was calibrated on current and historical data.

Suryani et al. (2010) modeled air passenger demand and passenger terminal capacity expansion using a system dynamics approach. The casual loop diagram (Fig.2.1) represents the relationship between population, GDP growth, level of service impact, airfare impact, runway utilization and required passenger space. The study concentrated on single airport level. Their model predicted when an airport should expand runway and passenger terminal capacities and determined the total airport area needed to meet future demand.

Alam and Karim (1998) addressed the present condition of the air transportation system in Bangladesh. They analyzed the operation and level of service of the system, supply structure and

the network configuration. A stepwise multiple linear regression analysis, using multiple time series collected over five years, was utilized to calculate total passenger trips per week along existing routes. The regression model in their study was established as follows (Eq.2.3).

$$T_{ij} = e^{a}(P_j)^{b}(E_j)^{c}(R_{ij})^{d}(X_j)^{e}(G_j)^{f} \qquad \text{(Eq.2.3)}$$

Where T_{ij} is the total passenger trips per week between cities i and j, P_j is the total population for city j, E_j is the employees number in city j, R_{ij} is the travel time ratio for traveling between cities i and j, X_j is a dummy variable, G_j is the GDP per capita, a, b, c, d, e and f are elasticity parameters.

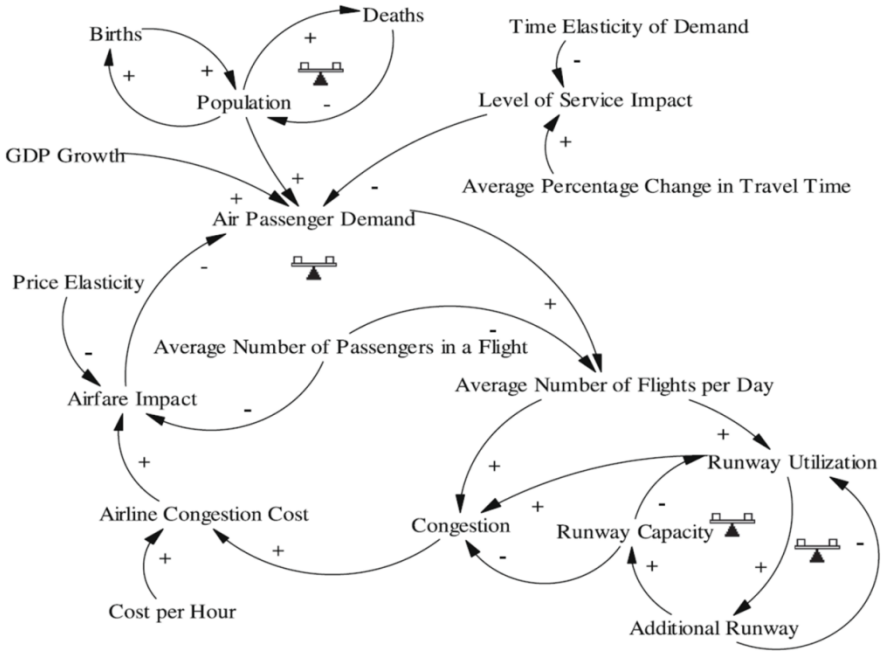


Fig.2.1. Causal loop diagram of air passenger demand and passenger terminal capacity expansion (Suryani et al., 2010).

Grosche et al. (2007) presented two gravity models for estimating air passenger volumes between city pairs. The estimation was based on socio-economic and geographic factors for the fixed number of city pairs. The basic gravity model was presented as follows (Eq.2.4).

$$V_{ij} = e^{\varepsilon} P_{ij}^{\pi} C_{ij}^{\chi} B_{ij}^{\beta} G_{ij}^{\gamma} D_{ij}^{\delta} T_{ij}^{\tau} \qquad \text{(Eq.2.4)}$$

Where V_{ij} is the total passenger volume between cities i and j, P_{ij} are the populations of cities i and j, C_{ij} are the airport catchment areas in cities i and j, B_{ij} is the average buying power index based on an airport's catchment area in cities i and j, G_{ij} is the gross domestic product of the country of the airports in cities i and j, D_{ij} is the distance between two airports, T_{ij} is the travel time between two airports, $\varepsilon, \pi, \chi, \beta, \gamma, \delta$ and τ are elasticity parameters.

The extended gravity model (Eq.2.5) took into account multi-airport cities adding variables that describe the spatial characteristics to the basic gravity model (Eq.2.4).

$$V_{ij} = e^{\varepsilon} P_{ij}^{\pi} C_{ij}^{\chi} B_{ij}^{\beta} G_{ij}^{\gamma} D_{ij}^{\delta} T_{ij}^{\tau} N_{ij}^{\upsilon} A_{ij}^{\alpha} W_{ij}^{\omega} \qquad \text{(Eq.2.5)}$$

Where N_{ij} is the number of competing airports in cities i and j, A_{ij} is the average distance to competing airports, W_{ij} is the number of competing airports weighted by their distances, υ, α and ω are elasticity parameters. As seen, these two gravity models did not take into account the possibility of new city pairs within the air transport system. The gravity models for air travel demand estimation have a long history and studies include various variables such as population, income, distance, etc. (Brown and Watkins, 1968; Verleger, 1972; Fotheringham, 1983; Rengaraju and Arasan, 1992; Russon and Rilay, 1993; O'Kelly et al., 1995; Jorge-Calderon, 1997; Doganis, 2002; Shen, 2004). However, these studies do not consider the possibilities of changes in the connected city number.

In contrast, air transport has also been analyzed with complex networks. Zanin and Lillo (2013) gave an overview of complex network theory application to ATS where they showed that for future ATS challenges, the complex network theory would play a more significant role in tackling these challenges. The ATS analyses were carried out in accordance with different network representations. In some studies, the ATS network is considered as an unweighted network (a route network of 2001 with 27,051 connections was analyzed) in order to obtain specific ATS network metrics (Guimera et al., 2005) or to define the most efficient flights for airlines in terms of benefits and passenger mobility (Bania et al., 1998; Alderighi et al., 2007). In other studies, weighted ATS networks were analyzed assuming the weight is the available seat number (Barrat et al., 2004; Wu et al., 2006) or passenger number (Xu and Harriss, 2008), or flight number (Lillo et al., 2011) in order. The ATS network evolution is analyzed in a few

studies. Han et al. (2007) and Li and Cai (2004) analyzed the main metric changes in Austrian Airline Flights and China's airport network respectively over a seven-day period. In general, the majority of studies in the complex networks field focus on the topological properties of the flight networks. Such studies do not consider the evolution of these networks over time.

2.3. Conclusion

The studies mentioned above utilized a range of techniques and various levels of aggregation. Industry forecasts and academic studies show various methods to calculate the demand at particular airports, on particular routes, at regional or city level with fixed number of connections between settlements. The aforementioned forecasts mostly dealt with air passenger demand using gravity models. Although gravity models demonstrate a relatively high accuracy (Grosche et al, 2007), they have to be calibrated for different types of city pairs (e.g. short-haul, medium-haul, long-haul, international, regional, local, etc.). Thus, when dealing with larger numbers of city pairs, the complexity of the calibration requirements of these models increases.

The studies which used complex network theory mainly analyzed current flight networks. Yet again, the long-term evolution of these networks was not considered. However, this could be a useful starting point for a symbiosis between the air transportation network and complex networks in order to forecast the future state of the air transport network which has not received the same level of attention.

It can be concluded that current studies do not provide a method of forecasting the evolution of air passenger demand between settlements at global level. They fail to take into account the potential for change in the number of airport-connected settlements when forecasting demand within an air transport system.

3. Research objectives

The literature review in Chapter 2 showed that there is no appropriate air passenger demand model containing settlements worldwide which is able to assess passenger flows at city level, while allowing the changes in the number of connections between these settlements to be forecasted. Thus, in order to capture these changes at city level, a model based on socio-economic indicators has to consider representative settlements worldwide (e.g. with at least one airport) and simulate the air passenger demand between them. This approach would provide a more realistic and detailed understanding of the air passenger demand evolution within the forecast period, based on a socio-economic development scenario of the settlements worldwide.

This chapter is organized as follows: Section 3.1 presents the key research objectives of this thesis and Section 3.2 shows the research methodology that was developed to fulfill the key research objectives.

3.1. Key research objective

As shown in Chapter 1, the ability to forecast passenger number on settlement pairs and capture the changes in passenger settlement pair topology is significant for assessing future air transport development. In particular, changes in air passenger demand topology and passenger number on origin destination settlements are an important basis for airline companies to ultimately decide whether to operate air service on specific settlement pairs or not and, therefore, to adapt their route network accordingly. Furthermore, the interface between origin-destination passenger demand and route network is essential within AIRCAST so that passenger streams

with airline supplies and structures can be connected. Thus, the research objective of this study is to develop an APD forecasting model to forecast the evolution of air travel passenger demand between settlements worldwide based on a socio-economic scenario, taking into consideration the probability of changes in the number of APD connections between settlements within an ATS, i.e., the number of settlement pairs that have any demand between them (regardless of flight network). In other words, the proposed model has to forecast air passenger demand and changes to the settlement pair connection topology of an air passenger demand network, within a forecast period as well as adequately respond to disruptive events described in scenarios and model their impact on the APD.

Following this approach for modeling APD on a worldwide scale, the APD forecasting model should contain two main parts: forecasting the potential for demand between settlement pairs and forecasting realized APD on new and existing connections. The first part of the APD forecasting model should determine whether the potential for demand between a given settlement pair exists or not. The second part, based on the potential for the APD existence between settlements, seeks to forecast the expected APD between these settlements. Since the demand for air travel originates in places with high population density and airports are one of the means to realize given demand, this thesis only considers the generation of the potential demand between settlements. Although the APD between settlements is directed, to limit the scope of the study the APD, it is assumed to be undirected. In other words, directed APD is the demand in direction from point A to point B in a time period *t*, but does not include the demand in direction from point B to A in the same time period *t*. The sum of the directed APD on directions A-B and B-A in the time period *t* represents the undirected APD between points A and B. Thus, the undirected demand is the total APD between two points regardless of direction. Moreover, the APD on settlement pairs varies depending on the time of year, specific events (e.g. a world sport competition in a city or a country) or conflicts in particular areas and other cases. However, their impact on the APD on settlement pairs is not significant compared to the global number of passengers traveling by air. At the same time, it is almost impossible to generate a precise forecast of possible conflicts in the future and where they will occur. Thus, to limit the scope of the study, this thesis considers the APD between settlements at yearly intervals. In other words, the minimum time frame in the APD forecasting model is one year.

A special methodology has been developed in order to fulfill the objective of this study. This approach is able to simulate the APD for a various number of socio-economic scenarios on a yearly basis. Based on socio-economic indicators from the scenario which develop over time and information from the base year as the input, the APD forecasting model simulates the APD between settlements worldwide. The modeling approach allows forecasting changes in the topology of the worldwide APD between settlements and defining the passenger number on these settlement pairs at yearly intervals.

In order to validate the APD forecasting model, it is checked against historically observed data. The APD for a past year (e.g., 2012) is simulated using a known APD network and socio-economic indicators from an earlier year (e.g., 2002). Simulated results are then compared to the actual APD network.

The validated model is then applied to the four socio-economic scenarios to obtain a consistent forecast of the APD between settlements from 2012 to 2042. Then the results of the Air Passenger Demand Model of the APD topology and the number of passengers are transferred to RPKs based on the historical trend and compared to existing Airbus, Boeing and ICAO FESG forecasts.

3.2. Research methodology

In order to fulfill the key research objective of this study, the research methodology is separated to three steps: model development, model validation and model application.

1. Model development. The APD forecasting model based on a socio-economic scenario and a base year simulates consistent changes in the APD network topology and the passenger number on these connections on a yearly basis. The methodology of the proposed model and detailed descriptions of sub-models are shown in Chapter 4 and Chapter 5.

2. Model validation. The model validation[2] is conducted by modeling the APD from 2002, 2007 and 2011 all connecting to 2012 and then compared to the actual 2012. The model is validated on almost 4,000 settlements worldwide and the actual socio-economic indicators of 2012 are used as the socio-economic scenario. The detailed APD forecasting model validation procedure and the results are presented in Chapter 5.

3. Model application. The validated APD forecasting model is applied to four UN GEO-4 socio-economic scenarios. The model provides simulation of the APD from 2012 to 2042 per decade. The model uses 4,251 settlements in 180 countries worldwide. The APD forecasting model application, results verification[3] at global level as well as the detailed APD analysis on settlement level for China are shown in Chapter 6. The final conclusion is discussed in Chapter 7.

[2] In this study validation is considered as "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model" (Thacker et al, 2004)

[3] In this study verification is considered as "the process of determining that a model implementation accurately represents the developer's conceptual description of the model and its solution" (Thacker et al, 2004)

4. APD forecasting methodology

This Chapter presents the APD forecasting model methodology which simulates the realized air travel passenger demand between settlement pairs worldwide based on various socio-economic scenarios. These scenarios determine whether the APD between settlements changes. The APD between a settlement pair could appear or disappear or the number of passengers could increase or decrease. The overall configuration of the APD between settlements is able to change in response to the socio-economic situation in a particular settlement, country or region. In order to capture these changes in the APD, a special framework has been developed. Section 4.1 provides a general modelling framework which demonstrates how sub-models and their inputs are connected to each other. In Section 4.2, all sub-models from the modeling framework are described in detail.

4.1. Modeling framework

The proposed APD forecasting method between settlement pairs worldwide aims to forecast the APD topology network and passenger numbers on these APD connections from the base year y_0 to the last year of a socio-economic scenario y_t with a given discrete time interval n. In order to implement this approach, the method contains three steps: 1. City clustering in y_0 and identifying the cluster dynamics from y_1 to y_t (Terekhov and Gollnick, 2015); 2. Forecasting the topology of the APD network from y_1 to y_t (Terekhov, et al., 2016); 3. Calculating the APD demand on existing and new connections from y_1 to y_t (Terekhov, et al., 2015-1). The first step of the method is settlement clustering in the base year y_0. It divides settlements according to the

similarity of their socio-economic indicators and, based on the socio-economic scenario, defines the content of every cluster from y_1 to y_t, i.e. the cluster dynamics. The second step determines whether the demand connection between a given settlement pair in every cluster pair from y_1 to y_t exists. This is done by implementing a weighted similarity-based algorithm. The weight is represented by a combination of socio-economic information of settlements in pairs, and the distance between them. The third step of the method, based on the existence of the APD connections between settlements and cluster pairs, seeks to forecast the APD between these settlements from y_1 to y_t.

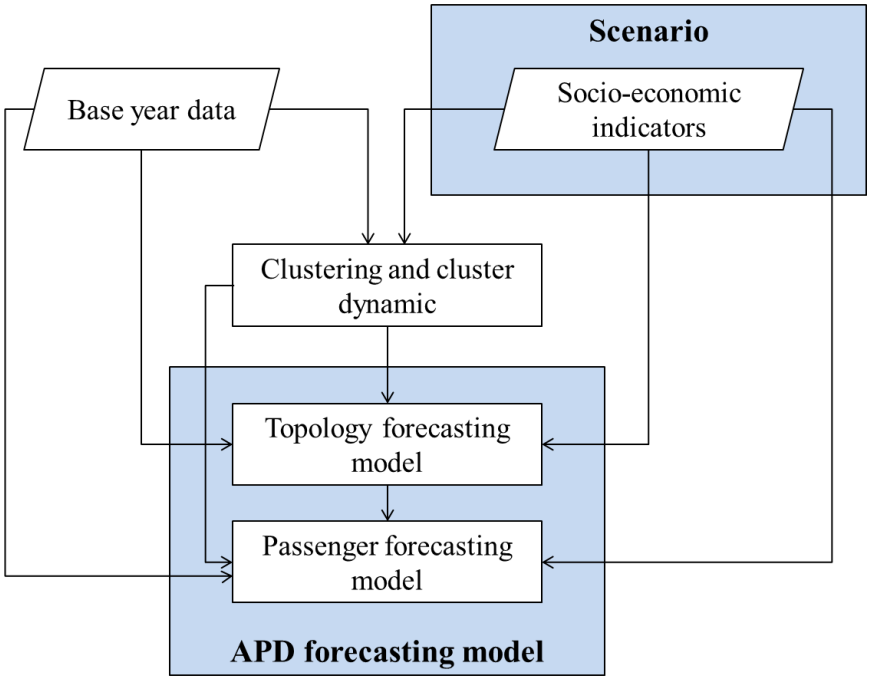


Fig.4.1. General approach of forecasting origin-destination air passenger demand between settlements worldwide based on socio-economic indicators

The general representation of the method is depicted in Fig.4.1. The settlement clustering process uses the settlement's socio-economic indicators to separate them into groups in the base year y_0 and defines a future content of clusters for every year of the socio-economic scenario – the cluster dynamics. The topology forecasting model has the following input information: settlement clusters, socio-economic indicators of settlements in the base year and the scenario inputs from y_1 to y_t. The output of the topology forecast model is a modeled configuration of the APD network for every discrete slice in the scenario. Using settlement socio-economic indicators

from a scenario and the topology forecast model, the passenger forecasting model computes passenger numbers from y_1 to y_t for every settlement pair of the modeled APD network configuration.

4.2. Description of sub-models

This section describes each model shown in Section 4.1. The input variables to the APD forecasting model are described in Sub-section 4.2.1. Sub-section 4.2.2 demonstrates the clustering approach and the cluster dynamics. The topology forecasting model is shown in Sub-section 4.2.4. Sub-section 4.2.5 describes the passenger forecasting model.

4.2.1. Inputs into the APD model

As shown in existing studies (Belobaba et al., 2009; Dray, 2010; Airbus, 2014; Boeing, 2013; Dray, 2014), the APD has a clear correlation with economic and social indicators. Thereby, in order to estimate future APD between settlements, the development of these indicators has to be known. Thus, external scenarios of the future development of socio-economic indicators are considered as input for the APD forecasting model. In order to obtain an accurate APD forecast, the scenario should contain a consistent set of numerical indicators over time, describing the future development path. Moreover, since the modeling approach is considered at settlement level, the scenario has to contain variables at settlement level respectively. In other words, the scenario could be presented as a database, where a set of variables for each time segment in a given time period for a required settlement number are provided. The scenario variables are the input for each sub-model of the APD forecasting model: clustering and cluster dynamics (described in Sub-section 4.2.2), the topology forecast model (Sub-section 4.2.3) and the passenger forecast model (Sub-section 4.2.4).

In addition, the APD forecasting model should take airfares into account, since they have an undoubted impact on the APD (Boeing, 2015): the lower the airfare between settlements, the more passengers would travel and vice versa. The average airfare is the input for the topology forecasting model (Sub-section 4.2.3) and the passenger forecasting model (Sub-section 4.2.4).

4.2.2. Clustering

This sub-section describes the clustering and cluster dynamics from the APD forecasting model. As shown by Zheleva et al (2012), for the topology forecast, node community (groups of settlements for the purpose of this study) features improve link prediction performance (Lü and Zhou, 2011). In other words, settlement partition into groups increases the accuracy of the APD topology forecast, as shown by Terekhov et al. (2016). Furthermore, it is likely that the APD generation process varies depending on the settlements. For instance, the APD formation in rich, highly developed settlements and poor settlements from rural areas has different mechanisms where various reasons and possibilities for travel depend on income levels and the number of inhabitants. Therefore, these settlements could be allocated to a number of groups according to their socio-economic indicators, where settlements in each group possess similar patterns. Thus, clustering defines qualitative and quantitative features of these groups in the base year (the starting point of forecasting) and the dynamic of settlements changing groups.

In addition, the socio-economic indicators of the settlements change over the forecast period. These changes affect the probability of membership of a given city to certain clusters. This reveals the settlement distribution changes within the clusters over time. Thus, this process is called "cluster dynamics" (Terekhov and Gollnick., 2015). Cluster dynamics calculates the probability that a given element (settlement) will appear within a given cluster at a particular point of time. This approach shows how the settlements are allocated to the various clusters for any forecast year based on socio-economic settlement indicators. During the forecast period, cluster centers remain fixed as for a base year and do not change. In other words, in this study, affiliation calculations are made from a base year perspective.

In contrast to clustering, cluster dynamics is based on discrete time series data from socio-economic settlement indicators. Cluster dynamics which use the cluster centers from the base year define the allocation of the settlements in clusters for every interval of the discrete time series data. An abstract clustering and cluster dynamics example is shown in Fig. 4.2. A settlement set is given in the base year T. For instance, the settlements which use clustering are assigned to four clusters: *Cluster A*, *Cluster B*, *Cluster C* and *Cluster D*. A socio-economic scenario from T to $T+2k$ with time interval k is given with known indicators in time intervals $T+k$

and *T+2k*. It is assumed that settlements in this scenario do not change their socio-economic indicators, except city *X*. Within the scenario, the cluster centers remain fixed and do not change. The allocation of city *X* is defined for every time interval in the scenario. Based on socio-economic indicators of city *X* in time interval *T+k*, this city changes from *Cluster A* to *Cluster C*. Since the socio-economic indicators of city *X* are changed in time interval *T+2k*, this city is assigned to *Cluster D*. Thus, the definition of city *X* allocation in time intervals *T+k* and *T+2k* is the cluster dynamics which is based on the clustering in the base year in time *T*.

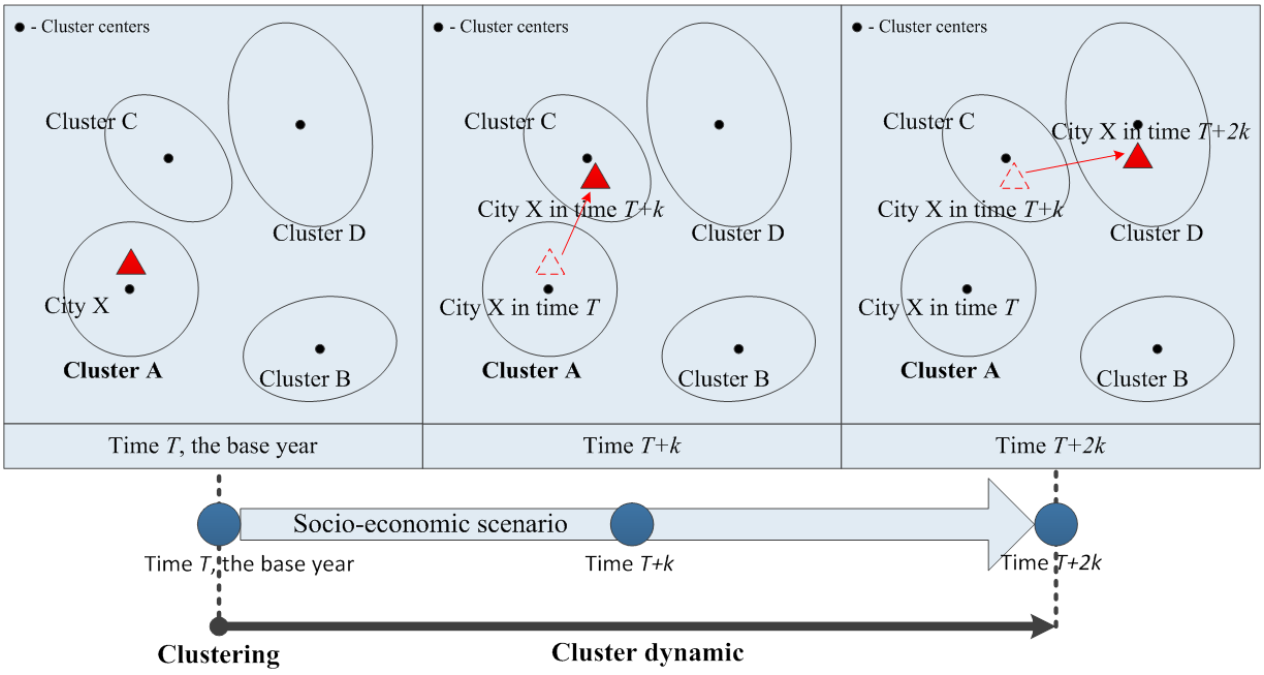


Fig.4.2. Clustering and cluster dynamics

Clustering is an important part of the APD forecasting model. Clustering results help to understand the APD generation within and between different settlement groups. Clustering results are utilized in the topology forecasting model and the passenger forecasting model. In the topology forecasting model, clustering has a positive significant impact on link prediction accuracy in the APD network (Terekhov et al., 2016), defining boundary conditions of connections adding and elimination processes for every cluster pair. The clustering algorithm is described in detail in Section 5.1.

4.2.3. Topology forecasting model

This sub-section presents a general description of the topology forecasting model. The model defines the future APD network based on the topology from the previous time interval and socio-economic indicators from a given scenario.

The socio-economic characteristics of settlements change throughout the duration of the scenario. Therefore, based on the scenario, the original APD network topology transforms, as a result of new APD connections appearing and disappearing. These topology changes must be taken into account when determining the APD between settlements. Furthermore, the number of settlements with airports changes throughout the duration of the scenario. A settlement is added to or eliminated from the demand network when an airport or airports appears or disappears.

The APD network could be described as a network where settlements are represented as nodes and settlement pairs as connections. Thus, network theory could be applied (Newman, 2003) and the problem of forecasting the potential APD between settlement pairs turns to a link prediction between nodes in a given year of a socio-economic scenario. In addition, the ADP network is a weighted network. The weight could be presented as a single indicator (number of passengers on a connection, a ticket price on a connection, distances between settlements, etc.) or as a combination of indicators. Since the APD has a clear correlation with the socio-economic indicators (Belobaba et al., 2009; Dray, 2010; Boeing, 2013; Airbus, 2014; Dray, 2014), the weight could be presented as a combination of socio-economic settlement indicators. This combination could be interpreted as an 'attractive force' between settlements in pairs and can be presented as a gravity model. The general weight representation w between settlements *x* and *y* is shown in Eq. 4.1., where Φ is the gravity model and a and b are indicators of settlements *x* and *y* respectively:

$$w_{xy} = \Phi(a_1, b_1, \dots, a_n, b_n), \; n = 1 \dots m \qquad \text{(Eq. 4.1)}$$

Thus, in this thesis, the topology forecasting model simulates the changes in the APD network topology using weights. The APD network is presented as a set of nodes linked by weighted connections, where weight is the combination of settlements indicators. The changes in topology are based on the socio-economic scenario. Changes in socio-economic indicators lead to

the connection weight changes. This means that the attractive force between settlements changes accordingly: the higher the socio-economic parameters, the stronger the attractive force and vice versa. The changing attractive force leads to the APD topology evolution. The new connections could appear in the APD network when the attractive force between settlements is strong enough and disappear when the force is quite weak. However, the terms "strong" and "weak" are relative and could vary significantly depending on the settlement in terms of their socio-economic indicators. Since settlements possess diverse socio-economic characteristics, the APD connection process generation is correspondingly different per settlement type. Based on clusters obtained from Sub-section 4.2.2, settlements could be assigned to a number of groups according to their socio-economic indicators. This helps to distinguish settlement groups from each other and give a clear definition of "strong" and "weak" for connection weights between settlements in these groups.

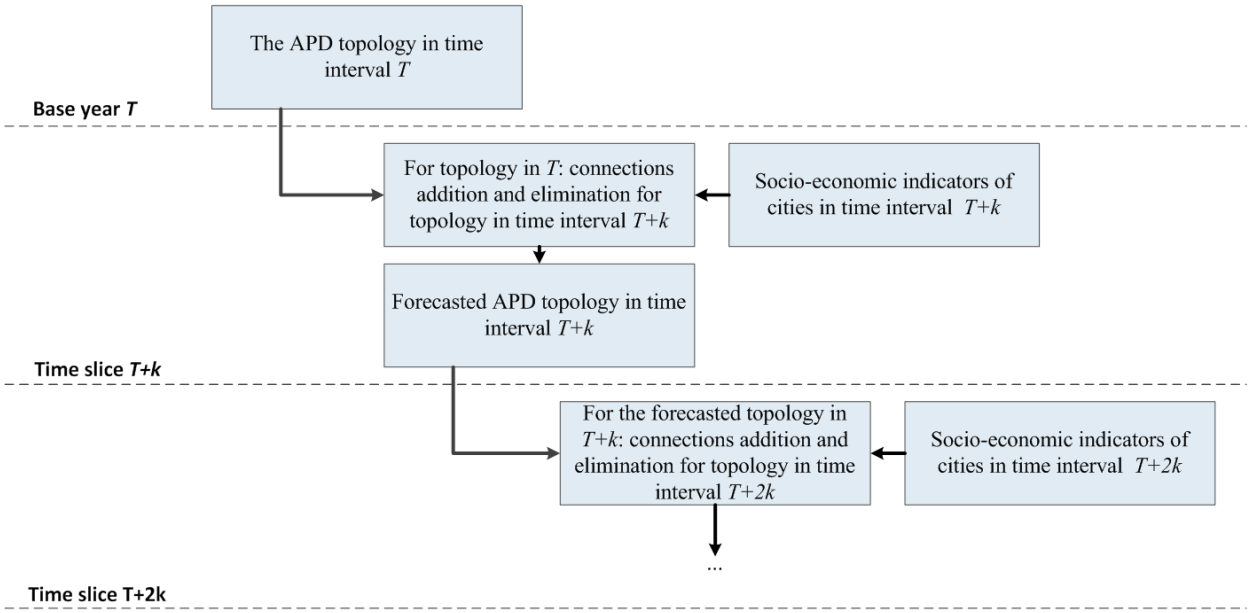


Fig.4.3. Topology forecasting model framework.

The general topology modeling process is sequential: the forecast of year *T+k* is based on year *T*, the forecast of year *T+2k* is based on year *T+k* etc. The topology forecasting model's basic structure is shown in Fig. 4.3. In the base year *T*, the actual APD topology and clustering is known. The socio-economic scenario is provided from the base year *T* to the last year in the scenario *N* with time intervals *k*. The socio-economic indicators from the scenario of year *T+k*

are assigned to the settlements from the APD network of year *T*. Thus, this network is incomplete for the year *T+k*. Then, the topology forecasting model, based on the network topology and the new settlement socio-economic indicators, defines which connections should be added into the network and which connections should be eliminated. After this process, the topology for the year *T+k* is obtained. This process is repeated analogically for year *T+2k* and so on, up to the last year of the scenario. Following this procedure, the topology for every time interval is obtained. The next step is to define the passenger number on every connection for every given topology. The topology forecasting model is described in detail in Section 5.2.

4.2.4. Passenger forecasting model

The APD forecast between global settlements pairs (Terekhov et al, 2015-1) consists of a sequential set of discrete intervals by years at the time scale up to the forecast horizon. When the APD network topologies are determined for the forecast period, the air passenger number on settlement pairs has to be defined. Since the APD has a clear correlation with socio-economic parameters, as mentioned above, the air travel passenger number estimation on settlement pairs is conducted based on socio-economic indicators.

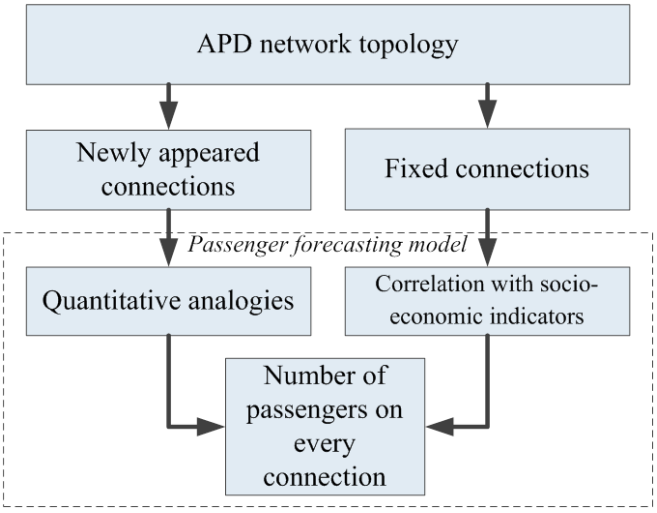


Fig.4.4. The passenger forecasting model basic framework

The passenger forecasting model basic framework is depicted in Fig.4.4. The forecasted APD networks contain two connection types: newly-appeared connections and connections

which are still fixed from the previous APD network. Since there is a clear correlation of socio-economic settlement indicators and the passenger number on connections, the passenger number for fixed connections is defined by socio-economic indicator growth correlation from settlement pairs in year *T+k* and year *T*. However, for the newly-added connections, this approach does not work due to the absence of these connections in the previous time interval. Thus, the passenger number for newly-appearing connections in year *T+k* which does not exist in base year *T* is defined by assigning the passenger numbers of similar connections in base year *T*. Since the base year is the only one year in the model with the actual data set, the analogy connections for the modeled time intervals are considered from the base year. The connection similarity could be characterized by socio-economic settlement and spatial closeness of connections in the base year and the forecasted year. Thereby, the number of passengers on a given settlement pair in year *T+k* is defined by searching for the closest set of conditions of a settlement pair in the previous year *T*. In other words, the number of passengers is defined using the Quantitative Analogies (QA) method for new connections of the year *T+1* and set of connections in the previous year *T*. In general, the QA approach requires an expert who has to identify the situation and find out the analogies for the newly-appeared connection in the analogies set in the base year data. However, this qualitative concept could be presented as a quantitative approach. A simple expert model can be developed in order to identify the set of analogies and the analogy for the newly-appeared connection in the base year and, thereby, determine the passenger number on this connection. The model's mathematical interpretation is presented in Eq. 4.2 as a calculation of the shortest distance between settlement pairs in a forecast year and settlement pairs in the base year.

$$\min d(x,y) = \sqrt{\left(X_1 - Y_{1,1}\right)^2 + \cdots + \left(X_n - Y_{k,n}\right)^2} \qquad \text{(Eq.4.2)}$$

Where X_n represents the condition *n* of settlement pair x in a given year within the forecast period, $Y_{k,n}$ represents condition *n* of settlement pairs Y_k in the base year, and *k* is the number of connections in the base year. An example of QA is depicted in Fig.4.5. Connection *X* in time segment T+1 and five connections *Y* are shown in the base year. Using Eq.4.2, the distances in multidimensional space (in Fig.4.5 two detentions are shown) between connection *X* and all connections *Y* could be defined. As could be seen, the minimum distance is between connection *X* and Y_5. This means that a newly-appeared connection X in time interval *T+k* which does not

exist in year *T*, is the most similar by the socio-economic indicators to connection Y_5 in the base year. Thus, the passenger number on connection *X* is assumed to be equal to the passenger number on connection in the base year Y_5.

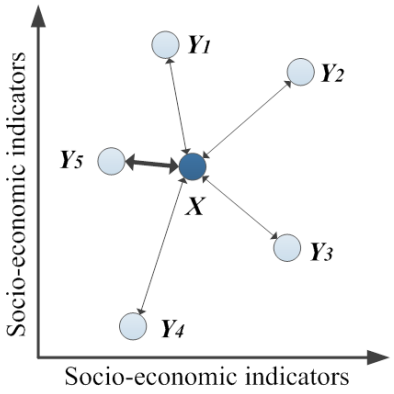


Fig.4.5. An example of the QA approach

In addition, the passenger generation between settlements is different in different settlement groups. Thus, based on clustering described in Sub-section 4.2.2, the settlement affiliations to clusters for every time interval are known. Thereby, for every newly-appeared connection in the network, cluster pairs based on settlements clusters are also known. In order to define the passenger number on a newly appeared connection in time interval $T+k$ in cluster pair *C*, the analogy by socio-economic indicators is searched within connections in the base year in cluster *C*.

For fixed connections in the APD network, correlation between passenger growth and socio-economic indicators growth is used. The number of passengers on this type of connection is defined as the number of passengers on a connection in year *T* plus passenger growth equal to the socio-economic indicators growth between years $T+k$ and *T*. Thereby, the number of passengers is calculated for all connections in every APD network within the forecast period. The passenger forecasting model is described in detail in Section 5.3.

4.3. Conclusion

This chapter presented the APD forecasting model concept. Based on the socio-economic scenarios, the proposed model simulates the future APD network topology and passenger number on all APD connections.

The introduced modeling framework was developed in order to cover the first step of the proposed research methodology, *Model development* (see Section 3.2), and objective which is indicated in Section 3.1. The next step in this thesis is to describe the proposed model in detail and validate it based on the actual data.

5. APD forecasting model development and validation

This Chapter presents detailed descriptions of the APD sub-models and their validation. In order to validate the proposed methodology described in Chapter 4, a special validation framework has been developed. This framework includes the preparation module and the validation module, as demonstrated in Fig. 5.1.

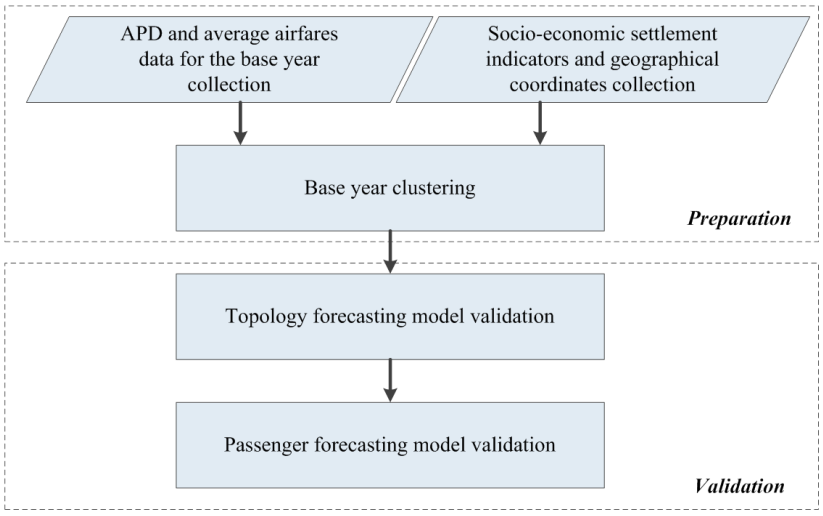
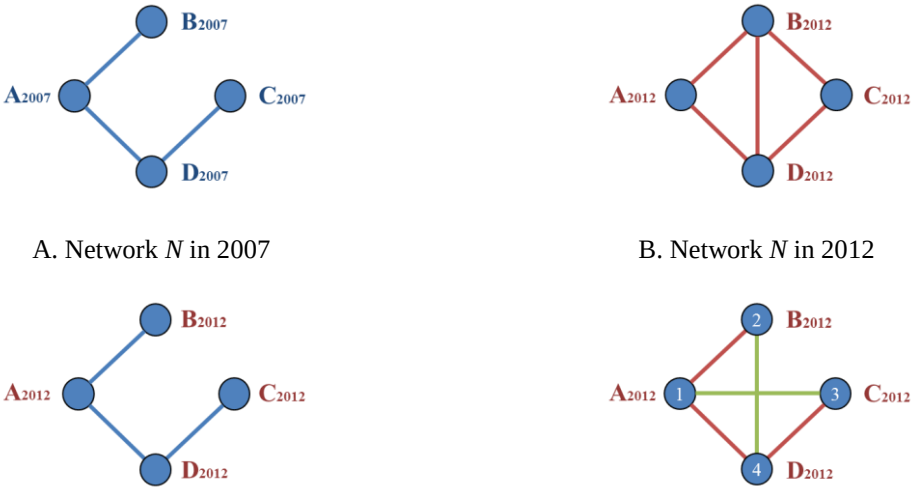


Fig.5.1. The validation framework

The preparation module aims to collect necessary data sets and generate the required inputs for APD model validation. The basis for the validation process is the base year, which, in this thesis, is the year where all required information for the validation process is known. The base year contains the following information: the APD on settlement pairs worldwide, socio-economic settlement indicators, geographical settlement coordinates and average airfares on the settlement

pairs. The validation of the topology forecasting model and the passenger forecasting model are conducted on the basis of the preparation module. The validation procedure is as follows: socio-economic indicators and cluster affiliation of the base year are assigned to settlements in year *X*, which was relative to the base year in the past. Thus, the APD network of year *X* becomes an incomplete network of the APD network in the base year. For the model validation, simulated APD connections and the passenger number in the base year from year *X* are compared with the actual real data in the base year. For example, four nodes and network topologies are known for 2012 (the base year) and 2007 (year X), as demonstrated in Fig.5.2B and Fig.5.2A respectively. The 2012 node attributes are assigned to the 2007 network nodes (Fig.5.2C). Thus, the network topology in Fig.5.2C is an incomplete 2012 topology where two connections are missing: $B_{2012}C_{2012}$ and $B_{2012}D_{2012}$. New connections are then simulated based on the incomplete network and known 2012 node attributes (Fig.5.2D). The modeled 2012 topology from 2007 in Fig.5.2D is compared to the actual 2012 topology. Therefore, the accuracy of the method can be obtained. In addition, since the accuracy assessed combining two models is based on different approaches, the measures for predicting accuracy, such as the mean absolute percentage error (MAPE) or the symmetric mean absolute percentage error (SMAPE), which are mainly used for the regression models assessment, are not applicable in the case of this study.



A. Network *N* in 2007

B. Network *N* in 2012

C. Network *N* with 2007 topology and node attributes from 2012

D. Forecasted 2012 network *N* from 2007, where the green connections are modeled

Fig.5.2. The validation approach

The basic principle for the method validation is depicted in Fig. 5.3. The APD network topology on settlement level in year *X* is known as well as the topology in the base year. The conditions required for the appearance of new settlements (new nodes) in the APD network are not clear and difficult to predict (Lü and Zhou, 2011). Thus, sets of settlements from two networks are reviewed. Settlements which are presented in these networks are allocated to the set of common settlements. Thereby, there is a constant set of common settlements for both networks. These settlements are allocated to a number of groups based on their socio-economic indicators in the base year using the clustering. Then, the socio-economic indicators and the cluster affiliation are assigned to the year *X* network.

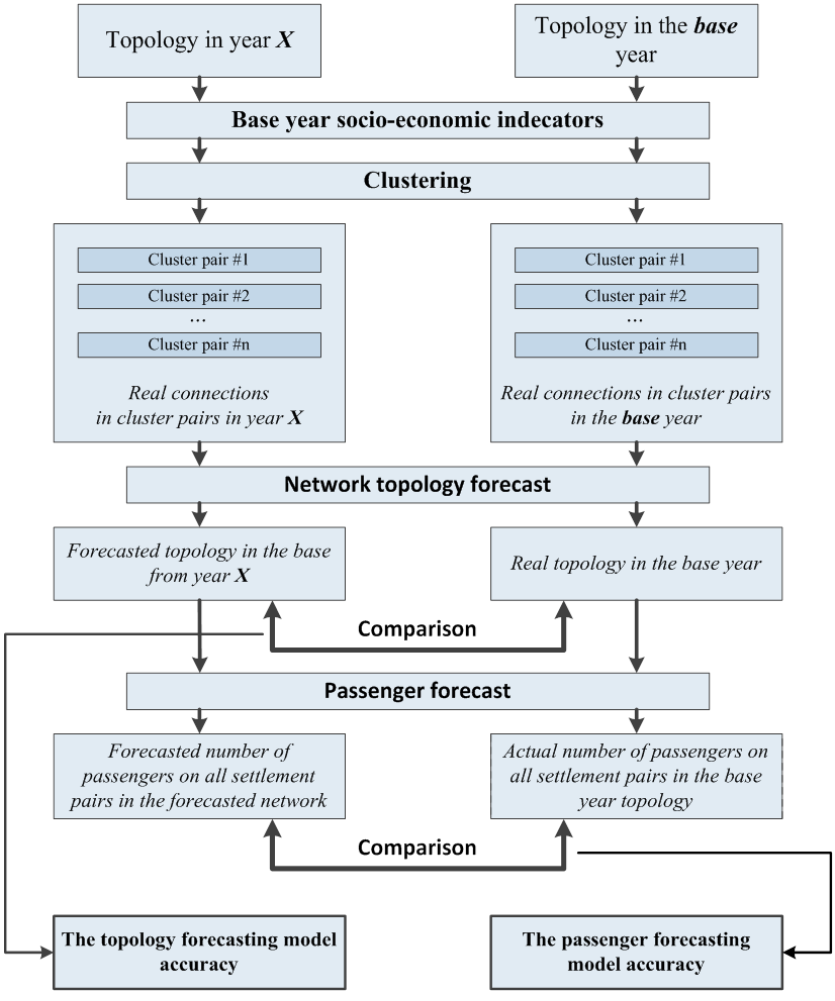


Fig.5.3. The basic principle for the forecasting method validation

The network topology model is applied to every defined cluster pair in year X. This forecast model simulates the base year APD topology from year *X* and adds new connections and eliminates old connections in year *X* APD network. The modeled connections from year *X* to the base year from every cluster pair are compared to the actual real connection number in the base year. Then, the passenger number in the modeled network on every settlement pair is calculated using the passenger forecasting model. The simulated passenger number from year X to the base year is compared to the actual passenger number in the base year. Thus, the accuracies of two models, and, therefore, the APD forecasting model, are assessed by comparing modeled results and actual data.

The main source for APD data worldwide between settlements (topology) is the Sabre Airport Data Intelligence (ADI) database (Sabre Airline Solutions, 2014). The database includes annual data on the APD and the average airfares between airports worldwide. The ADI database contains booking information from the Global Distribution System (GDS), its primary data source, and other external data sources. In addition, the ADI database contains two quality sets: preliminary data and final data. The preliminary data starts from year 2002 and the final data starts from 2009. Although in the ADI final data mistakes are corrected and it is more accurate, the ADI preliminary data cover more years. The detailed analysis of the final and preliminary data is described by Ghosh and Terekhov (2015). When the study started, the latest available data from the final set were for 2012. Thus, 2012 has been chosen as the base year for this thesis, due to data accessibility. The method is validated on real world data for 2012 from 2002, 2007 and 2011. These years cover intervals of one, five and ten years respectively. This allows the APD forecast model to be assessed on a short, middle and long forecast time frame.

Since the APD forms in populated areas and not in airports as it is shown above, the APD data between airports is modified to the APD between settlements associated with these airports. In the case that there are a few airports in a settlement, their APD is merged to this settlement. Thus, the set of common settlements for the validation process contains more than 3,600 settlements from more than 190 countries worldwide. In addition, the APD between settlements is considered undirected for this study.

In order to define weights in the APD network, the socio-economic indicators have to be defined. Due to a large number of settlements in the model, a detailed set of economic and social indicators for the settlements requires a manual search and in most cases this information is not accessible. Thus, as socio-economic indicators, settlement GDP and population are chosen based on previous studies (Grosche et al, 2007) and database accessibility. Thus, for the base year, for every settlement, the population is obtained using (UN, 2014-1; MaxMind, 2014), GDP data collected from (UN, 2014-2; The World Bank, 2014). Settlement GDP is calculated using GPD per capita in a country. GDP per capita is defined based on country GDP and country population obtained from (UN, 2014-1; The World Bank, 2014). Since some databases adjust their economic indicators to 2005 US dollars and in order to have homogeneous economic data, all economic indicators in this study are adjusted to 2005 US dollars. Geographical settlement coordinates are collected from Our Airports (2014) and OpenFlights (2014).

This Chapter is organized as follows: Section 5.1 justifies and describes the clustering method. Section 5.2 describes and validates the topology forecasting model versus the actual 2012 APD topology, thus providing the model accuracy. Section 5.3 presents the passenger forecasting model and demonstrates the assessed accuracy. Section 5.4 shows the overall APD forecasting model accuracy and the propagation error assessment for the short, middle and long forecasting terms. Section 5.5 provides conclusions to this chapter.

5.1. Clustering

As shown in existing studies (Zheleva et al, 2012; Lü and Zhou, 2011), a partition of elements into groups improves link prediction performance and, thereby, increases the accuracy of the APD topology forecast between settlements. Furthermore, studies (Belobaba et al., 2009; Dray, 2010; Airbus, 2014; Boeing, 2013; Dray, 2014) show that the APD has a clear correlation with economic and social indicators. Thus, it is likely that the process of the APD generation is different for different settlements. These settlements can therefore be allocated to a number of groups by their socio-economic indicators, where settlements in each group possess similar patterns. The goal of the clustering is to determine a finite set of groups (clusters) to describe a

dataset according to similarities among its elements (Hruschka et al., 2009; Berkhin, 2006). This allows appropriate methods to forecast the APD for each cluster pair to be determined and, thereby, for the accuracy of the whole APD forecast method to increase.

5.1.1. Clustering methods

There are various clustering methods which can be divided into two main groups: hierarchical and partitional. Hierarchical clustering (HC) unites clusters based on their proximity and forms a hierarchical tree (Xu and Wunsch, 2005). At first, HC assumes every element is a cluster. Then, the two nearest clusters are combined and assumed to be a new cluster. This procedure continues until there is only one cluster containing all elements (SAS Institute Inc., 2014). The result is a hierarchical tree of clusters, also known as a dendrogram (Fig.5.4). Thus, HC builds a system of nested clusters instead of one partition into disjointed clusters. Using this method, clusters could be retrieved by cutting the dendrogram at different levels. However, HC is appropriate for small sets of data, up to several thousand elements. The method is very sensitive to noise and outliers in data. Furthermore, HC algorithms are not capable of correcting possible previous misclassifications. Once an object is assigned to a cluster, it will not be considered again (Xu and Wunsch, 2005). Moreover, HC does not work well in overlapping areas (SAS Institute Inc., 2014) (in these areas, elements from several clusters share the same space).

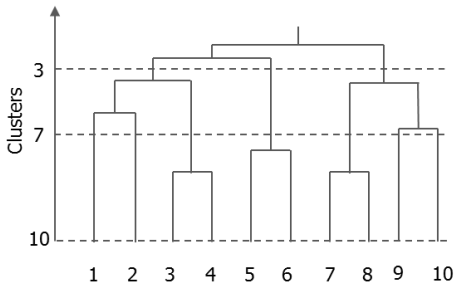


Fig.5.4. Hierarchical clustering

In partition methods, two approaches can be highlighted: exclusive clustering (EC) and probabilistic clustering (PC). Both approaches require the number of clusters to be determined in advance. In the EC approach, elements only belong to certain clusters and cannot be included in others (hard clustering). One of the most commonly used algorithms in the EC approach is the *k-means* algorithm, based on a certain number of pre-defined clusters. The main idea of *k-means* is

to define means for every cluster (Fig.5.5). This algorithm selects the randomly chosen number of elements in the initial set, equal to the number of pre-defined clusters. These elements are assumed to be cluster means. This is an iterative process. The algorithm recalculates cluster means until a specified criterion is met. The affiliation to clusters is defined for every element in the set by defining the minimum distance between means and elements. The *k-means* algorithm is appropriate for large sets of data, up to hundreds of thousands of elements. However, the appropriate number of clusters is unknown and it is necessary to specify a number of clusters before starting the algorithm (SAS Institute Inc., 2014). The algorithm is sensitive to the selection of the initial partition (Jain et al., 1999). In addition, it does not work well in overlapping areas (SAS Institute Inc., 2014).

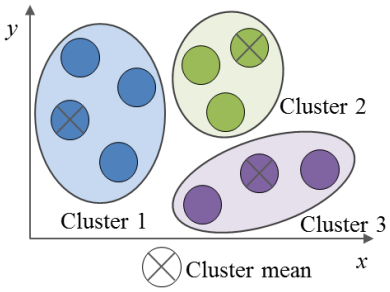


Fig.5.5. Exclusive clustering

In the PC approach, each cluster can be present as a parametric distribution. Thus, the initial set of elements is modeled by a mixture of these distributions. In contrast to k-means, where elements are deterministically assigned to one, and only one, cluster (hard clustering), the PC approach assigns elements to clusters with certain probabilities (soft clustering) (Fig.5.6).

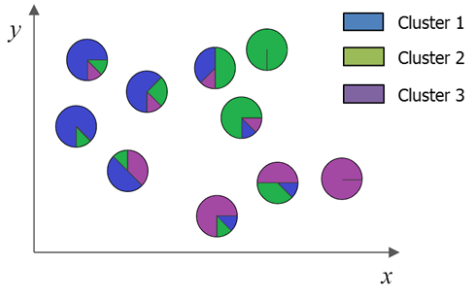


Fig.5.6. Probabilistic clustering

The most commonly used PC algorithm is a normal mixture or a mixture of Gaussians. The normal mixture algorithm is similar to k-means. The normal mixture uses expectation-maximization (EM) algorithm (Krishnan and McLachlan, 1997), where at the expectation step (E-step) expected values of the cluster membership for each element are calculated. Here, probabilities for all elements are calculated. Then, maximization step (M-step) recalculates the parameters of each Gaussian to maximize the probabilities found on E-step. These steps repeat until convergence. The normal mixture algorithm based on the probabilistic approach performs well in overlapping areas (SAS Institute Inc., 2014). However, it is sensitive to the selection of the initial partition (Xu and Wunsch, 2005).

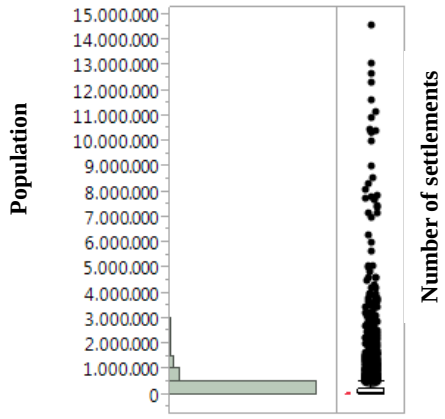


Fig.5.7. Settlement distribution by population

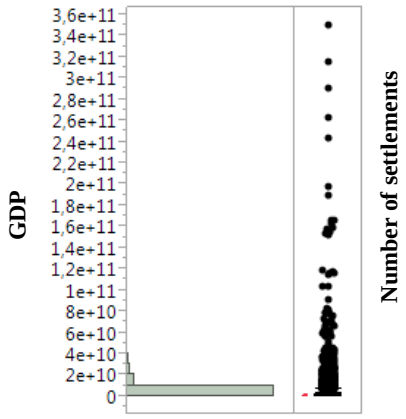


Fig.5.8. Settlement distribution by GDP

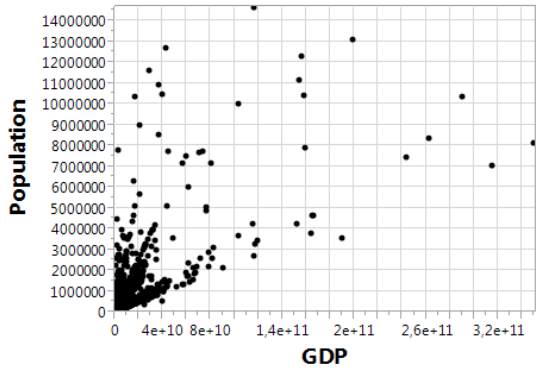


Fig.5.9. Settlement distribution by population and GDP

Percent of settlements	Quantiles	Population	Settlement
100.0%	Maximum	14,608,512	Shanghai, China
75.0%	Quartile	206,570	Annaba, Algeria
50.0%	Median	50,675	Mweka, DR Congo
25.0%	Quartile	7,716	Fort Dix, US
0.0%	Minimum	2	Portage Creek, US

Tab.5.1 Settlement quantiles by population

Percent of settlements	Quantiles	GDP (billions)	Settlement
100.0%	Maximum	350	New-York, US
75.0%	Quartile	3	Pekanbaru, Indonesia
50.0%	Median	0.744	Arcata, US
25.0%	Quartile	0.103	Lakselv, Norway
0.0%	Minimum	0.00007	Kadhdhoo, Maldives

Tab.5.2. Settlement quantiles by GDP (indicated here in constant 2005 US dollars)

The base year is 2012 which contains 4,435 settlements based on the APD network obtained from the ADI database. The next numerical attributes are obtained for every city: GDP and population in 2012 and geographical coordinates. Settlement distributions by population and GDP are presented in Fig.5.7 and Fig.5.8 respectively. Based on these distributions, settlements' quantiles by population and GDP are presented in Tab.5.1 and Tab.5.2 respectively. Settlements possess various socio-economic indicators, but they are not separated well, as can be seen in Fig.5.9. Most settlements are concentrated in areas of low GDP and population numbers. Thus, settlements in this area have similar GDP and population values. It is hard to understand which group they could be assigned to. As the overview above suggests, the PC is more effective in the overlapping areas, defining certain probabilities to clusters for every element. Using this approach, the cluster, the element is closest to, is seen and the changes in the probability proportions for a certain element are tracked. Thus, for clustering settlements in the APD model a PC algorithm of normal mixture is used. However, for this type of clustering it is necessary to define the appropriate number of clusters.

5.1.2. Normal mixture clustering application

The PC of normal mixture is chosen to group settlements into clusters. As mentioned, the APD generation process is likely to be different for different settlement clusters. Thus, it is important to define the appropriate clusters number as well as the settlement parameters number for clustering. Applying a few parameters could lead to high bias and missed opportunities for

cluster insight. Such clustering is not flexible enough to describe the sample well. In contrast, clustering with too many parameters is not able to fit the observed data well, and is too closely tailored to it. Such models may generalize poorly (Dziak et al., 2012). The base year data contain settlements GDP and population. Based on these parameters, it is possible to add one more parameter – GDP per capita. As shown in existing studies (Jain et al.,1999; SAS Institute Inc., 2014; Xu and Wunsch, 2005), a larger number of parameters can help adequately recognize the pattern and accurately define cluster means. Therefore, GDP per capita allows normal mixture algorithm to describe clusters with higher precision. Thus, due to data restrictions, to define settlement group numbers with similar socio-economic indicators, clustering in this thesis is made by using city GDP, population and GDP per capita.

For the normal mixture, the number of clusters has to be set in advance. This is a typical issue for the normal mixture clustering approach (Jain et al., 1999; Xu and Wunsch, 2005; Berkhin, 2006). It is solved through measurements of standard metrics for different number of clusters. In this study two standard metrics are used: the Bayesian information criterion (BIC) (Schwarz, 1978) and the Akaike information criterion (AIC) (Akaike, 1974). Both these metrics are penalized-likelihood information criteria. BIC and AIC choose the model with a particular number of clusters which demonstrates the best penalized log-likelihood. BIC and AIC is a variation of a penalty weight A_n in the information criterion:

$$IC(k) = -2l + A_n p \qquad \text{(Eq.5.1)}$$

Where k is number of clusters; l is the log-likelihood; p is the number of parameters in the model. For AIC $A_n = 2$, and for BIC $A_n = \ln(n)$; n is the sample size. BIC and AIC penalize more for models with additional parameters. The penalty of BIC depends on the sample size and it is usually more "heavy" than AIC. The number of clusters n minimizing BIC and AIC is considered to be the optimal number of clusters for a given set. BIC and AIC for 4,435 settlements in the base year 2012 of the APD forecasting model is presented in Fig.5.10. Clustering of these settlements is made by their GDP, population and GDP per capita.

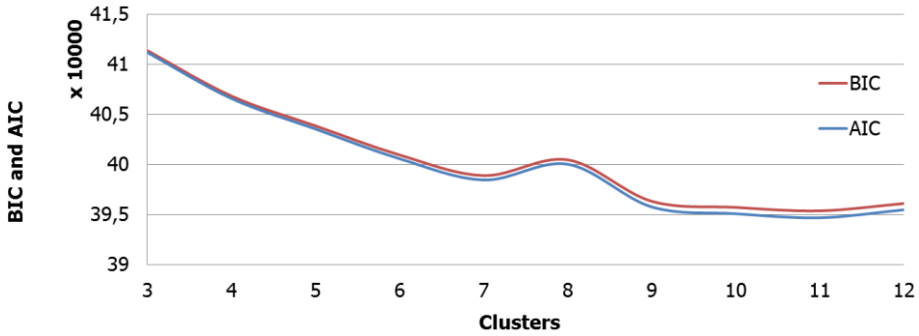


Fig.5.10. BIC and AIC metric for different cluster numbers for ADP forecast model settlement set

For clustering, 20 independent estimation process restarts with different starting values are used. This process prevents a local solution from being found. The maximum number of iterations for the EM algorithm's convergence stage is 200. The converge criterion is the difference in the likelihood at which the EM interactions stops and it is equal to10^{-8}. Based on the AIC and the BIC in Fig.5.10, separation into 11 clusters demonstrates best results. However, some means of these clusters are close to each other and interpreting their significance poses a challenge. Thus, three separations of the smallest AIC and BIC into 9, 10 and 11 clusters are considered. Cluster means of these separations are depicted in Fig.5.11 based on population and GDP per capita. As can be seen, the clustering algorithm detects groups of settlements with the largest socio-economic indicators every time. The main changes in separations are in settlement groups with populations of less than 1 million. The cluster means for these separations are depicted in Fig.5.12, Fig.5.13 and Fig.5.14 based on population and GDP per capita.

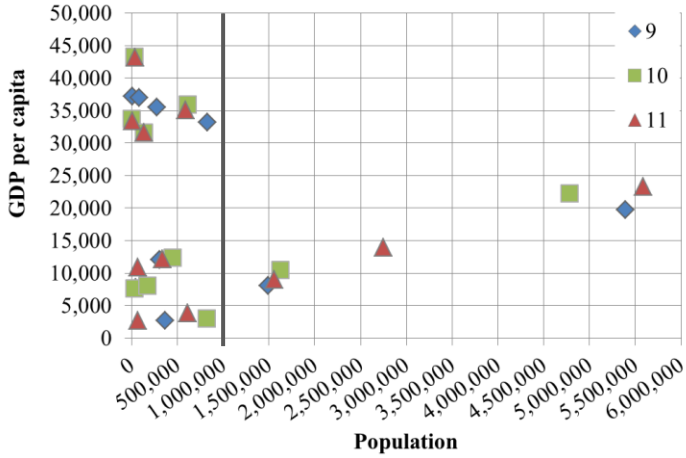


Fig.5.11. Cluster means for 9, 10 and 11 clusters by population and GDP per capita

For separating into 11 clusters, a few means are similar to each other. These settlement groups have relatively small populations with high GDP and GDP per capita. Furthermore, there are two proximate settlement groups of small settlements with low GDP and GDP per capita (Fig.5.12). The same proximity groups are generated when separating into 10 clusters (Fig.5.13). However, the situation is different for the 9 cluster separation (Fig.5.14) as these cluster means are easily distinguishable and their meaning is easy to interpret. Thus, separation to 9 clusters is chosen despite the fact that it does not demonstrate the best AIC and BIC.

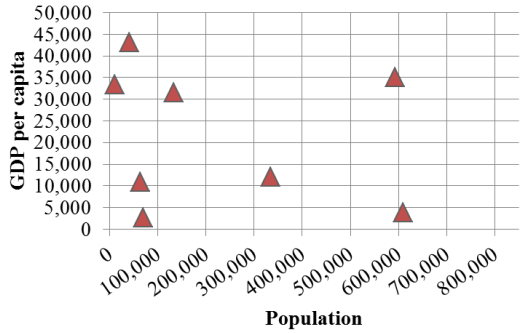


Fig.5.12. Cluster means for 11 clusters with settlements of less than 1 million inhabitants

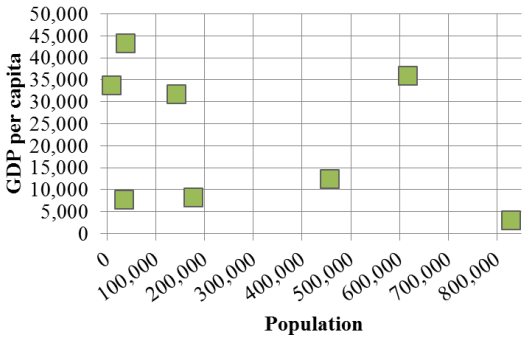


Fig.5.13. Cluster means for 10 clusters with settlements of less than 1 million inhabitants

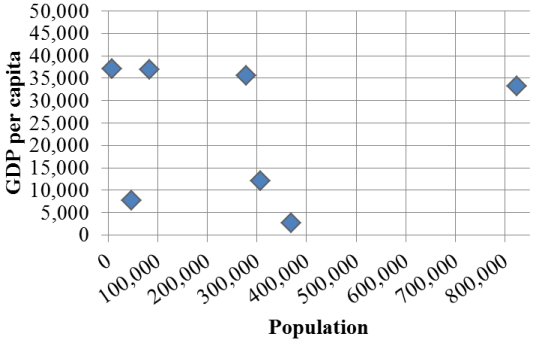


Fig.5.14. Cluster means for 9 clusters with settlements of less than 1 million inhabitants

All nine cluster centers are well separated. These clusters cover "small", "middle" and "large" sized settlements according to population and "poor", "middle" and "rich" settlements according to wealth. Based on these 9 clusters, the APD network in 2012 can be presented as a set of 45 cluster pairs[4]. For the purpose of the study, short-hand cluster names, derived from cluster means (population, GDP and per capita GDP), are adopted (i.e., *very small* and *rich* settlements, *small* and *poor* settlements, etc.). Tab.5.3 reflects the number of settlements in each cluster, cluster means and cluster names. For the PC of normal mixture, a complex formula is obtained to find the probabilities of every settlement affiliation to each cluster. Thus, by using this formula it is possible to retrieve settlement affiliation probabilities for various clusters for ongoing, developing socio-economic indicators, and tracking how settlement clusters change within a certain time period from the base year perspective (cluster dynamics). In other words, during the forecast period, cluster centers remain fixed, as in the base year. This process reveals the settlement distribution changes within the clusters over time.

Cluster #	Population	GDP, billions	GPD per capita	Settlement number in cluster	Proportion	Size	Wealth
1	8,520	0.3	37,134	1,453	0.32191	Very small	Rich
2	47,010	0.3	7,729	1,055	0.22774	Small	Poor
3	824,546	27	33,219	108	0.02487	Large	Rich
4	307,440	3	12,066	417	0.09684	Middle	Middle
5	5,394,129	77	19,767	76	0.01748	Megacities	
6	82,790	2	37,010	565	0.13312	Small	Rich
7	1,493,549	11	8,032	238	0.05451	Large	Poor
8	278,644	9	35,547	207	0.04738	Middle	Rich
9	369,340	1	2,744	316	0.07615	Middle	Poor

Tab.5.3. Cluster centers, settlement distribution among clusters and cluster names. GDP and GDP per capita indicated here in constant 2005 US dollars

5.1.3. Conclusion

This sub-chapter demonstrates the qualitative and quantitative features of different settlement groups in the base year 2012. It presents methods for grouping settlements into clusters according to their socio-economic indicators and the possibility of tracing changes in the cluster content within a socio-economic scenario. For settlement grouping, three main clustering

[4] The number of cluster pairs is defined as follows: $p = 0.5*n*(n-1) + n$, where p is the number of pairs and n is the number of clusters. Clusters here are interconnected. Therefore, 9 clusters form 45 cluster pairs.

approaches are considered: hierarchical, exclusive and probabilistic. After analyzing the advantages and disadvantages of these approaches, probabilistic clustering of normal mixture is chosen to separate settlements for the APD forecasting model as it performs better than others in overlapping areas. This is essential in the case of the APD forecast model, since many settlements have close socio-economic values. Clustering is based on socio-economic settlement indicators including settlement GDP and population. Thus, due to the data restrictions for the considered settlement number, the three parameters settlement GDP, population and GDP per capita are defined to fit the data observed. Using these parameters and special metrics such as AIC and BIC, separation into 9 clusters is chosen. Notwithstanding that the separation does not demonstrate the best AIC and BIC, cluster means are well distinguished and their meaning is easy to interpret.

Clustering is an important part of the APD forecasting model and the results are used within the whole study i.e. for the topology forecasting model and the passenger forecasting model. It is extremely likely that clustering increases the accuracy of the proposed APD model, and, specifically, accuracies for link prediction in the topology forecasting model. Moreover, clustering results can help to understand the APD generation within and between different groups of settlements in further studies.

5.2. Topology forecasting model

The APD network dynamically evolves over time. This network contains a number of settlements (nodes) with links between them. As previously mentioned, in this thesis, the APD network is considered as an undirected, weighted network (Terekhov et al, 2015-1) where each link is characterized by a parameter or a set of parameters. As shown, the APD has interdependences with economic and social indicators (Boeing, 2013; Dray et al, 2010). Thus, the weight of a link could be considered as a combination of socio-economic indicators between settlements in pairs. During the forecasting period, the socio-economic indicators of settlements vary. Therefore, the link weights also change. This variation over time has an impact on the APD network and, accordingly, the topology of the network is likely to change. For example, where the socio-economic settlement indicators (e.g. GDP, population and oil price) show a rapid

increase, it is likely that a number of connected settlements with a significant APD will appear where no APD connections previously existed.

There are three main link prediction method groups (Lü and Zhou, 2011) for forecasting connections in the network: similarity-based algorithms, maximum likelihood (ML) and probabilistic models (PM). Similarity-based algorithms are divided into local, global and quasi-local indexes (Lü and Zhou, 2011) and are the mainstream class of algorithms for link prediction. ML methods and PM are complex and very time consuming. ML is able to handle networks with up to a few thousand nodes in a reasonable time (Lü and Zhou, 2011). Furthermore, ML methods do not demonstrate the best accuracy (Lü and Zhou, 2011). PM performs well when a network contains different types of nodes and attributes related to them. However, adopting this approach is a long-winded and complicated process. Mostly, studies consider link prediction in non-weighted networks. Studies on link prediction in weighted networks are mainly conducted using weighted local similarity indexes (Murata and Moriyasu, 2011; Lü and Zhou, 2011). In addition, the APD network is a high-clustered network, as shown by Ghosh and Terekhov (2015). For highly-clustered networks, the common neighbor-based indexes demonstrate relatively good prediction with low complexity (Lü and Zhou, 2011). Thus, in this study, only weighted local similarity indexes are considered.

The underlying principle of weighted and non-weighted indexes of similarity-based algorithms is the same. These algorithms assign a score to each non-existing link in a given network. The links are then ranked in descending order according to their score and links with the highest score should appear in the network. Here, two significant problems arise as one index in the network can perform well where another fails (Lü and Zhou, 2011). Thus, the first problem is to define which weighted local similarity index shows the best performance in the APD network. The second problem is to define a criterion for adding new connections to the network with the highest score from the top of the ranking list. In other words, a boundary condition in the ranking list of non-existing links has to be defined: links from the ranking list between the first link and a boundary link will be added to the network.

In addition, as shown by Zheleva et al (2012), the combination of network structure, node attributes, and node community features improve link prediction performance. In the APD

network, the network structure and node attributes are known. For node communities, settlements are distributed into groups according to the proximity of their socio-economic indicators. For example, settlements with large GDP and population are assigned to the *large-rich* group and settlements with high population and small GDP are united to the *large-poor* group. Since settlements generally possess different socio-economic indicators in these groups (clusters) (Terekhov et al, 2015-1), the process of link appearance in each APD network cluster pair could be different. Thus, a similarity-based algorithm which shows the best performance in one cluster is probably different in another cluster. For example, different weighted similarity algorithms could perform better between *large-rich* settlements and *small-poor* settlements than between *megacities* and *middle-rich* cities. Furthermore, it is likely that every cluster pair has its own boundary. In this section, the performance of similarity-based algorithms for each cluster pair is analyzed. The boundary for each cluster pair is defined utilizing the algorithm with the best performance.

5.2.1. Weighted local similarity index identification

Two standard metrics are used to identify the appropriate index for each cluster pair: the area under the receiver operating curve (AUC) (Hanely and McNeil, 1982) and precision (Herlocker et al, 2004). In order to identify the most suitable weighted local similarity index, the data from the base year 2012 is used. Thus, here these metrics are applied to the 2009 and 2012 APD topology. For accuracies and boundary identification, a set of forecasts of the APD network is made: from 2002 to 2012, from 2007 to 2012 and from 2011 to 2012.

For 2002, 2007, 2011 and 2012 the APD networks and related data are obtained from the various databases mentioned above in Section 5.1. The initial set of 4,435 settlements from the base year 2012 obtained from the ADI data base is divided into 9 clusters, as mentioned in Section 5.1**.**

In the APD network, every cluster is defined as a set of settlements and weighted connections. These connections link settlements in one cluster with settlements in other clusters and settlements within a cluster. Weights in this thesis are considered as a combination of average airfare (Ghosh and Terekhov 2015), distance between settlements and main socio-

economic indicators such as settlement GDP and population. The weight on the connection between settlements *x* and *y* is presented as follows:

$$w_{xy} = (g_x * g_y)^{\alpha} * (p_x * p_y)^{\beta} * (l_{xy})^{\gamma} * (t_{xy})^{\delta} * \varepsilon + \theta \qquad \text{(Eq.5.1)}$$

Where $g_{x,y}$ are the GDPs of city *x* and *y*; $p_{x,y}$ are the populations of city *x* and *y*; $l_{x,y}$ is the distance between city *x* and *y*; $t_{x,y}$ is the average airfare between city *x* and *y*; $\alpha, \beta, \gamma, \delta$ are elasticities of GDP, population distance and average air fare respectively; ε is a dummy variable and θ is a free parameter. For this study, it is assumed that $\alpha = 1$, $\beta = 1$, $\gamma = -1$, $\delta = -1$, $\varepsilon = 1$ and $\theta = 0$ in order to avoid unnecessary complexity. Thus, Eq.5.1 turns to a variation of Newton's gravity model (Newton et al., 1833) and the weight could be interpreted as an abstract attractive force between settlements. Furthermore, the gravity model has been used in a number of studies (Grosche, 2007; Dray, 2014) to predict the APD between city pairs.

Based on assumptions in Eq.5.1 the weight between settlements *x* and *y* could be presented as:

$$w_{xy} = \frac{g_x * g_y * p_x * p_y}{t_{xy} * l_{xy}} \qquad \text{(Eq.5.2)}$$

Within this study, nine indexes of similarity-based algorithms are analyzed. Based Lü and Zhou' (2011) study, the weighted common neighbors (WCN), weighted Adamic-Adar index (WAA) and weighted resource allocation index (WRA) are applied to the APD network. In addition, similarity indexes for unweighted networks such the Salton index (Salton and McGill, 1983), Sorensen index (Sorensen, 1948), hub depressed index, hub promoted index (Ravasz et al., 2002), Leicht-Holme-Newman index (Leicht et al., 2006) and preferential attachment index (Barabasi and Albert, 1999) are adapted for weighted networks utilizing Murata and Moriyasu's (2007) proposed simple method integrating link weights into the nine indexes. These similarity indexes are presented in Tab.5.4.

Two standard metrics, the *AUC* (Hanely and McNeil, 1982; Lü and Zhou 2011) and *precision* (Herlocker et al, 2004; Lü and Zhou, 2011) are used to determine the accuracy of each index. Initially, for an undirected weighted network, all existing and non-existing links are known. From this set of existing links, a group of links – the probe set – is excluded and the

remaining existing links are the testing set. The score of each index in the network formed by the testing set is calculated for all non-existing links and the probe set.

The *AUC* shows the probability that a randomly chosen link from the probe set has a higher score than a randomly chosen link from the set of non-existing links. According to Lü and Zhou (2011) the *AUC* is as follows:

$$AUC = \frac{n' + 0{,}5 * n''}{n} \qquad (5.3)$$

n' shows how many times links from the probe set have a higher score then randomly chosen links from the non-existing links set. n'' denotes how many times links from the probe set have the same score as randomly chosen links from the non-existing links set. n is a number of independent comparisons.

Weighted common neighbors (WCN)	$s_{xy}^{WCN} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} (w(x,z) + w(z,y))$	(Eq.5.4)
Weighted Adamic-Adar (WAA)	$s_{xy}^{WAA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\log(1 + s(z))}$	(Eq.5.5)
Weighted Recourse Allocation (WRA)	$s_{xy}^{WRA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\mathrm{s(z)}}$	(Eq.5.6)
Weighted Salton index (WSA)	$s_{xy}^{WSA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\sqrt{s(x) * s(y)}}$	(Eq.5.7)
Weighted Sorensen index (WSO)	$s_{xy}^{WSO} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{2(\, w(x,z) + w(z,x)\,)}{\mathrm{s(x)} + \mathrm{s(y)}}$	(Eq.5.8)
Weighted hub promoted index (WHPI)	$s_{xy}^{WHPI} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\min\{\mathrm{s(x)}, \mathrm{s(y)}\}}$	(Eq.5.9)
Weighted hub depressed index (WHDP)	$s_{xy}^{WHDI} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{\max\{\mathrm{s(x)}, \mathrm{s(y)}\}}$	(Eq.5.10)
Weighted Leicht-Holme-Newman index (WLHN)	$s_{xy}^{\mathrm{WLHN}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(z,x)}{s(x) * s(y)}$	(Eq.5.11)
Weighed preferential attachment index (WPA)	$s_{xy}^{WPA} = s(x) * s(y)$	(Eq.5.12)

Tab.5.4 Weighted similarity-based algorithm indexes

For the *precision* metric, the set of probe links and non-existing links is categorized in descending order, according to their scores. From this list, the top-L links are selected as the predicted ones. Among these links, the L_r links are right (links from the probe set). The *precision* is a ratio of L_r to L. Thus, higher *precision* means higher prediction accuracy (Lü and Zhou, 2011). Both metrics are numbers between 0 and 1. The closer the metric is to 1, the better the performance of the index in a given network. In the equations in Tab.5.4, *w(x,z)* indicates the connection weight between settlements *x* and *z*; *s(z)* indicates the strength of settlement *z*. In other words, it is a sum of all connection weights of settlement *z*.

In this study for *AUC* and *precision* calculations, the 2012 and 2009 APD networks are used. These data are used for an objective picture of *AUC* and *precision* in different years. The year 2009 is chosen because the final data in the ADI database starts from 2009 and this year is the closest available year to the 2008 financial crisis, which had a negative impact on the APD. The year 2012 is chosen because this is the base year in this study. For 2009 and 2012, 3,918 and 4,435 settlements are obtained respectively. These settlements are allocated to 9 clusters according to their socio-economic indicators, applying the clustering approach in Section 5.1. Based on settlement clusters, 471,824 and 533,170 real connections in 2009 and 2012 are distributed between 45 cluster pairs. Non-existing links are obtained for each cluster pair. For the *AUC* and *precision* calculation, sets of existing and non-existing links are used.

Following existing studies (Lü and Zhou, 2010, 2011) the network is divided into two sets: testing and probe in 90% and 10% proportions, respectively. Each *AUC* and *precision* value is obtained by averaging 10 realizations with independent random separations of random and probe sets. The metrics for the whole network and each cluster pair are calculated for different indexes in addition to their standard deviations. The results for the whole network and the average of 45 cluster pairs' metric values are presented in Tab.5.5. The *AUC* and *precision* are used to determine the accuracy of each index for the whole network and clusters. The index with the best metrics values is chosen for the topology forecast in the APD network. The closer the metric is to 1, the better the performance of the index in a given network.

The data in Tab.5.5 demonstrates that only the weighted hub promoted index (WHPI) in 2009 and the WLHN in 2012 have a higher *precision* value in the whole network than the cluster

average (Fig.5.15 and Fig.5.17). However, these values are low compared to other indices. All other indexes show higher *AUC* and *precision* numbers in clusters than in the whole network. This proves the necessity of separating settlements into groups according to their socio-economic indicators, so as to improve link forecasting performance. The best *AUC* number for the whole network is WSO for 2009 and 2012 (Fig.5.16 and Fig.5.18) but this figure is lower than the *AUC* for WRA in clusters. The WRA index shows the best *AUC* and *precision* results in clusters pairs. This is expected, since the WRA gives a higher score to a non-existing connection between two nodes if these nodes have many common neighbors with large weights. It is important to note that the WRA index has the best *AUC* and *precision* performance in each cluster pair. This disproves the assumption that cluster pairs in the APD network have different similarity indexes demonstrating the best performance.

Based on the aforementioned analysis, the WRA index is chosen for the topology forecast in the APD network. The score for each non-existing link in each cluster pair is calculated using the WRA index. Next, it is necessary to validate the method based on historical data.

Metrics		Year	WCN[5]	WAA	WRA	WSA	WSO	WHPI	WHDI	WLHN	WPA
AUC	The whole network	2009	0.731	0.761	0.662	0.779	**0.852**	0.643	0.823	0.449	0.656
		2012	0.704	0.71	0.755	0.745	**0.786**	0.656	0.765	0.659	0.691
	Cluster average	2009	0.843	0.948	**0.964**	0.85	0.87	0.725	0.879	0.639	0.824
		2012	0.741	0.773	**0.806**	0.747	0.771	0.694	0.76	0.683	0.745
Precision	The whole network	2009	0.79	0.847	**0.912**	0.824	0.819	0.485	0.788	0.183	0.662
		2012	0.637	0.643	**0.687**	0.659	0.709	0.596	0.691	0.609	0.631
	Cluster average	2009	0.911	0.988	**0.991**	0.921	0.866	0.482	0.924	0.579	0.886
		2012	0.718	0.729	**0.78**	0.693	0.744	0.617	0.733	0.584	0.728

Tab.5.5. AUC and precision values for the whole APD network and average values for cluster pairs for 2009 and 2012

[5] The equations for the weighted similarity-based algorithm indexes are presented in Tab.5.4

2009

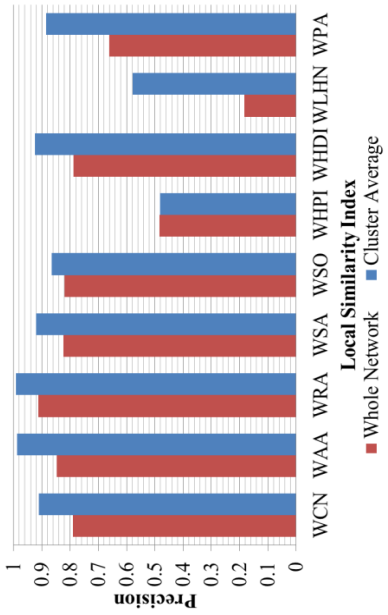


Fig.5.15. Precision metric for nine weighted indexes for the 2009 APD network

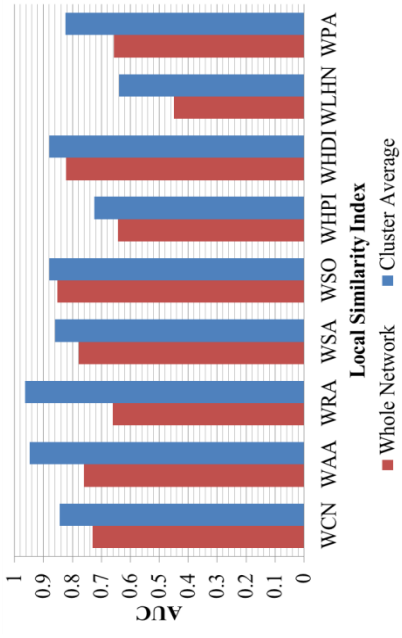


Fig.5.16. AUC metric for nine weighted indexes for the 2009 APD network

2012

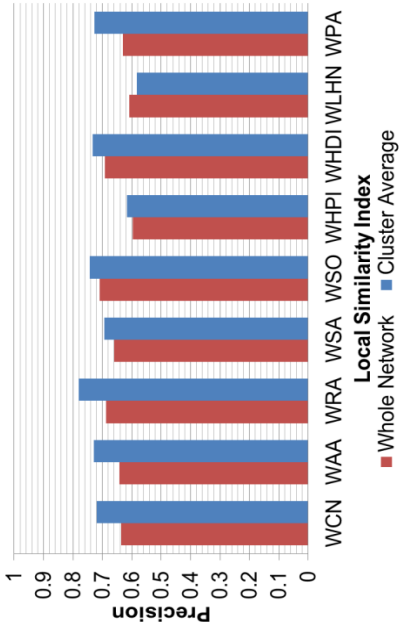


Fig.5.17. Precision metric for nine weighted indexes for the 2012 APD network

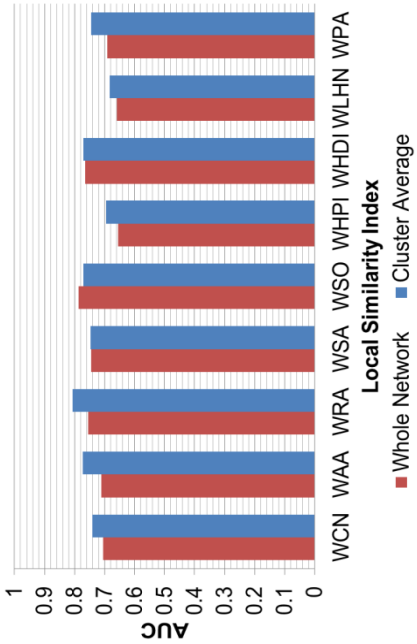


Fig.5.18. AUC metric for nine weighted indexes for the 2012 APD network

5.2.2. Analysis and verification of the similarity based algorithm using WRA

The APD topologies of four years, from 2002 to 2012, are used for the validation and data for these years from the ADI database (2002, 2007, 2011 and 2012 APD networks) are retrieved. Socio-economic data and geographical coordinates for settlements from the same databases as 2012 are obtained. The conditions required for the appearance of new settlements in the APD network are not clear and difficult to predict (Lü and Zhou, 2011). Thus, for the analysis, settlement sets from four networks are reviewed. Settlements which are presented in a given year and 2012 are allocated to the set of common settlements. Thereby, there are three sets of common settlements for all three networks. In Tab.5.6, topological characteristics of three networks with common settlements and connection numbers are presented.

Year	No. of common settlements in the 2012 APD network	No. of connections with common settlements	No. of non-existing connections with common settlements
2002	3,699	426,150	6,413,301
2010	3,896	500,020	7,087,440
2011	3,667	521,171	6,200,440

Tab.5.6. Topological characteristics of three APD networks with original and common settlements

Three analyses based on modified networks with common settlements to define the APD topology forecast accuracies are made. New connections are calculated within the analyses using the WRA index and then compared with real data. These APD connections are calculated for the following topologies: 2012 from 2002, 2012 from 2007 and 2012 from 2011.

For all three analyses, predicted APD topologies are compared with real APD topologies. For example, newly calculated APD connections in 2012 from 2002 are compared with the real APD topology from 2012. The analysis procedure is as follows: the 2012 socio-economic indicators and cluster affiliations are assigned to settlements in 2002. Thus, the 2002 APD network becomes an incomplete network variation of the 2012 APD network. The score for all non-existing connections in every 2002 network cluster pair is calculated using the WRA index. Connections are sorted in descending order according to their score. The calculated data are then compared to the real data. The real newly added connection number for every cluster is already known. Thus, from the calculated APD connections in the sorted

list, the same amount of connections is added to the modeled APD network. In other words, from the real 2002 to 2012 data in cluster *A*, for example, *x* new connections are added. This means that from the calculated sorted list in cluster *A*, the same *x* connection number is added from the top, assuming that the real newly added connections have a high score. All of these newly calculated connections from every cluster are added to the 2002 APD network, which forms an extended 2002 APD network. In addition, some of the connections are removed from the network because the socio-economic indicators change and APD between settlements disappear. The elimination process follows a similar procedure to the newly connection addition process. The score is calculated for every connection in every cluster pair from the extended 2002 APD network and then the connections are sorted in descending order as per their score in this second list. The real connection number in every cluster in 2012 is known. From the second sorted list, the same connection numbers as in the real connection numbers in 2012 is added to the final network. The remaining connections are eliminated from the APD network. Thus, the APD forecasting method has two sequential steps: the connection addition process and the connection elimination process. It is possible to define the accuracy of these processes. The connection addition accuracy is defined as a ratio between the amount of real newly added connections in 2012 and the number of real new connections in the sorted list. The connection elimination accuracy is defined as a ratio between the modeled eliminated APD connection number and the real eliminated APD connection number. This number is between 0 and 1. The addition and elimination processes have a higher accuracy the closer the ratio is to 1. Accuracies for the addition and elimination processes for every cluster pair for 2012 from years 2002, 2007 and 2011 are presented in Fig.5.19 and Fig.5.20 respectively. Cluster pairs are ordered in descending order according to accuracy in 2002. As can be seen in Fig.5.19, the accuracy for strong cluster pairs for the 10-year interval is high. It shows that for the strong clusters, the socio-economic indicators and the long interval of 10 years play a determining role in establishing new connections. The accuracy is not at the same level for the weak cluster pairs and shorter time intervals. This is most likely due to a number of random effects influencing APD which are a challenge to identify and quantify. In other words, these impacts have a greater influence on the newly added connection accuracy of weak cluster pairs than on strong cluster pairs.

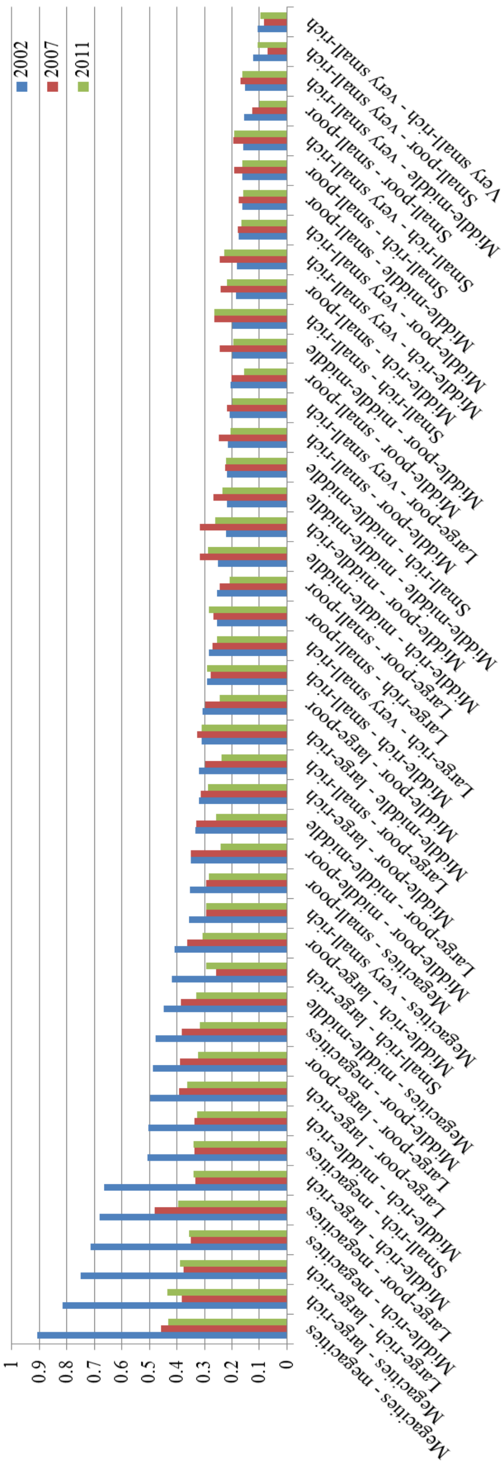


Fig.5.19. Accuracies for connection addition for every cluster pair in 2012 from 2002, 2007 and 2011

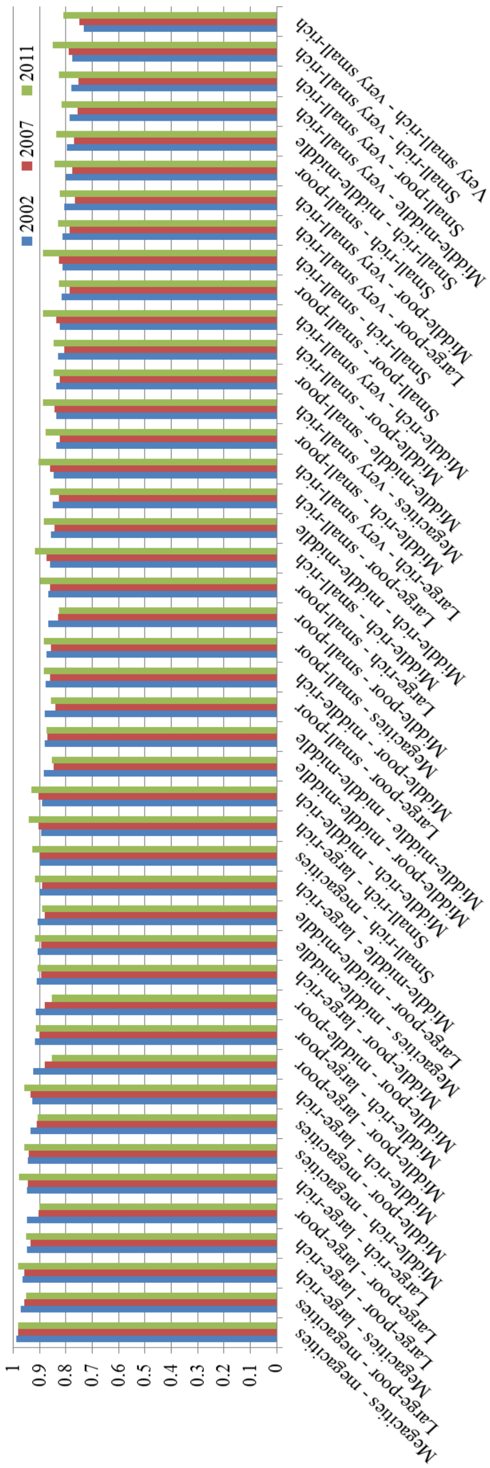


Fig.5.20. Accuracies for connection elimination for every cluster pair in 2012 from 2002, 2007 and 2011

In this study, all average accuracies for the modeled newly added connections from the years 2002, 2007 and 2011 to 2012 are below 50%, meaning that the prediction contains errors amounting to more than 50%. For connection elimination, average accuracies are above 50%, meaning that there is less than 50% error in the elimination prediction. The final average accuracy for all clusters of the APD network forecasts is above 0.6 and it is higher than in Murata and Moriyasu's (2007) study of link prediction in a weighted network of Question-Answering Bulletin Boards, for instance. The average accuracies for newly predicted connections, eliminated connections and the final accuracies of the forecasted 2012 ADP networks from 2002, 2007 and 2011 are presented in Tab.5.7. The final accuracy of every cluster pair for 2012 from 2002, 2007 and 2011 is shown in Fig.5.21.

After the addition and elimination processes, the forecasted APD networks of 2012 and the real APD network of 2012 are compared. This comparison shows that the forecasted networks have accuracies higher than 0.6 (Tab.5.7). The accuracy is high even when the average addition accuracies show poor results. The reason for the over 0.6 accuracy lies in the proposed approach when it models that more than 50% of connections in 2012 remain in the network from 2002, 2007 and 2011 APD networks and accuracies in most clusters are high (Fig.5.21).

	2002	**2007**	**2011**
Addition	0.344	0.283	0.261
Elimination	0.873	0.859	0.888
Total	0.682	0.748	0.812

Tab.5.7. Average accuracies for newly predicted connections, eliminated connections and the final accuracies of the forecasted 2012 ADP networks from 2002, 2007 and 2011

The total accuracy increases the shorter the time period between the base year and the forecasted year 2012. Although the accuracy results for cluster pairs are scattered, the accuracies for "strong" cluster pairs is higher in 2002. These are, in particular, higher in *megacities – megacities*, *large-poor – megacities*, etc. cluster pairs (see Fig.5.19 and Fig.5.20). This is probably related to the 2008 economic crisis and predictions from 2007 and 2011 to 2012 are more likely to show the higher impact of this crisis, when the world economy had not fully recovered, than from 2002.

The accuracy figures are different for various cluster pairs. Relationships between settlements with "strong" socio-economic indicators in some cluster pairs (for example *small - rich – megacities*) are better described using Eq.5.3 than for "weak" cluster pairs (for example *large-poor – small-poor*). Therefore, the accuracy is higher for settlements with relatively high socio-economic indicators. In addition, it should be noted that cluster pairs are not equal in terms of number of passengers. For the APD forecasting model, it is necessary to have a high accuracy for connections with a high APD in order to accurately predict new APD connections. Tab.5.8 presents the final accuracies for the accumulative passenger number according to cluster pair in Fig.5.21 from the forecasted 2012 APD networks from 2002, 2007 and 2011. The figures in brackets indicate accumulated cluster pair numbers corresponding to a given accumulated percentage of passengers. Therefore, the proposed topology forecasting validation using historical data shows acceptable results and seems to be adequate for further modeling and model application. However, the accuracy could probably be enhanced by defining appropriate coefficients in Eq.5.1 for every time interval per cluster pair. Next, it is necessary to analyze the WRA index boundary criteria in the sorted lists of non-existing connections for each cluster pair.

	50% passengers	**75% passengers**	**90% passengers**	**100% passengers**
2002	0.899 (5)	0.859 (12)	0.813 (19)	0.682 (45)
2007	0.929 (5)	0.899 (12)	0.861 (19)	0.748 (45)
2011	0.949 (5)	0.927 (12)	0.897 (19)	0.812 (45)

Tab.5.8 Average accuracies for 2002, 2007 and 2011 for a given percentage of passengers. Figures in brackets indicate the cluster pair number, generating a given percentage of passengers

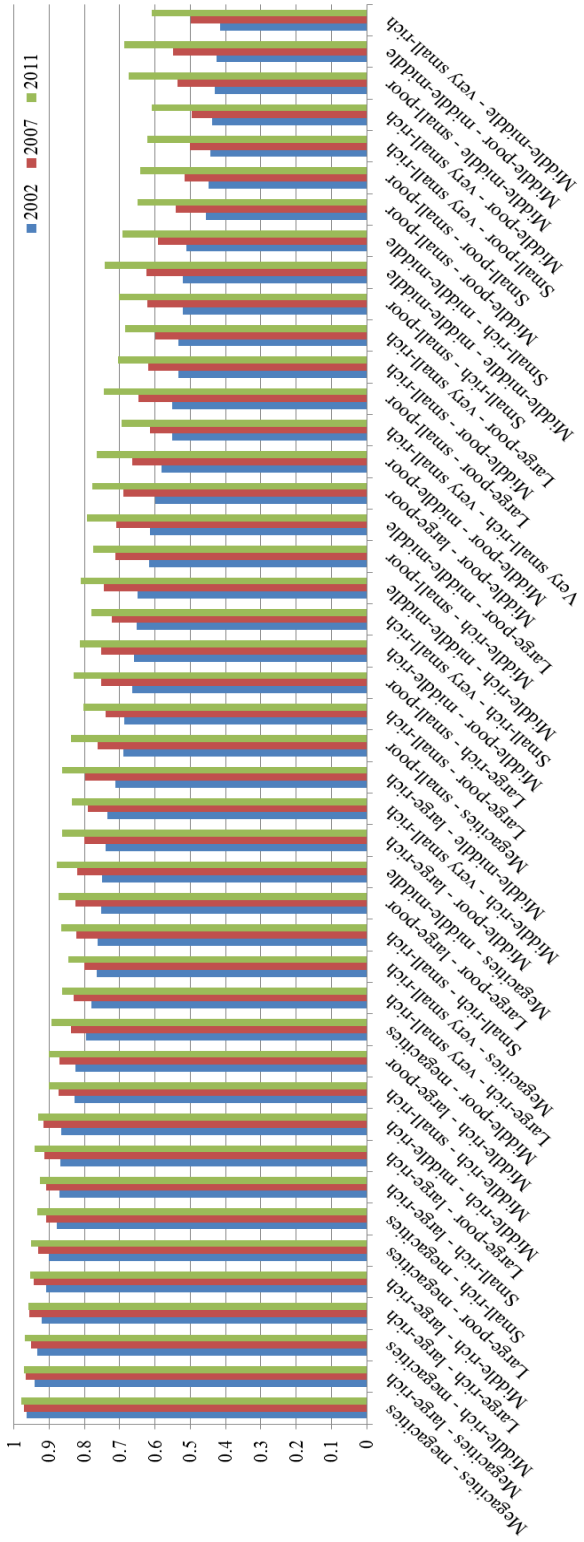
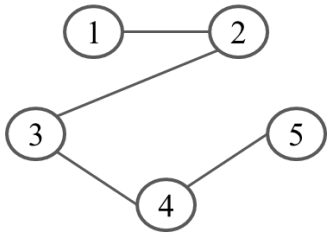


Fig.5.21 Final accuracies for connection in every cluster pair in 2012 from 2002, 2007 and 2011

5.2.3. Boundaries for addition and elimination connections processes

Based on the aforementioned analysis, there are two ways to define boundaries in the sorted list of connections: either using the fixed number of new added/eliminated connections in each cluster pair or the boundary scores for each cluster pair. In other words, for the first method, a fixed number of connections are added to or eliminated from the network from the sorted non-exiting connections list, which is arranged in descending order according to the score. In the second method, for the addition connection process into the APD network, all connections where the score exceeds the boundary score (boundary for adding connections) in the sorted list with all possible connections is added to the APD network. For the elimination process, all connections where the score does not exceed the boundary score (boundary for eliminating connections) in the sorted list are eliminated from the APD network. For example, the APD network topologies of a cluster pair in year *y* (Fig.5.22A) and the next year *y+1* (Fig.5.22B) are known. Settlement socio-economic indicators from year *y+1* are assigned to the same settlements as in year *y*. Scores for all non-existing connections are calculated using the WRA index (Fig.5.22D). Connections are sorted in descending order according to their score (Fig.5.22E). The accuracy of the method can be defined using new real added connections to network in year *y+1*. The number of forecasted links from the top of the list is equal to the number of new real added connections. The accuracy is defined as the ratio of relevant connections in the list of non-existing connections to number of new real added connections. There are two criterion types for adding connections to the APD network. The first criterion is a fixed amount of connections. This amount added every year is equal to the number of new real added connections from year *y+1*. The second criterion is the boundary score. Each connection with a score higher than the boundary score in year *y+1* is added to the network. The next step in forecasting is the elimination of connections from the network. Forecasted links from the top of the list (from Fig.5.22E) are added to the existing links in year *y* (from Fig.5.22D) and sorted by their score in descending order. The number of forecasted links from the top of this new list is equal to the number of real connections in year *y+1*. Connections with a score lower than the boundary score for the elimination process are eliminated from the network (Fig.5.22F). Thus, after the addition and elimination processes, the forecasted APD network for year *y+1* is obtained (Fig.5.22C).

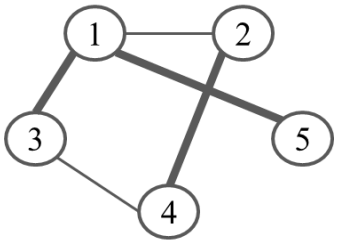
In this study, the link addition and elimination possibilities are considered. It is assumed that APD connections are able to appear and disappear in the APD network. However, there are two adding and elimination connection approaches from the network: a fixed amount of connections or boundary scores. In both approaches, a situation could arise where all settlements within a cluster pair are connected to each other, although this is not very likely. However, this could in fact occur when the first method is used, i.e. using the number of newly added connections in each cluster pair. For example, the cluster pair *middle-rich – small-rich* in 2012 has 207 and 565 settlements respectively. This cluster pair has 27,628 connections in 2012, including 3,538 new connections added from 2011. The number of non-existing connections is 89,327. If it is assumed that the number of added connections will remain fixed, all cities in this cluster pair will be connected to each other within approximately 25 years. For the second method, applying boundary scores, the year when all of the settlements are interconnected in the cluster is hard to predict. It depends on various factors such as network configuration, settlement clustering, socio-economic scenarios, etc. Thereby, in this thesis, it seems reasonable to use the second method of boundary definition for the APD topology forecast – boundary scores. In addition, it is important to note that each cluster pair has different boundaries either for the fixed number of connections method or the boundary score method. This proves the assumption that each cluster pair has its own boundaries which are defined within the validation process. Thus, the boundaries for APD addition and elimination connections for every cluster pair and for every considered time interval (one, five and ten years) are defined. This allows the topology forecasting model to be applied to a socio-economic scenario in order to forecast the existence of the APD between settlements.



A. The APD network topology of a cluster pair in year *y*

Existing connections	Non-existing connections	Score of non-existing connections
1-2	1-3	S13
2-3	1-4	S14
3-4	1-5	S15
4-5	2-4	S24
	2-5	S25
	3-5	S35

D. Existing connections and all non-existing links in the APD network in year *y*. The score for each non-existing link is calculated.



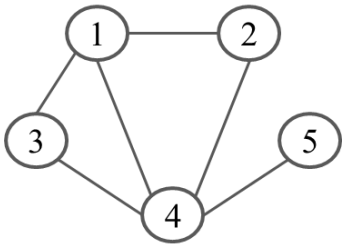
B. The APD network topology of the cluster pair in year *y+1*. Thick lines depict new real added connections to the year *y* network

All possible connections	Score for connections	Real added connections
2-4	S24	1-3
1-4	S14	1-5
1-3	S13	2-4
2-5	S25	
3-5	S35	
1-5	S15	
...	...	

Fixed amount of connections

Boundary score for adding

E. Non-existing connections are sorted in descending order based on their score. Two types of boundaries, based on the number of new real added connections are presented: fixed amount of connections and the boundary score. The forecast predicts two actual connections out of three. Thus, the accuracy in this case is 0.6666.



C. The forecasted APD network of year *y+1*

Existing connections	Score of every connection	Real connections
3-4	S34	1-2
2-4	S24	1-3
1-4	S14	1-5
1-3	S13	2-4
4-5	S45	3-4
1-3	S13	
2-3	S23	

Boundary score for eliminating

F. The new three connections for year *y+1* from *E* are added to the existing connections of year y from *D* and sorted in descending order according to their score. Connections with a score of less than S45 (boundary for the elimination process) are eliminated from the network.

Fig.5.22 The APD network topology forecast example

5.2.4. Conclusion

This section presents the topology forecasting model validation in the APD network using a socio-economic scenario. The study shows that the Weighted Resource Allocation (WRA) index demonstrates the best performance. *AUC* and *precision* metrics are higher for cluster pairs than for the whole APD network. This proves the necessity of separating settlements into groups according to their socio-economic indicators to improve link forecasting performance. Thus, the WRA index is used to calculate scores for all non-existing links in each cluster pair. This disproves the assumption that cluster pairs in the APD network have different similarity indexes demonstrating the best performance. For years with available APD data, modeling is applied and the results are compared with real data. The accuracy of the similarity-based algorithm for the APD network is higher than in related studies. The study shows two methods of adding new connections from the ordered score list of non-existing connections. The first method is to add a fixed number of connections based on the historical analyses. The other method is to use a score number from the ordered list as the boundary. Both methods prove the assumption that each cluster pair has its own boundary. It seems reasonable to use the second approach with the boundary score, since in this case the boundary depends on the network topology and socio-economic indicators changes. Thus, this approach is capable of taking these changes into account and providing a more realistic view of the APD topology changes. Despite the low average accuracy for predicting new connections in the APD network, the topology forecasting model validation demonstrates a high accuracy for elimination connections and the final accuracy for the forecasted APD networks (Tab.5.7). However, this accuracy is strongly related to the fact that more than 50% of connections in 2012 stay in the network from 2002, 2007 and 2011. The APD topology forecast approach is tested for 1, 5 and 10-year time intervals. It is furthermore believed that the accuracy of the similarity-based algorithm can be enhanced. The topology forecasting model can be improved by defining appropriate coefficients in Eq.5.1. It is likely that every cluster pair has its own coefficients. In addition, the main network metrics should be analyzed (e.g. average weighted degree, average path length, modularity, etc.) and compared with the metrics obtained from historical data described by Ghosh and Terekhov (2015). This may help to understand latent processes for the APD connections generation. Nevertheless, the obtained accuracy shows

acceptable results and seems to be adequate for further modeling and model application. For the model application the weighted similarity index is defined – WRA. Using this index, the score for all APD connections could be calculated. The APD topology could be defined using the obtained boundary scores. Thus, the proposed APD forecasting model is ready to forecast the APD appearance or disappearance between settlements for various time frames.

5.3. Passenger forecasting model

Once the APD network topology is defined, the passenger number on every connection is determined by the passenger forecasting model. As shown in Section 5.2, the connections in the APD network are separated into two categories: newly appeared connections and connections remaining in the network from the previous time interval. The passenger forecast process is different for these two categories since the passenger number on the remained connections for the previous time interval is known and for the newly appeared connections these numbers do not exist. As already shown in this thesis, the APD has a strong correlation with socio-economic indicator growth, especially with GDP growth. This correlation shows good predicting accuracy and is used in a wide range of studies (e.g. Grosche et al, 2007; Boeing, 2014). Thus, to define the passenger number on the connections remaining in the network from the previous year, the correlation of the GDP growth and the APD growth is applied. For the newly appeared connections, the Quantitative Analogies (QA) method (Armstrong and Green, 2005) is used. Since the passenger number in previous time slice is unknown, it is not possible to use the correlation of GDP growth and APD growth. However, this is possible to determine for a newly appeared connection with assigned socio-economic indicators from a scenario and analogous connection in the year with actual data. This year could be the last actual year, and within this study, this is the base year. The passenger number on the newly appeared connection is determined by the closest analogous connection in the base year. The passenger number for the new connection is assumed to be the same as the closest analogy in the base year. Therefore, taking these two passenger calculation methods (settlement GDP correlation and QA) into account, the aim of this section is to validate the passenger forecast model and define the accuracy.

For the validation, the passenger number is calculated based on the 2012 forecasted networks. The passenger forecasts for 2012 from 2011, 2007 and 2002 are assessed versus the actual data of 2012. In other words, the forecasted passenger number of 2012 on connections is compared with the real passenger number on real connections in 2012. Since the exact passenger prediction on connections requires particularly complicated effort, the accuracy is assessed at given intervals. In other words, if the real number of passengers falls to a given range of $\pm p$ of the forecasted passenger number, the forecast for this connection is assumed to be correct. The passenger number data for every connection in years 2002, 2007, 2011 and 2012 as well as the APD topology are obtained from the ADI database. The basic statistic is provided in Tab. 5.9. The detailed statistic on the cluster pair level is provided in Appendix A.

	2002	**2007**	**2011**
Total passenger number	1,407,265,975	1,952,984,315	2,427,549,288
Total APD connection number with common settlements in 2012	426,151	500,020	521,171
Total APD connection number	518,750	526,310	527,013
Number of remaining APD connections in 2012	324,487	386,970	423,563
Number of new added APD connections in 2012	194,264	139,340	103,450
Number of eliminated APD connections in 2012	101,664	113,050	97,608
Total passenger number on added APD connections	30,255,220	11,017,267	813,941
Total passenger number on eliminated APD connections	5,084,006	12,121,643	3,223,820

Tab.5.9. Basic statistics for years 2002, 2007 and 2011

The passenger number constantly increases and from 2002 to 2011 the growth is 73%. Tab.5.9 demonstrates the total connection numbers with common settlements within the 2012 APD network. As seen, the connection number also increases. In addition, the longer the time interval, the more connections are added to the network. For example, for 2012 from 2002, 37% connections from the total connection number in 2012 with an average of 156 passengers per new connection are added, when, for 2012, from the 2011 to 2012 comparison, only 20% of connections with an average of eight passengers per new connection are added. The eliminated connection numbers do not demonstrate a decreasing trend. However, the obtained

data do not allow conclusions to be drawn on the issue. In addition, the passenger number on added connections follows the same trend as the added connection number. It must also be related to the time interval difference between years. The passenger number on eliminated connections does not follow the trend coinciding with the eliminated connection number situation. Thus, the basic statistical analysis of years 2002, 2007 and 2011 shows the dynamic for connection and passenger numbers and the importance for correct passenger forecast, both on newly appeared and remaining connections. The connection and passenger number for eliminated, added and remaining connection in 2012 from 2002, 2007 and 2011 for every cluster pair are shown in Appendix A in A.8. In order to assess the forecasting passenger model, this section is organized as follows. In Sub-section 5.3.1, the accuracy for the quantitative analogies for the newly added connections into the network is determined. In Sub-section 5.3.2 the accuracy for the correlation between GDP and the passenger number on connections is shown. A conclusion is provided in Sub-section 5.3.3.

5.3.1. Quantitative analogies approach validation for the newly added connections

As shown in literature (Armstrong and Green, 2005), within the range of quantitative forecasting methods, the quantitative analogies approach is most suitable if there is a poor knowledge of relationships between variables and the data are presented in a cross-sectional format. This is particularly true regarding APD forecasting on newly appeared connections. The relationships between future socio-economic settlement indicators and the APD on these connections are latent, requiring complicated analysis. The data, obtained from the ADI database, contain APD information on an annual basis to form the cross-sectional data. In addition, in order to reduce potential biases in QA application, a representative sample of analogies should be used. The annual ADI data contain more than 400,000 connections in every year. These sets present various types of connections. Using the proposed clustering approach in Section 5.1, the data in every year contain analogies between nine groups of settlements separated by their socio-economic indicators. As shown in Sub-section 4.2.4, the QA can be presented as a simple expert model, where the analogies are determined by calculating the shortest distance between settlement pairs in a forecast year and settlements pairs in the base year. In other words, the passenger number on a newly appeared connection is defined by the analogous connection in the base year with the closest socio-economic

indicators to the new connection. Similarity values for this approach are as follows: GDPs and populations of origin and destination settlement, average ticket price, and distance. This similarity is defined in Eq. 5.13.

$$pax_a = \min\{d_{a,n}; d_{a,n+1} \dots d_{a,m}\}, \quad n = 1 \qquad \text{(Eq.5.13)}$$

where: $d_{an} = \sqrt{(g_{On} - g_{Oa})^2 + (g_{Dn} - g_{Da})^2 + (p_{On} - p_{Oa})^2 + (p_{Dn} - p_{Da})^2 + (t_n - t_a)^2 + (l_n - l_a)^2}$

(Eq.5.14)

Where *pax* is the passenger number on a connection; *a* – a newly added connection in 2012; *n* – connections in the previous year 2011; *m* – the total connection number where analogies are searching; *d* – the distance between two connections; *O* – the first settlement in a city pair; *D* – the second settlement in a settlement pair; *g* – the settlement GDP; *p* – the settlement population; *t* – the average airfare; *l* – the distance between settlements.

Sets of the newly added connections in 2012 from 2002, 2007 and 2011, taking into account the actual passenger number on these connections in 2012, are analyzed. Using the cumulative curves for every connection set for 2012, it could be seen that newly appeared connections emerge mostly between 95-100% of passengers (Fig.5.23, Fig.5.24, Fig.5.25 and Tab.5.10). Thus, the newly appeared connections do not have a large number of passengers. In other words, if a connection between two settlements has appeared, the number of passengers will be comparable to the passenger numbers on connections in the base year which are found in 95-100% of connections on the cumulative curve.

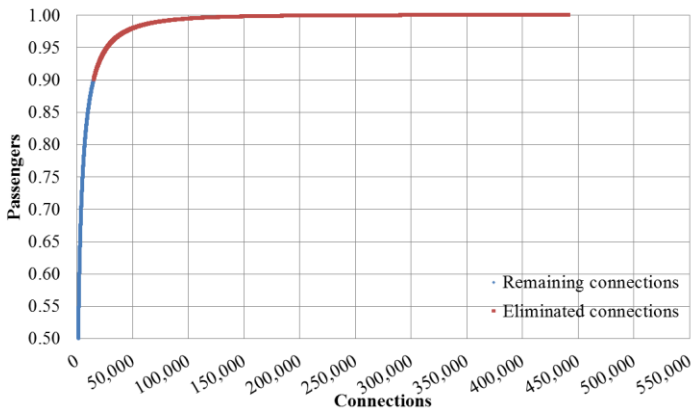


Fig.5.23. Newly appeared connections from 2002 on the cumulative curve for the 2012 base year

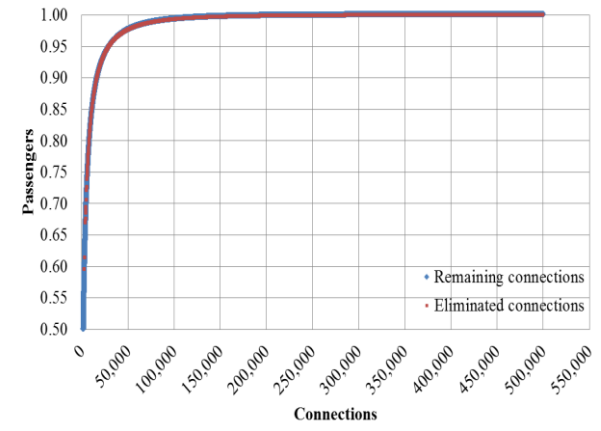


Fig.5.24. Newly appeared connections from 2007 on the cumulative curve for the 2012 base year

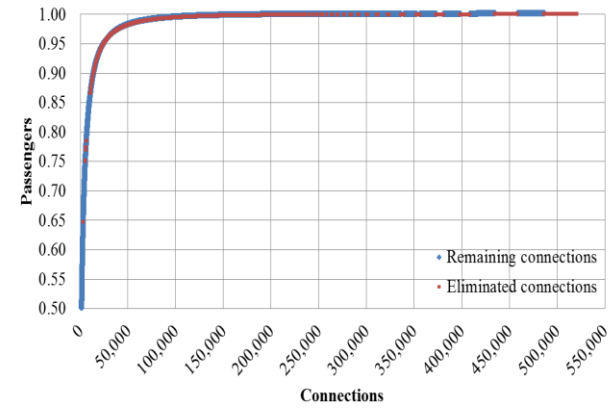


Fig.5.25. Newly appeared connections from 2011 on the cumulative curve for the 2012 base year

%	2002	2007	2011
0 – 5	0	0	0
5 – 10	0	0	0
10 – 15	0	0	0
15 – 20	0	0	0
20 – 25	0	0	0
25 – 30	0	0	0
30 – 35	0	0	0
35 – 40	0	0	0
40 – 45	0	1	0
45 – 50	0	0	1
50 – 55	0	0	0
55 – 60	0	1	0
60 – 65	0	1	1
65 – 70	0	9	0
70 – 75	0	10	1
75 – 80	0	32	4
80 – 85	0	38	0
85 – 90	0	96	16
90 – 95	335	292	71
95 – 100	113,340	112,570	97,506

Tab.5.10. The eliminated number of connections from the 2002, 2007 and 2011 APD network in 2012 at a given interval on the cumulative curves (Fig.5.23, Fig.24 and Fig.25)

Cluster pairs	2002, con	2002, pax	2007, con	2007, pax	2011, con	2011, pax
Very small-rich – very small-rich	13,724	1,719,021	14,925	2,011,567	14,197	2,232,425
Small-poor – very small-rich	4,825	385,390	6,454	598,060	6,797	723,627
Large-rich – very small-rich	17,855	3,087,777	19,003	3,957,721	18,689	4,699,087
Middle-middle – very small-rich	4,147	235,878	5,650	387,487	6,115	380,788
Megacities – very small-rich	12,464	1,878,222	13,880	2,511,050	13,980	2,686,011
Small-rich – very small-rich	29,732	3,050,222	31,929	3,588,989	30,474	3,827,359
Large-poor – very small-rich	8,026	436,707	10,418	654,521	11,602	755,968
Middle-rich – very small-rich	23,149	3,391,058	24,073	3,997,566	23,282	4,654,152
Middle-poor – very small-rich	2,503	103,234	3,498	163,306	4,188	219,552
Small-poor – small-poor	2,175	490,151	2,944	689,219	3,234	800,740
Large-rich – small-poor	8,047	1,189,277	9,976	1,881,490	10,796	2,323,915
Middle-middle – small-poor	3,648	586,016	4,898	946,358	6,155	1,070,624
Megacities – small-poor	6,999	1,344,415	8,865	1,935,623	9,835	2,273,385
Small-rich – small-poor	8,624	683,238	11,504	1,139,948	11,823	1,448,885
Large-poor – small-poor	6,014	947,720	8,787	1,675,593	10,627	1,995,410
Middle-rich – small-poor	9,332	1,151,749	11,390	1,697,863	11,986	2,150,513
Middle-poor – small-poor	2,099	248,693	3,181	477,200	4,447	539,001
Large-rich – large-rich	2,432	1,220,093	2,715	1,609,038	2,656	2,321,595
Middle-middle – large-rich	8,569	1,066,081	10,720	1,650,673	11,848	2,012,571
Megacities – large-rich	3,593	2,180,907	3,919	2,943,577	4,046	4,520,006
Small-rich – large-rich	17,462	5,286,444	18,593	6,666,513	18,162	8,345,580
Large-poor – large-rich	10,290	2,591,192	12,343	3,985,961	12,959	5,383,334
Middle-rich – large-rich	8,943	3,400,922	9,735	4,552,165	9,749	6,017,922
Middle-poor – large-rich	5,550	813,642	6,945	1,262,377	7,843	1,702,312
Middle-middle – middle-middle	3,008	629,979	3,819	925,934	5,101	1,202,613
Megacities – middle-middle	6,428	1,286,046	8,380	2,025,078	9,426	2,248,839
Small-rich – middle-middle	8,448	522,183	11,580	856,757	12,172	883,563
Large-poor – middle-middle	7,132	1,296,528	10,159	1,897,420	12,699	2,610,327
Middle-rich – middle-middle	10,163	880,337	12,580	1,309,818	13,577	1,575,802
Middle-poor – middle-middle	1,626	84,571	2,645	198,409	3,780	246,799
Megacities – megacities	1,172	782,834	1,170	1,049,991	1,263	1,575,555
Small-rich – megacities	13,792	3,213,152	15,114	4,350,359	15,248	5,126,749
Large-poor – megacities	7,598	2,548,175	9,005	4,017,901	10,015	5,704,249
Middle-rich – megacities	7,365	3,296,018	7,916	4,450,524	8,237	5,718,530
Middle-poor – megacities	5,043	1,106,221	6,417	1,718,839	7,401	2,364,692
Small-rich – small-rich	19,546	2,750,017	20,608	3,635,428	19,478	3,882,708
Large-poor – small-rich	16,185	1,034,808	20,426	1,529,092	21,921	1,641,001
Middle-rich – small-rich	26,098	6,331,979	27,311	7,955,655	26,278	9,444,498
Middle-poor – small-rich	5,508	207,754	7,510	331,466	8,443	431,241
Large-poor – large-poor	5,335	1,071,076	7,584	1,895,881	8,796	2,557,279
Middle-rich – large-poor	14,624	2,213,835	17,486	3,292,999	18,516	4,089,005
Middle-poor – large-poor	4,392	481,404	6,610	901,066	8,865	1,459,336
Middle-rich – middle-rich	7,355	2,426,223	7,866	3,176,847	7,868	4,045,895
Middle-poor – middle-rich	6,442	521,443	8,031	791,719	9,033	987,801
Middle-poor – middle-poor	1,302	191,438	2,002	355,040	3,035	498,744

Tab.5.11. The number of analogy connections in every cluster pair for the base year 2012 from 2002, 2007 and 2011

In addition, the sets of connections between 95-100% on the cumulative curve are divided into subsets based on cluster pairs. Since the passenger number depends on the socio-economic indicators, as mentioned before, the analogy for a newly appeared connection is sought in the 95-100% interval on the base year cumulative curve and then in the similar cluster pair subset. The number of analogy connections in every cluster pair and the passenger numbers for the base year 2012 from 2002, 2007 and 2011 are presented in Tab.5.11.

Thus, the QA approach for defining passenger number on a newly appeared connection follows the next procedure. For the new connection, the affiliation to a cluster pair is known. Thus, in the identified connection set between 95-100% on the cumulative curve in the base year, the connection subset in a given cluster pair is retrieved. Calculations are made to determine distances between the new connection and all connections in the cluster pair by applying Eq. 5.14. Then, Eq.5.13 is used to define the connection in the base year with the shortest distance to the new connection. Thus, the passenger number on the connection in the base year is assigned to the newly appeared connection.

The accuracy of this approach is calculated to forecast passenger numbers from the three years (2002, 2007 and 2011) to 2012. The actual passenger data for these years are obtained from the ADI database. The number of newly appeared connections from every year to 2012 is known. Socio-economic indicators for 2002, 2007 and 2011 are obtained from the same databases as for the base year 2012. However, due to data limitations, the settlement population for these years is calculated based on the settlement population in the base year and the urbanization rates in 2002, 2007 and 2011. The passenger number for these years is calculated and compared to the actual 2012 data.

In order to obtain the QA approach accuracies, the passenger number on newly appeared connections are assessed in intervals. The interval borders are defined as an integer number. In other words, the actual passenger number is the middle value in integer interval ±p. For example, if the interval is ±5 and the actual passenger number on a connection is 9, then the forecasted passenger number will be assessed in the interval [4, 14]. If the forecasted passenger number on this connection is inside the interval [4, 14], the forecast is assumed to be correct. If not, the forecast is wrong. The intervals change in order to assess the accuracy of

the forecast. In addition, the forecast is more accurate if the interval is small and the number of connections in this small interval is high.

The intervals are chosen based on the historical data statistics provided in Tab.5.12 where the number of newly added APD connections in 2012 from a given year, the total passenger number on added APD connections and the average passenger numbers on connection for 2012 from 2002, 2007 and 2011 are presented.

	2002	**2007**	**2011**
Number of newly added APD connections in 2012 from a given year	194,264	139,340	103,450
Total passenger number on added APD connections	30,255,220	11,017,267	813,941
Average passenger number per connection	156	79	8

Tab.5.12. The connection and passenger numbers and the average passenger number on newly appeared connections in 2012 from 2002, 2007 and 2011

As seen, the average passenger number decreases as the time period between the years decreases. This shows that the newly appeared connections for the one-year period (from 2011 to 2012) appear with a low passenger number and this passenger number does not differ significantly compared to all other connections. For the long term of ten years (from 2002 to 2012), the analysis shows that newly appeared connections have different development in terms of passengers over that decade. For example, in 2002 a settlement had a few APD connections and from 2002 the settlement experienced rapid socio and economic growth. The connection number changed within these ten years as well as passenger numbers on them. Thus, in 2012 the settlement had a number of APD connections which were generated in a 10-year interval from 2002, but the year in which connections appeared is unknown. Basically, the period from 2002 to 2012 could be presented as a black box. The inputs to 2002 and outputs in 2012 are known, but the information about these ten years in between is missing.

The QA approach accuracies were assessed within a ±50 passenger interval, based on the average passenger number per connection. In this interval, the QA approach predicted more than 70% connections correctly, as follows: 75% of connections for the long term of ten years (from 2002 to 2012), 74% of connections for the middle term of five years (from 2007 to 2012) and 71% of connections for the short term of one year (from 2011 to 2012). These correctly predicted, newly appeared connections contain less than 10% of the actual passenger

number: 3% for the long term of ten years, 4% for the middle term of five years and 9% for the short term of one year. In order to have a QA accuracy overview, the approach was assessed in discrete intervals from 0 to 500. In Fig.5.26, the average accuracies for all newly appeared connections in 2012 from 2002, 2007 and 2011 are presented. On the vertical axes, accuracies from 0 to 100% are indicated, and the absolute integer intervals are represented at the outer circumference. The diagram shows the total amount of correctly predicted connection numbers and, covered by these connections, the actual passenger number at given intervals.

As demonstrated, the QA approach works well predicting almost 90% of new connections correctly within a ±300 passenger interval. However, these connections only cover about 10% of passengers on all actual newly appeared connections. For 2002, the forecast accuracy is less than for 2007 and 2011. The number covered by these connections passengers for 2002 is also smaller than in 2007 and 2011. This is directly related to the time period between these years. The large periods cause lower accuracy. On average, 2011 demonstrates a higher accuracy level. The correctly predicted, newly appeared connections cover more than 15% of passengers. Thus, the QA approach demonstrates a high accuracy for newly added connections with a relatively low passenger number, since connections with a large passenger number do not fall into given intervals. This can be explained by the poor correlation of socio-economic indicators for this connection type, as shown in Eq.5.13 and Eq.5.14. In other words, passenger generations on connections with large passenger numbers are not likely to be supported mainly by socio-economic indicators because other potential reasons which are challenging to identify and quantify, such as cultural features of connected settlements or political decisions could come into effect.

In addition, the average estimated accuracies are assessed in every cluster pair. The average accuracies for all intervals from 0 to 500 are depicted for 2012 from 2002, 2007 and 2011 (Fig.5.27, Fig.5.28 and Fig.29 respectively). The detailed accuracies for every cluster pair and the covered passenger number on correctly predicted connections are shown in Appendix A in A.2, A.3, A.4, A.5, A.6 and A.7 for one, five and ten time intervals.

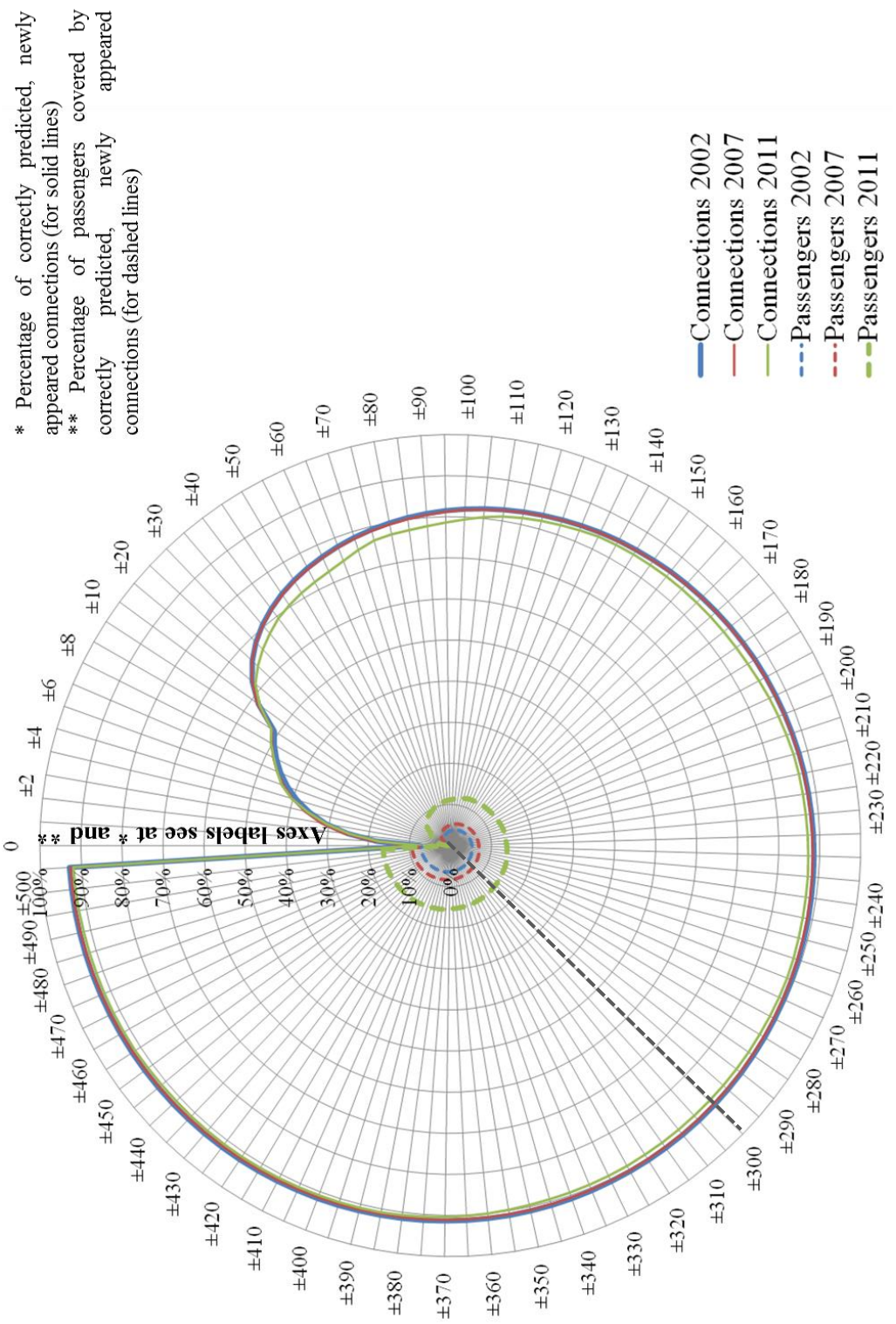


Fig.5.26. Average accuracy of the total number of newly appeared connections at the given intervals

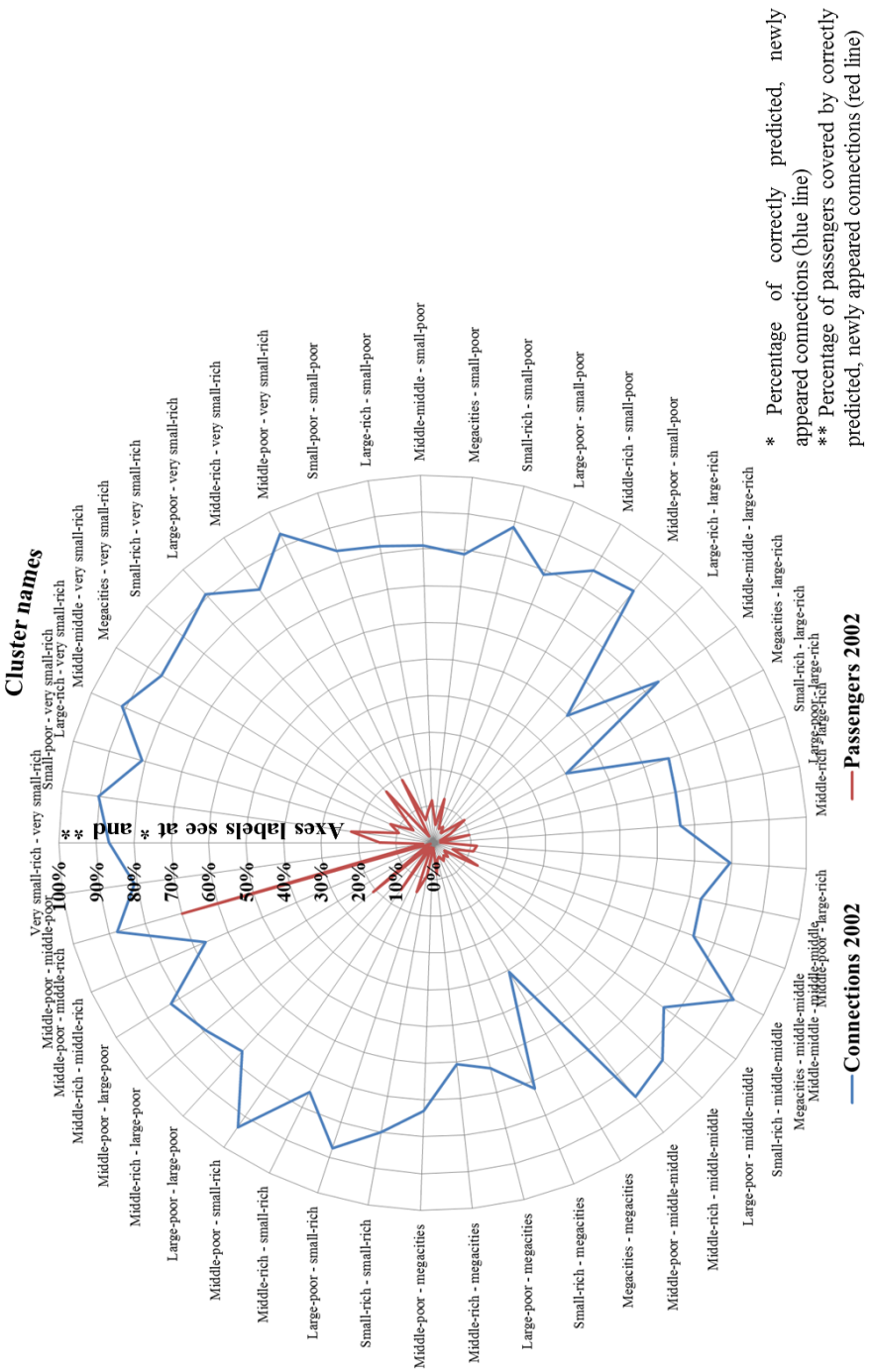


Fig.5.27. The QA average accuracy for forecasted newly appeared connections for 2012 from 2002 to within ±500 passengers

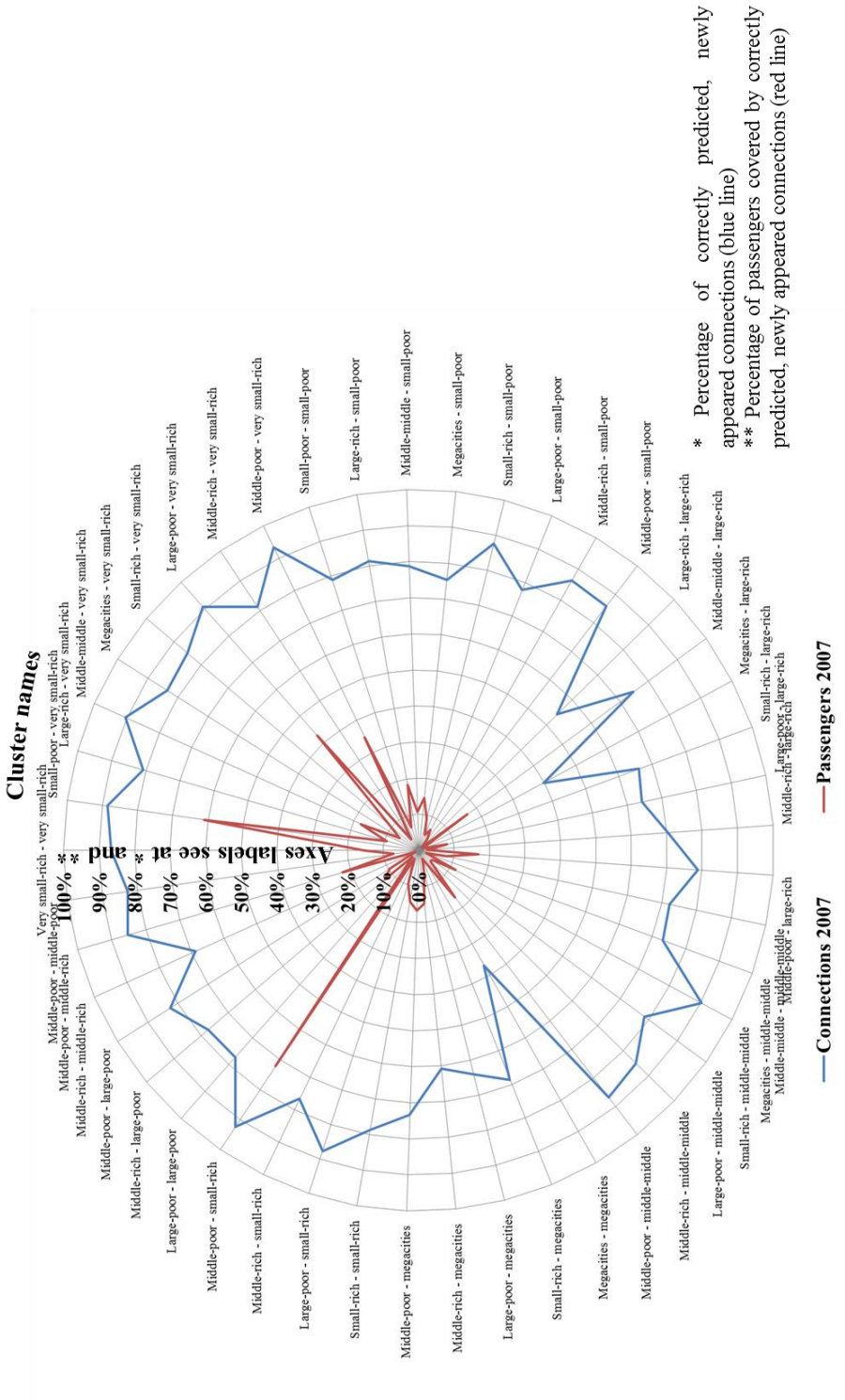


Fig.5.28. The QA average accuracy for forecasted newly appeared connections for 2012 from 2007 to within ±500 passengers

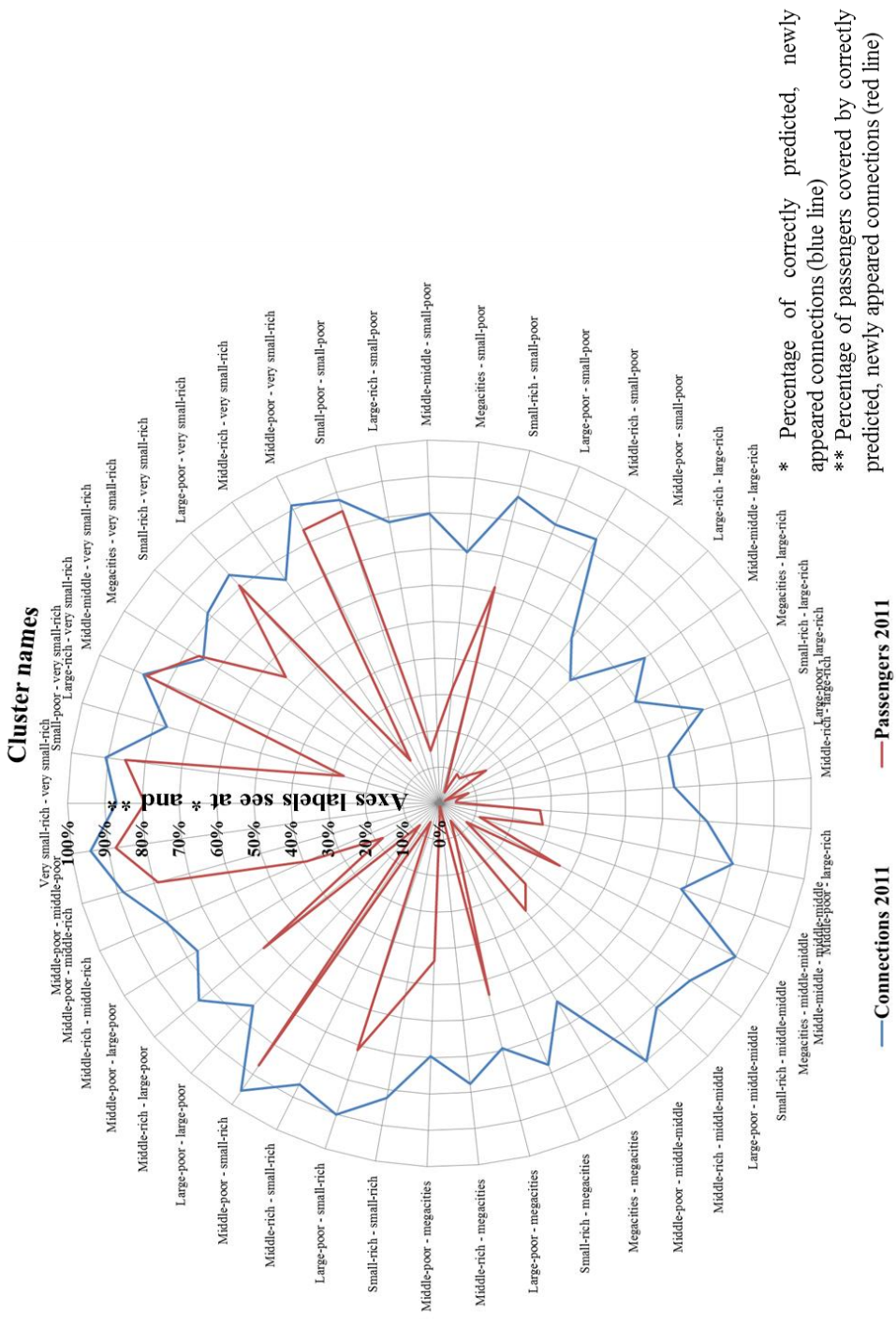


Fig.5.29. The QA average accuracy for forecasted newly appeared connections for 2012 from 2011 to within ±500 passengers

On Fig.5.27, Fig.5.28 and Fig.5.29, the correctly predicted connection with average accuracies of discrete intervals from 0 to ±500 for every cluster pair are presented (detailed accuracies for every cluster pair for the intervals are shown in Appendix A, Tab.A.2 - A.7). The passenger numbers covered by correctly predicted connections for every cluster pair are also indicated in these figures. As seen, more than 60% of connections in almost all cluster pairs for all of the year groups are predicted correctly. However, the total passenger number covered by these connections in every cluster pair increases due to the prediction time term. For 2012 from 2002, the accuracy for the connection number with correctly predicted passenger numbers shows high average accuracy, however cluster pairs which include *megacities* and *large-rich* settlements demonstrate low results. This is related to the QA approach: connections with a large APD in these cluster pairs probably do not have clear correlations with socio-economic indicators such as settlement GDP and population. It is likely that there are some additional factors and indicators which play a significant role in the connection establishment. The passenger number for 2012 from 2002 shows a poor result. The correctly predicted connections in every cluster cover, as an average of all clusters, about 7% of the actual passenger number on actual, newly added APD connections in 2012 from 2002. Only one cluster pair *middle-poor – middle-poor* shows a very high accuracy of 70% for covered passengers on the correctly predicted connections. However, the poor results for the remaining cluster pairs are likely to be related to the latent processes for connections with a large passenger number. In addition, the prediction from 2002 to 2012 has a ten-year gap. The passenger generation processes in such a time frame are not clear. They probably dynamically change within these years and could hardly be cached with the QA approach. For the 2012 from 2007 prediction, the situation with the prediction is the same as for 2012 from 2002. The average passenger number of correctly predicted connections for 2012 from 2007 is 12% and it is higher compared to 2012 from 2002. Settlements from *very-small – rich* cluster settlements demonstrate the highest accuracy in terms of passengers. For the one-year time frame from 2011 to 2012, the prediction demonstrates the highest accuracy for the QA approach compared to five and ten year terms. The average passenger accuracy in terms of passengers is 87%. Here, settlements which are mainly from the clusters *small-rich*, *small-poor*, *middle-poor* and *large-poor* settlements show the best accuracy. A tendency can therefore be seen: the QA approach accuracy is higher when the prediction time frame is

smaller; the QA approach has a higher accuracy in poor settlements. The accuracy increases as follows: over the long term, the accuracy is low for all cluster pairs, over the middle term, the accuracy is high for clusters with *very small-rich* settlements and over the short term, the accuracy is high for clusters with poor small, middle and large settlements.

The forecasting time period for the QA approach plays an important role. The shorter the forecasting period between the base year and forecasted year, the higher the accuracy in terms of passengers on the correctly predicted connections. However, the central indicator for the QA approach accuracy is the connection number with the correctly predicted passenger number. In general, the poor accuracy of connections with high passenger numbers using QA could be explained by the poor correlation between passenger number on a connection and socio-economic indicators of settlements on this connection. In other words, passenger growth is less related to socio-economic growth and it is caused by other combined latent reasons (such as political or humanitarian reasons). The emergence of these reasons is hard to predict on a worldwide settlement scale. However, assessing the approach from the connection number with the correctly predicted passenger number point, the QA demonstrates high accuracy for all forecasting time periods. This is essential for the proposed forecasting model, since the APD forecast process is based not solely on passenger number, but mainly on the combination of the APD network structure, settlements socio-economic indicators and only then on the passenger number on the connections. Thus, the QA validation for forecasting the APD between newly appeared settlement pairs worldwide shows sufficient accuracy. The validation results therefore allow the presented method to forecast the passenger numbers on newly appeared settlements pair in future APD networks.

5.3.2. Validation of the APD correlation with GDP for remaining connections

As already shown in this thesis, APD demand has a strong correlation with socio-economic indicators. However, the aforementioned studies with high accuracy consider the limited number of airports with various socio-economic indicators such as population, income, proximity of a hub airport, tourism destination and so on in Jorge-Calderon (1997) or income, sales competition, phone calls, international passengers on domestic flights and so on in Brown and Watkins (1968). These data for modeling passenger numbers on connections between settlements are hardly available for the considered settlement level aggregation

worldwide and cannot be used in this study. However, based on the aggregate trend analysis, Ishutkina and Hansman (2008) demonstrate that the passenger number and GDP at country level have a correlation, but the passenger growth rates and processes behind the interaction differ for different economies. Although the underlying processes are unknown, the GDP development for every considered settlement in this is known from a scenario. In this study the passenger number on remaining connections is therefore calculated according to the average GDP growth between settlements in the forecasted year and the previously known year. The mathematical interpretation of this correlation could be presented as follows:

$$p_{ij,y_a} = p_{ij,y_b} * z_{ij,y_a,y_b} \text{ , where} \quad \text{(Eq.5.15)}$$

$$z_{ij,y_a,y_b} = \left(\frac{g_{i,a}}{g_{i,b}} + \frac{g_{j,a}}{g_{j,b}}\right) * 0.5 \quad \text{(Eq.5.16)}$$

p_{ij,y_a} – the forecasted passenger number in forecasted year y_a between settlements i and j; p_{ij,y_b} – the passenger number in previous known year y_b between settlements i and j; z_{ij,y_a,y_b} – the average GDP growth between years y_a and $y_b$between settlements i and j; g – the city GDP.

In order to define correlation approach accuracy, the approach is validated on the remaining connections in 2012 from 2002, 2007 and 2011. The number of remaining connections, passengers and the average passenger number per connection are presented in Tab. 5.13.

	2002	**2007**	**2011**
The number of remaining connections in 2012	324,487	386,970	423,563
The number of passengers on remaining connections in 2012	1,402,181,969	1,940,862,672	2,424,325,468
Average passenger number per connection	4,321	5,016	5,724

Tab.5.13. The connection and passenger numbers, the passenger mean values and the standard deviations on remaining connections in 2012 from 2002, 2007 and 2011

The number of remaining connections constantly increases. It is possible to observe that the shorter the time interval between years, the more connections remain in the network. In other words, over a long time interval, more connections are eliminated from the network, since there are more latent processes within this period. This correlates with the eliminated connection numbers and cumulative curves in Sub-section 5.3.1. In addition, the average

passenger numbers on connections decrease. This means that on short periods there are many connections with fewer passengers than on longer periods, where there are not as many connections, but more passengers on them.

The accuracy for the correlation approach is assessed in intervals using the same logic as for the accuracy assessment for the newly appeared connections in Sub-section 5.3.1. However, due to the significant difference between the average passenger numbers on the newly appeared connections and remaining connections in Tab.5.12 and 5.13, the intervals here are defined as a percentage of the actual passenger number. For example, if the interval is ±10% and the actual passenger number on a connection is 9,854, then the forecasted passenger number will be assessed in interval [8,869; 10,839]. If the forecasted passenger number on this connection is within this interval, the forecast is assumed to be correct, if not, the forecast is wrong. The intervals change in order to assess the accuracy of the forecast. In addition, the forecast accuracy is higher if the interval is short and the number of connections in this short interval is high.

The accuracy was assessed within a ±50% interval. In this interval, the correlation approach predicted more than 35% remaining connections correctly: 35% for the long term of ten years (from 2002 to 2012), 41% for the middle term of five years (from 2007 to 2012) and 57% for the short term of one year (from 2011 to 2012). These correctly predicted, remaining connections contain more than 45% of the actual passenger number: 46% for the long term of ten years, 61% for the middle term of five years and 83% for the short term of one year. In order to have an accuracy overview for the correlation approach, it was assessed in discrete intervals from 0 to ±150%. The total average accuracies for remaining connections in 2012 from 2002, 2007 and 2011 are shown in Fig.5.30. The diagram shows the total correctly predicted connection numbers which cover the actual passenger number at given intervals.

As seen, the correlation approach demonstrates high accuracy for all years for connections with correctly predicted passenger numbers in interval ±100%. In contrast to the QA approach, the correlation approach shows higher accuracy for connections with a large passenger number. However, the correlation approach prediction demonstrates the same accuracy trend as for QA: for 2002, the forecast accuracy is less than for 2007 and 2011 and

the number of passengers covered by these connections for 2002 is also lower than in 2007 and 2011. The results follow a similar logic as the long periods cause lower accuracy.

In addition, the average accuracies are assessed in every cluster pair. The average accuracies for all intervals from 0 to ±150% of actual passenger numbers for 2012 from 2002, 2007 and 2011 are depicted in Fig.5.31, Fig.5.32 and Fig.33 respectively. Detailed accuracy information for every cluster pair for the correlation approach is presented in Appendix A in A.9, A.10, A.11, A.12, A.13, A.14.

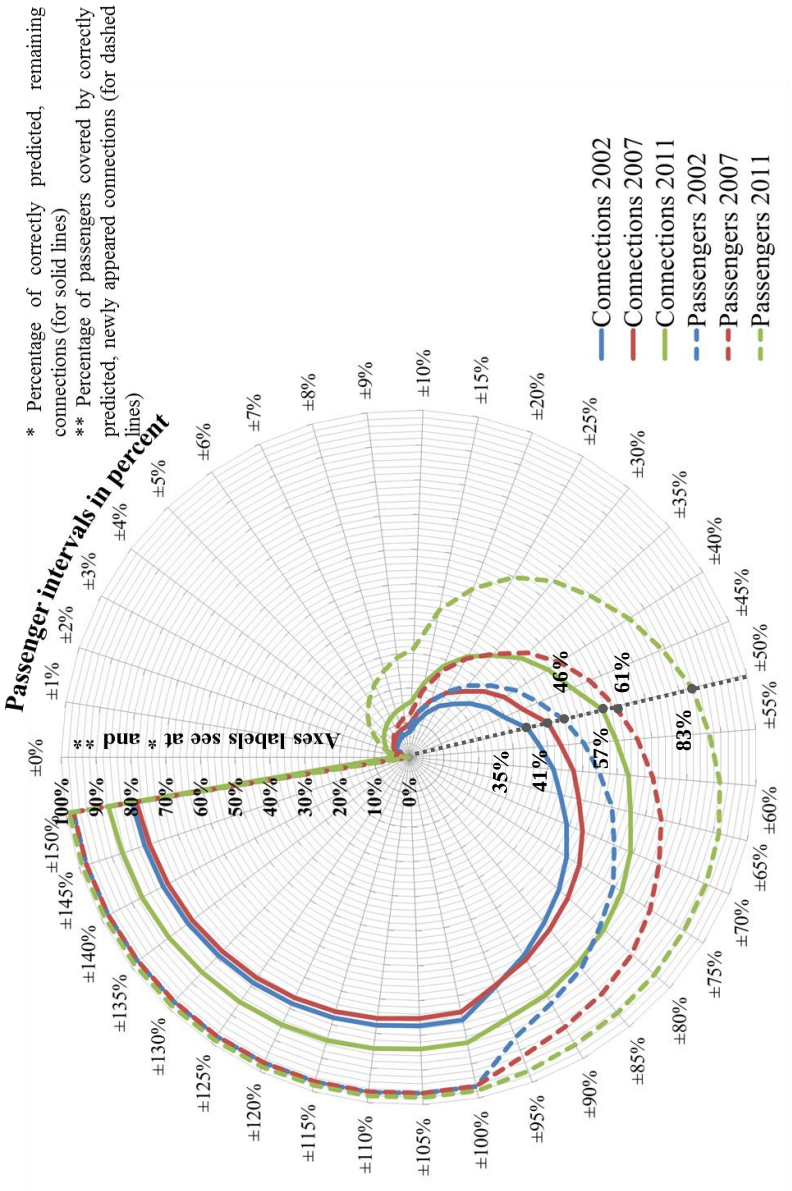


Fig.5.30. Average accuracy at the given intervals for the total number of remaining connections

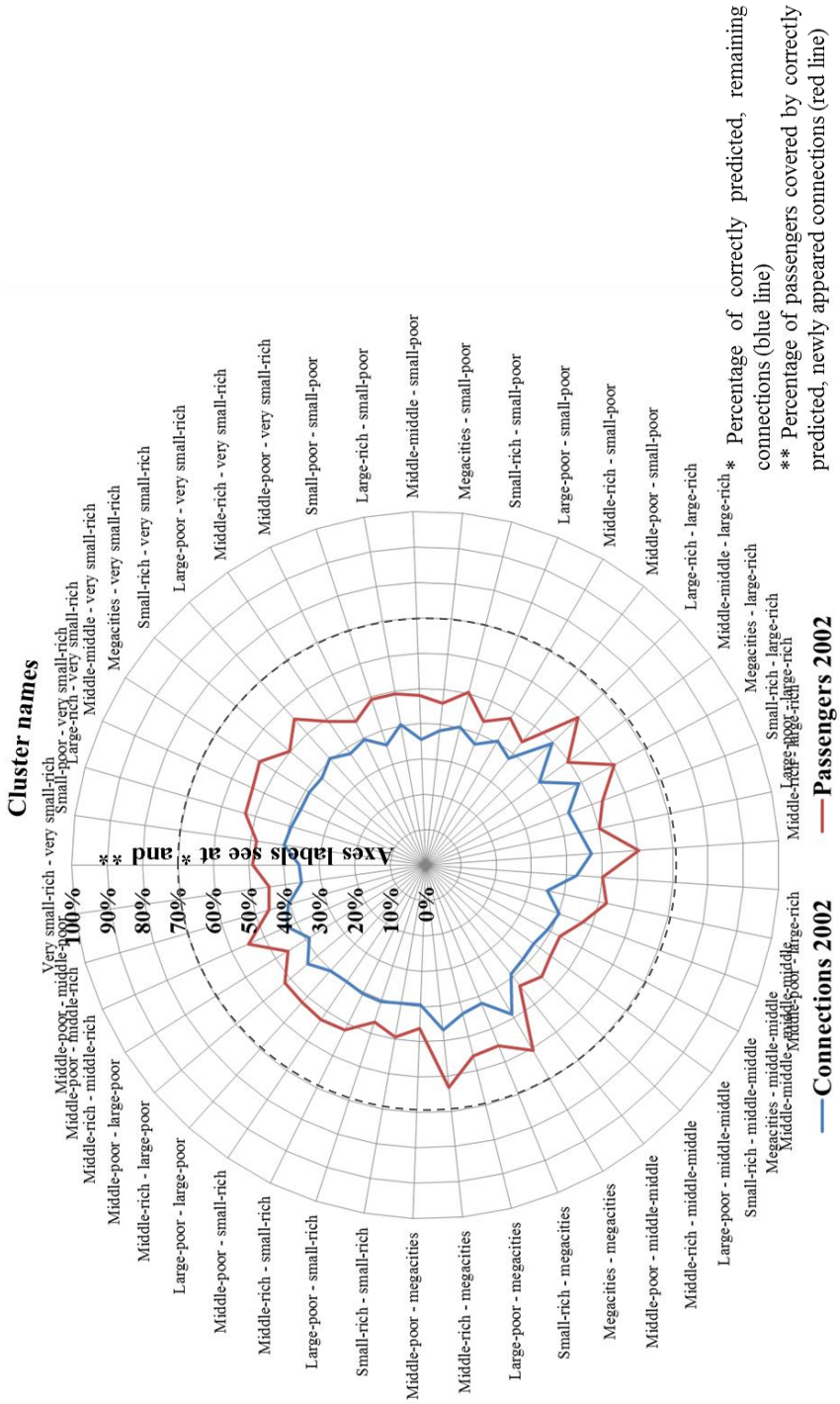


Fig.5.31. Correlation approach average accuracy for remaining connections for 2012 from 2002 to within ±150% passengers

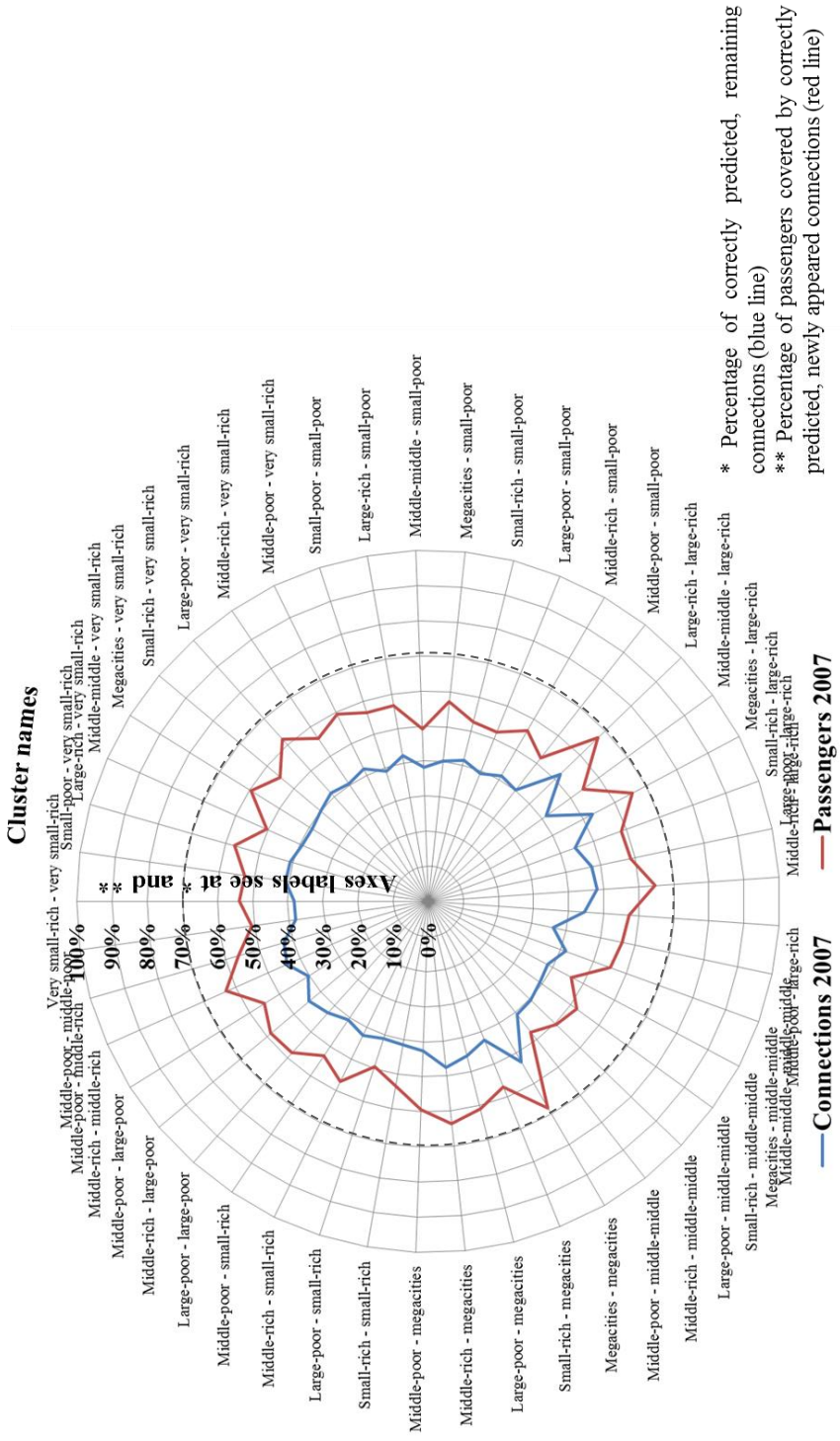


Fig.5.32. Correlation approach average accuracy for remaining connections for 2012 from 2007 to within ±150% passengers

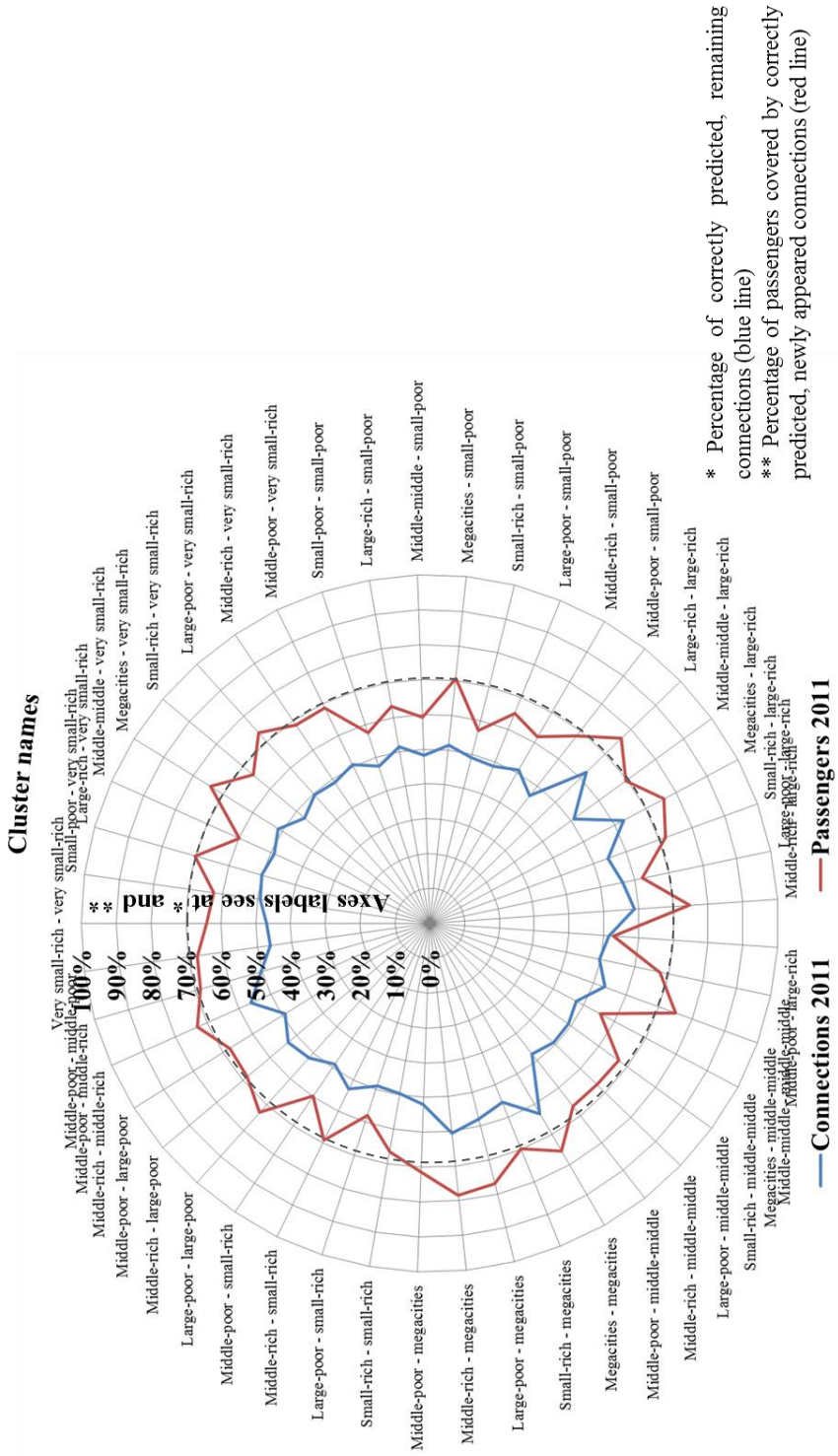


Fig.5.33. Correlation approach average accuracy for remaining connections for 2012 from 2011 to within ±150% passengers

In general, results show that correlation approach accuracy increases the shorter the time interval. Unlike the QA approach, the correlation approach shows higher accuracy in terms of covered total passenger number by connections with correctly predicted passenger numbers. The accuracy for every time interval for these connections is distributed approximately equally within 45 cluster pairs. However, for clusters with a large passenger number, the accuracy is higher. This is related to the validation approach of the intervals definition. Here, the larger the passenger number on a connection, the larger the interval for assessing forecasted passenger numbers on the connection. Thus, the cluster pairs with *large-rich, megacities, middle-rich* settlements demonstrate higher accuracy. For remaining connections in 2012 from 2002 the connection accuracy with correctly predicted passenger numbers is above 0.4. This means that passengers on 60% of connections are incorrectly predicted. However, correctly predicted connections cover more than 50% of the total passenger number in each cluster. For 2012 from 2007 and 2011, the accuracy for connections with correctly predicted passenger numbers is above 0.45 and 0.6, covering an average of 60% and 70% respectively. It is likely that the low results for longer time intervals are related to the hidden processes for passenger generation on some connections. There could be various aspects influencing passenger generation on connections between settlements (e.g. political decisions) which could barely be captured by the city GDP correlation approach. Essentially, the time periods between years might be considered as a black box. The longer the time period between two time points, the higher the distortion in the correlation. Thus, in order to enhance accuracy, it is necessary to introduce a number of stabilizing variables. These variables should compensate distortions and provide a corrected output using the influencing aspects and event impacts on passenger generation. In addition, in the correlation approach, a cluster specific in terms of correlation between passenger number growth and GDP growth is not considered. Thereby, the inclusion of the specific as well as stabilizing variables on this stage of study would add significant complexity into the model, since the correlation equations should be made taking into account these improvements for every cluster pair. It is probable that such procedures would increase accuracy. However, applying the correlation approach for all clusters demonstrates sufficient accuracy for the proposed approach assessment. The validation in this sub-section for the passenger number on connections remaining in the ADP network 2012 from 2002, 2007 and

2011 using the correlation between passenger growth and the GDP growth shows sufficient results and will be applied for further modeling.

5.3.3. Conclusion

This sub-section presents the passenger forecasting model validation for the APD model for 2012 from 2002, 2007 and 2011. There are two connection types: the newly appeared connections and connections remaining in the APD network. Each connection type has its own approach for passenger calculation. The analysis shows that for newly appeared connections, the quantitative analogies approach is reasonable to use. For the remaining connections, based on the literature review, the correlation with the GDP growth is used. For the validation, these connections and passenger numbers on them are extracted from the actual data and the passenger modeling approaches are applied, comparing them with the actual 2012 data. The modeling results are assessed at a number of intervals, since the exact passenger prediction on connections requires significantly complicated efforts. If the predicted passenger number on a connection is within the actual passenger number interval, the prediction is assumed to be correct. Integer intervals are used for the QA approach, and percentage intervals are applied for the correlation approach. The accuracies are assessed for all connections and for every cluster pair. The QA approach demonstrates that, on average, about 80% of connections for all intervals and all years are predicted correctly. However, these connections only cover an average of 10% of the actual total passenger number on newly appeared connections. Nonetheless, it is shown that for the newly appeared connections, a higher number of correct connections is more relevant than the total covered passenger number. For the correlation approach, the validation demonstrates that, on average, about 40% of connections for all intervals and all years are predicted correctly and they cover an average of around 60% of the actual total passenger number remaining connections. For both approaches, the time interval length plays an essential role: the larger the interval the lower the accuracy. The longer the time interval, the larger the random impacts on the passenger number volume on connections. These random impacts are difficult to determine as additional, detailed analysis would have to be performed. Moreover, additional datasets should probably be collected for equation calibration for every connection independently. However, the considered approaches demonstrate sufficient accuracy for applying the passenger forecasting model.

5.4. Overall model accuracy and error propagation analysis

The validation for the forecasting topology and forecasting passenger model were considered in the previous sub-section. Based on the analyses conducted, the overall accuracy for the APD model could be assessed. This accuracy took into account the topology forecasting model accuracy and the passenger forecasting model accuracy. Combining these two factors allow the overall model accuracy to be defined. The modeled results are then compared to the actual data by applying the entire proposed modelling procedure i.e. the topology forecast model for APD topology forecast is used at first and then, based on the model output, the passenger forecasting model is used. The overall accuracy is assessed for 2012 from 2002, 2007 and 2011 as a percentage of the actual passenger number, following the same procedure as for the remaining connections in Sub-section 5.3.2. The overall accuracies from these years to 2012 are defined as the ratio between correctly forecasted passenger numbers on the correctly predicted connections and the actual 2012 APD data. The overall accuracies for 2012 from 2002, 2007 and 2011 are depicted in Fig.5.34. On the vertical axes, accuracies are presented in percent for the correctly forecasted connection number (dashed lines) and for the passenger number covered by these connections (solid lines). Accuracies for correctly predicted connections and the passenger number covered by them correlate to the accuracies on the remaining connections shown in Sub-section 5.3.2. This relates to the large number of remaining connections and passengers on them, as shown in Tab.5.9.

As seen in Fig.5.34, the accuracy is higher when the time interval between two time points is lower. Thus, the one-year time interval from 2011 to 2012 demonstrates the highest accuracy for all intervals considered. The ten-year time interval shows the worst accuracy compared to 2007 and 2011 for 2012. It correlates with the topology forecast accuracy (Tab.5.7) where the accuracy for 2012 from 2007 is higher than from 2002 (0.68 and 0.75 respectively) and the overall accuracy from 2007 is higher than from 2002. Thereby, the "black box" term could be applied, as shown in Sub-section 5.3.1. The longer the time interval between two time points, the higher the likelihood that disturbing events such as *force majeure* or conflicts could occur and thus influence the APD generation. Accordingly, the highest accuracy in this analysis is the time interval with a length of one year. The five-year time

interval demonstrates low accuracy and the lowest accuracy is the ten-year time interval. Nevertheless, although the highest accuracy can be achieved from one-year time intervals, errors accumulate if the time intervals are extended. For long-term forecasts of, say, 30 or 40 years, accuracy levels attained from the one-year interval are very poor. Therefore, the error propagation for three time intervals, namely one, five and ten years, is analyzed for the APD forecast with different time horizons. The analysis is conducted as follows: the accuracies for all three time intervals at given intervals are known and assumed to be constant. In addition, the connection number is deemed to remain constant, as in the base year.

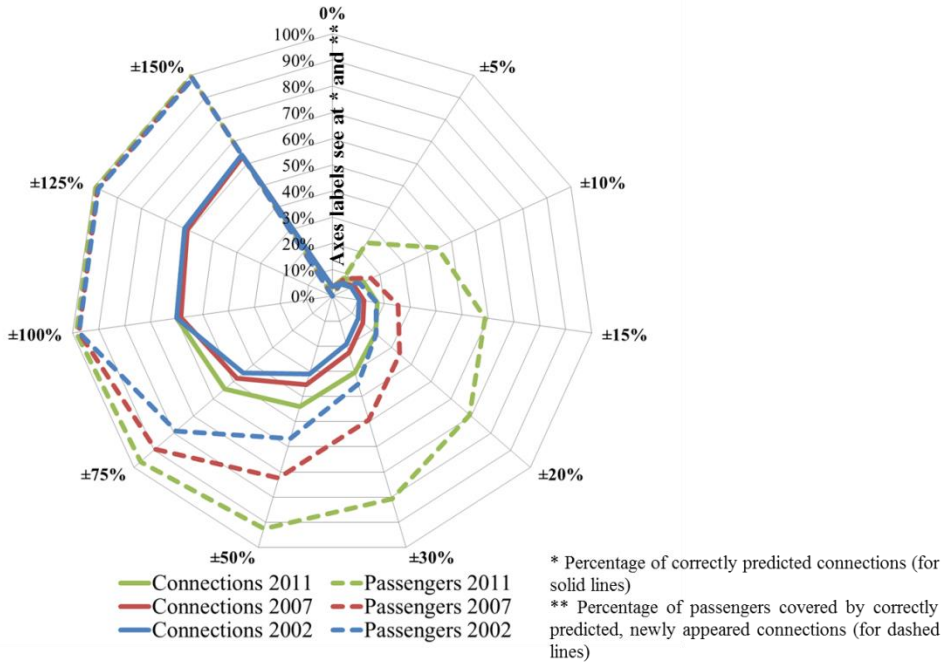


Fig.5.34. The overall accuracies for passengers and connections for 2012 from 2002, 2007 and 2011

For every time interval, the correctly predicted connections decrease according to the accuracy obtained. In other words, from the whole number of connections in the base year, the correctly forecasted connection number in the $y+n$ is known, where y is the base year and n is a time interval. Then, this correctly forecasted connection number in the $y+n$ is assumed to be 100% and the correctly forecasted connections in year $y+2n$ are defined, applying the known accuracy. This process continues up to the required time point. For example, if the base year is 2012 and the five-year time interval is selected, the accuracy would be 0.7. Thus, for 2017, the correctly predicted connection number would be 70% of the total number in 2012. For 2022, the correctly predicted connection number would be defined as 0.7*0.7*100=0.49 or 49% of

the total number in 2012. For 2027, the correctly predicted connection number would be 34.3%. For 2032 this would be 24% and so on. A simple example is shown in Fig.5.35, which depicts error propagation. The time interval is *n* and the accuracy is 0.5. The gray areas demonstrate a correct forecast in percent. Thus, the correct area decreases by 0.5 (the provided accuracy) at each time point. Therefore, with 0.5 accuracy at time interval *3n*, the correctly forecasted area would be 12.5% of the area in the base year. Using the same procedure, the error propagation is assessed for the APD model for one, five and ten-year time intervals.

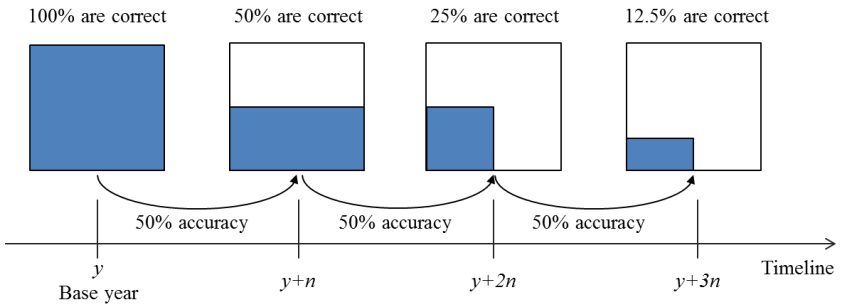


Fig.5.35. Example for the error propagation assessment procedure

In Fig.5.36, the average connection error propagations for intervals from 0 to 150% are presented. As shown in Section 5.3.1, the cumulative curves contain about 4.6% of the total connection numbers, which generate about 90% of the APD. These strong connections remain fixed in the network and are not eliminated. It is assumed that these strong connections remain in the APD at every time point which is why the curves in the figure reach their saturation at 4.6%, rather than progressing to 0. Fig.5.37 presents the average passenger percent covered by correctly predicted connections for a number of discrete intervals from 0 to 150%. As seen in both figures, the expected assessed accuracy decreases for all three time intervals. Nonetheless, the decreasing dynamics are different for various time intervals. The one-year time intervals show good performance for a short range forecast of one year. Within the analyzed time intervals, it is only possible to perform four-year forecasts with the one-year time interval (for every year from 2012 to 2016 for instance). However, the expected connection number accuracy for the years following 2013 years is very low. For example, the expected connection number accuracy in 2017 is close to 4.6% of the total connection number in the base year 2012, covering about 45% of the passenger number. The one-year time interval reaches the saturation of the total connection number in the base year, covering about

41% of passengers in 2019. Thus, this time interval does not demonstrate high accuracy for the middle term and, especially, for the long term forecasts. In contrast to the one-year time interval, the five-year time interval shows better middle term forecasting results. Although the five-year interval is not able to predict the APD for every year, the expected connection number accuracy in 2017 is about 30%, covering about 55% of passengers (for the one-time interval these numbers are 5.8% and 45% respectively). The five-year interval shows higher expected accuracies than the one-year interval. However, using this interval length, it is impossible to make a short-term forecast. In addition, in order to conduct a long-term forecast, it is necessary to use many intervals and the expected accuracy shows lower results than for the ten-year time interval.

The ten-year time interval allows long-term forecasts to be made. As seen in Fig. 5.36 and Fig.5.37, for example for 2020, the best expected accuracy is for the ten-year interval compared to one and five-year intervals. In addition, the ten-year interval application demonstrates the highest expected accuracy for connections and passengers covered. Thus, the analysis shows that time interval length has a significant impact on the expected accuracy. A short-time interval works well for a short-term forecast (2-3 years from the base year), a middle interval shows sufficient accuracy for a mid-term forecast (5-10 years from the base year) and a long time interval performs well for a long term forecast (10 to 30 years from the base year). Thereby, the APD forecast error propagation analysis shows that in order to have a satisfactory level of expected accuracy for the APD forecast, a minimum number of time intervals are necessary to achieve the forecast horizon. For example, for the base year 2012 and the forecast horizon 2032, it is better to use the ten-year interval. Firstly, the accuracy of this time interval is higher. Secondly, fewer efforts are required to archive 2032 from 2012: for ten-year intervals, only one intermediate interval is necessary, when using five there are three intermediate intervals and for one-year intervals there are 19 intervals. This means that the calculation time costs are minimal using the ten-year interval. However, if detailed intermediate results are required and an adequate time budget is available, the five-year interval could be chosen. Thus, the analysis shows that, based on the time horizon of the planned APD forecast and the defined level of detail the appropriate time interval should be chosen.

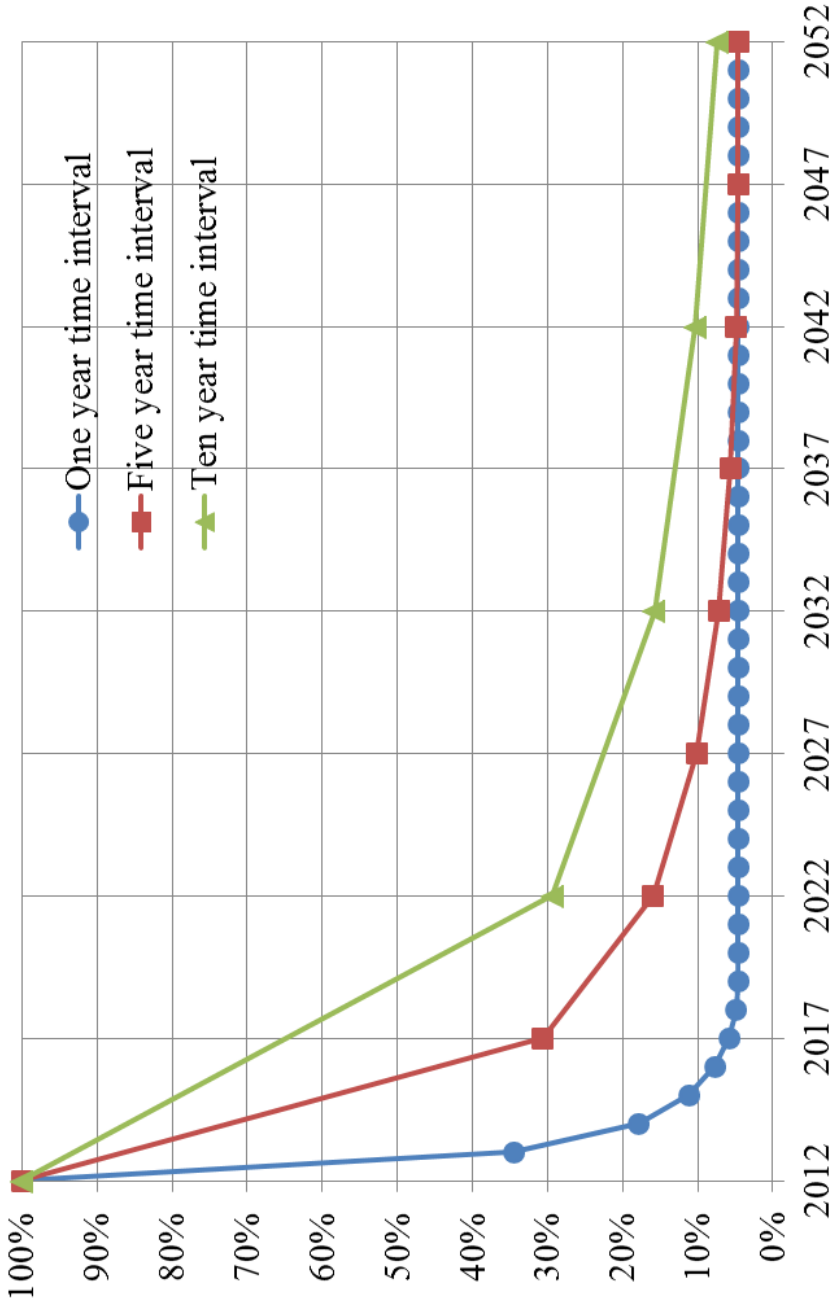


Fig.5.36. The average connection error propagations

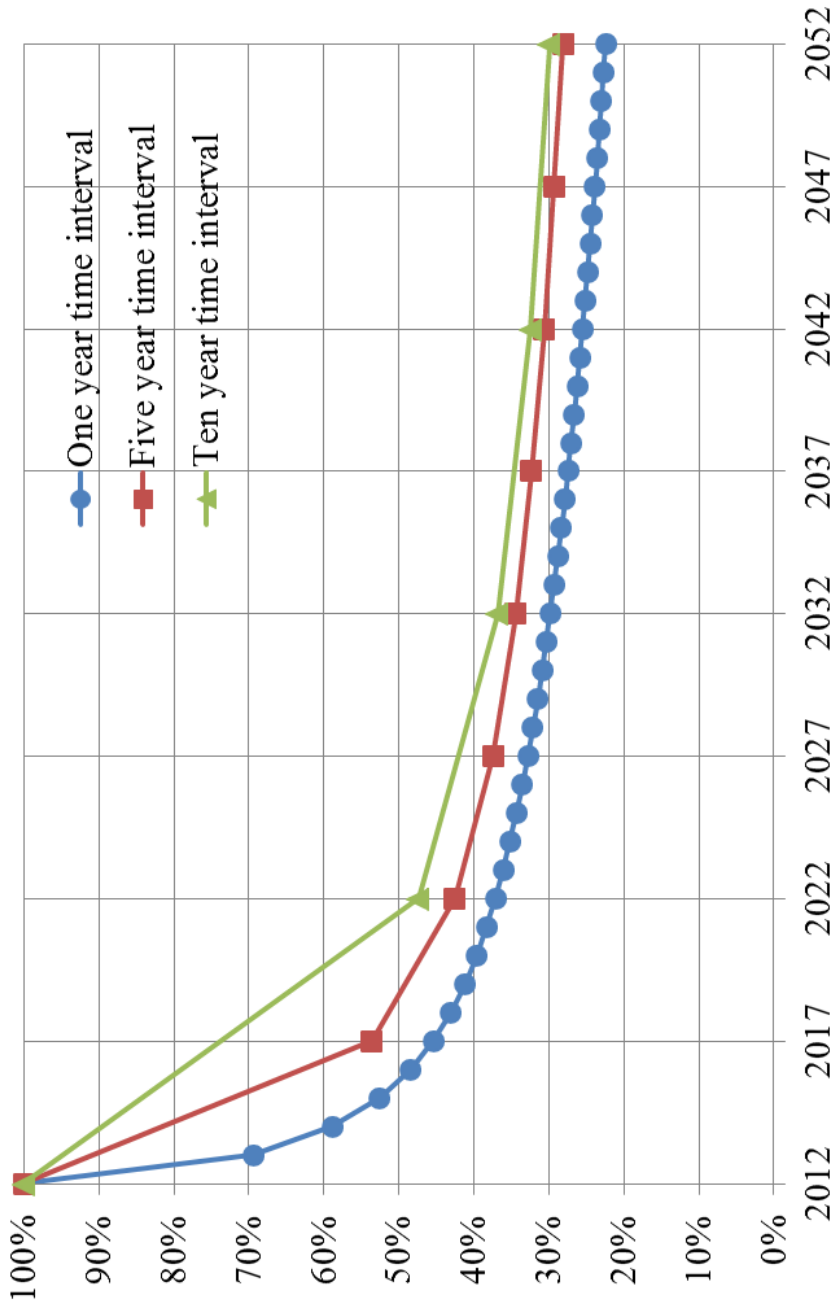


Fig.5.37. The average passenger percent covered by correctly predicted connections

5.5. Conclusion

In this chapter, the setup and validation of the APD model was considered. The proposed framework allowed the accuracy of the proposed APD model for a short, middle and long range time terms to be assessed. This was performed based on real data: the model was applied to 2002, 2007 and 2011 to forecast ADP topology and passenger number on every connection and then compared to the actual 2012 data. For validation preparation, the APD data and average airfares between settlements were collected from the ADI database, the settlements' GDP and population were obtained from the various databases. In addition, since the APD generation process is different for different types of settlements (in terms of wealth and population) and in order to increase the accuracy of the proposed method, the settlements were divided into clusters according to their socio-economic closeness. The analysis in Sub-section 5.1 showed that the probabilistic clustering of the normal mixture was more appropriate for settlement grouping than the other aforementioned approaches. In addition, it demonstrated the advantages of separating into nine clusters, since the cluster centers are well distinguished and their meaning is easy to interpret. Thus, the clusters were defined for the base year and clustering was applied to 2002, 2007 and 2011 to separate their settlements into groups.

The topology forecasting model validation was conducted based on the preparation module. The APD topology was presented as a weighed network. As discussed in Sub-section 5.2, among different topology forecast approaches, the similarity-based algorithms with local indexes for topology forecasting were most preferable for the APD network forecast in terms of accuracy and complexity. Nine indexes were considered which were modified for the weighed networks. The AUC and precision showed that the Weighted Resource Allocation index demonstrates the best performance since it gives a higher score to a non-existing connection if the nodes have many common neighbors with large weights. The boundary conditions were defined due to the specifics of the similarity-based algorithm with the WRA index. There are two boundary definition methods for adding and eliminating connections from the APD network. Either a fixed number of connections is added to or eliminated from the network at every time interval, or connections with a score of more or less than the

boundary score are added or eliminated at every time interval. In this study, the second approach was chosen for boundary definition. Since the APD network contains nine settlement types, they, in turn, have 45 cluster pairs. Thus, for every cluster pair, the boundary for adding and eliminating connections from the APD network was defined based on actual data from 2002, 2007, 2011 and 2012. Thereby, the boundary conditions for short, middle and long-term network topology forecasts were determined for every cluster pair. Using the defined WRA index and the boundary conditions, the forecasting topology model accuracies were assessed for 2012 from 2002, 2007 and 2011. The accuracy for newly added connections demonstrated poor accuracy on average, while the elimination accuracy was high. In addition, the forecasting topology model demonstrated a high accuracy for the remaining connections in the APD network. Despite the low average accuracy for predicting new connections, the high accuracy for remaining connections provided a high total accuracy for the forecasting topology model. Thus, it could be concluded that the APD connection disappearance and the connections remaining in the APD network are strongly connected to socio-economic indicator changes and the generation process for new APD connections is less related to these changes. Furthermore, the proposed clustering improved the topology forecasting model accuracy. The accuracy results obtained from the forecasting topology model were considered sufficient for applying the model.

The passenger forecast model was validated in Sub-section 5.3. There are two APD connection types: newly appeared and remaining connections. For newly appeared connections, the quantitative approach was applied and validated. The analysis of the cumulative curves showed that the newly appeared connections contain a lower passenger number on average in comparison to remaining connections. The accuracy assessment was conducted at given intervals. Based on the statistical analyses, the intervals for newly appeared connections were defined as integer numbers. Thus, the QA modeling result analysis for 2012 from 2002, 2007 and 2011 showed that the accuracy is higher the smaller the time interval. For remaining connections, the correlation between GDP growth and passenger number on every connection was applied. The accuracy assessment was also made in given intervals, but based on the statistical analyses of these connection types, the intervals were defined in percent. Analogically to the newly appeared connections, the analysis of the remaining

connections demonstrated that the accuracy is higher the smaller the time interval. Thus, based on the accuracies obtained, the total accuracy of the proposed APD modeling was assessed. This is performed in the same way as for the remaining connections – in intervals defined by percent. The one-year interval demonstrates the best accuracy from 2011 to 2012 and the ten-year interval shows the lowest accuracy from 2002 to 2012. The error propagation analysis was made according to these accuracies. The analysis showed that it is necessary to use various time intervals for different forecast horizons. Thus, for the short-term forecast, it is better to use the one-year interval, while the expected accuracy for the middle- and long-term forecasts is very low. The three analyzed time intervals, i.e. one, five and ten years, demonstrated sufficient expected accuracy for short, middle and long term forecasts respectively.

The validation of the proposed APD model demonstrated a sufficient accuracy value in terms of the correctly predicted connections and passenger number covered by them. The error propagation analysis showed the expected level of mistakes in the forecasts at various time intervals. Thus, using the validation results and databases obtained, the APD model can be applied to future socio-economic development scenarios in order to assess the future APD according to settlement level on a worldwide scale.

The conducted APD forecasting model development and validation covers the first and second step of the proposed research methodology shown in Section 3.2 – *Model development* and *Model validation* and objective which is indicated in Section 3.1. The next step in this thesis is to apply the validated APD forecasting model to socio-economic scenarios and model the future APD.

6. APD model application

In this chapter, the validated APD model is applied to the socio-economic scenarios in order to assess the future APD. These scenarios include socio-economic indicators of settlements worldwide (e.g. GDP and population). However, although scenarios are not available for settlements on a worldwide scale, they are broadly accessible at country level. Thus, the scenarios at country level were disaggregated to settlement level, to ultimately retrieve the GDP and population of the settlements. The APD was calculated based on the Global Environment Outlook 4 (GEO-4) socio-economic scenarios from the United Nations Environment Programme (UNEP) (2007). There are four different scenarios: Markets First, Policy First, Security First and Sustainability First. These scenarios allow the future APD for different considered developments to be evaluated. In order to assess future APD for a long term of 30 years, forecasting was performed from 2012 to 2042. Based on the model validation results, the forecast was made with the ten-year interval since it provides a higher expected accuracy of results. In addition, GEO-4 scenarios do not provide a scenario for airfares between settlements, thus, a simple airfare model was developed based on historical analysis.

This chapter is organized as follows: Section 6.1 presents the Simple Airfare Model (SAM) as developed by Ghosh and Terekhov (2015). Based on the analysis of the historical average air fare data on APD connections from 2002 to 2012 and the crude oil price, a simple equation for future average airfare according to the distance between settlements and the crude oil price was obtained. Using this equation, it is possible to define average airfare for every APD settlement pair. Section 6.2 describes the four GEO-4 scenarios and shows the modeled

APD results for each of them including cluster analysis, showing how settlements change clusters based on socio-economic indicators, the topology and the passenger forecasting model results for every time interval. Section 6.3 summarizes and verifies the results obtained in Section 6.2. In addition, the results are compared to other studies using a special transition metric which allow APD modeling results to be converted to revenue passenger kilometers (RPK).

6.1. Simple airfare model

This section describes the SAM. Since GEO-4 scenarios do not provide the average airfare on APD connections, a special model was developed. The basic framework for the SAM is depicted in Fig. 6.1.

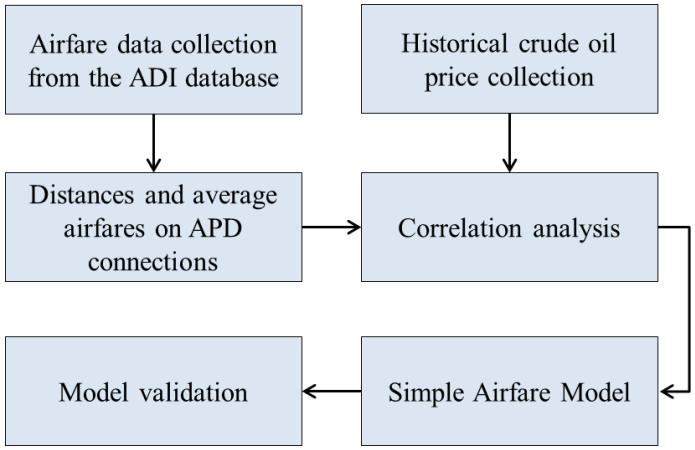


Fig.6.1. The basic SAM framework

A number of airfare models exist, such as the constant rate of return model used as part of the AIM project. The model allows fares to be calculated based on changes in operating costs which are required as an input (Dray et al, 2014). However, these models are not applicable within the presented APD modeling framework due to a lack of required input parameters. As shown by Ghosh and Terekhov (2015), the average airfare could be modeled and, therefore, assessed for the future based on indirect factors. GEO-4 contains scenarios for the crude oil price. The average airfare at settlement level, averaging all airfares for one year, depends on various main indicators such as crude oil price, distance between settlements and other factors. Thus, to reflect average airfare changes and to limit the scope of this study in

order to avoid unnecessary complexity, SAM takes the distances between settlements and the annual average crude oil price into account.

The historical airfare data on all APD connections from 2002 to 2012 was obtained from the ADI database. For the distance calculations, the geographical settlement coordinates were obtained from OpenFlights (2014) and OurAirports (2014). The historical data for the crude oil price were obtained from the U.S. Energy Information Administration (2014). In order to have consistency throughout the study, all prices were adjusted to US dollars as per 2005 values. The average airfares for 2002 to 2012 based on the distance between settlements are shown in Fig.6.2.

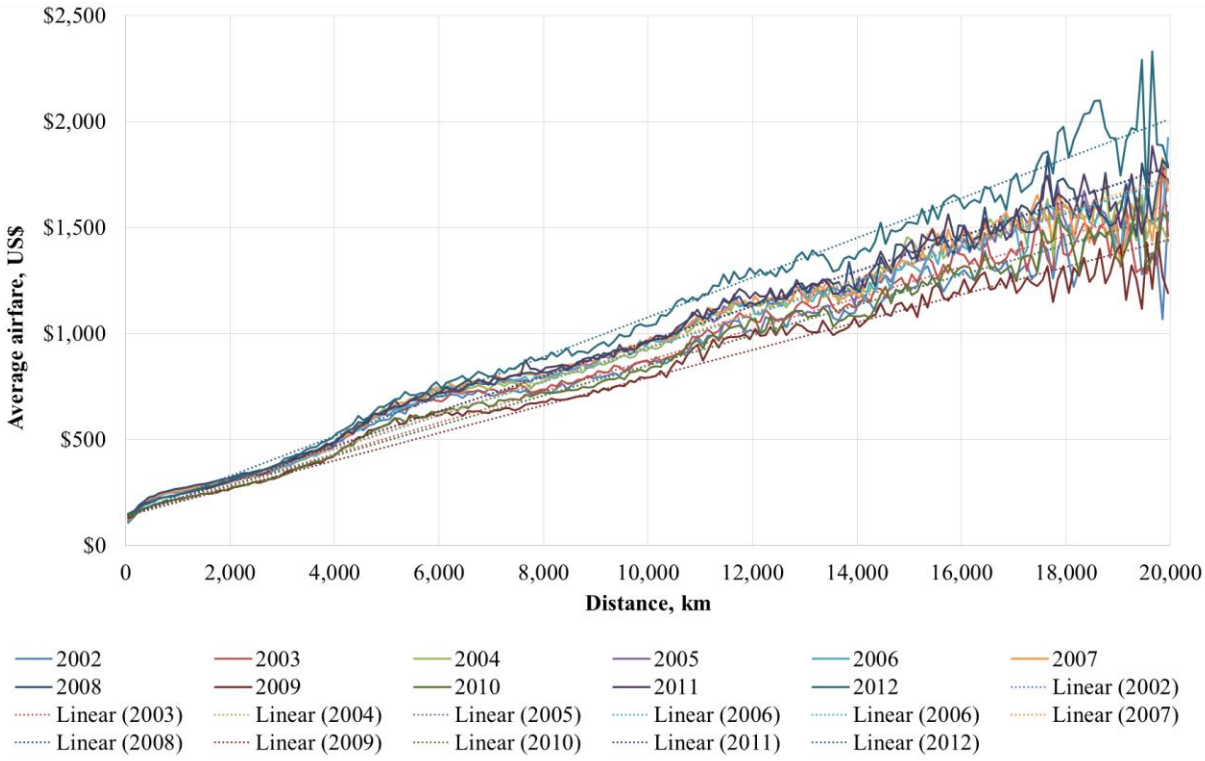


Fig.6.2. The average airfares for 2002 to 2012 based on the distance between settlements

Based on historical data, a simple linear equation was retrieved for the airfare (Eq. 6.1) according to the distance between settlements and the average airfare.

$$t_{xy} = a * d_{xy} + c \qquad \text{(Eq.6.1)}$$

Where t_{xy} is the average airfare between settlements x and y; a – the slope of the line; d_{xy} – a great-circle distance between settlements x and y; c – y-axis interception of the line. The great-circle distance between settlements x and y was calculated from the geographical settlement coordinates as follows:

$$d_{xy} = \text{acos}(\sin\phi_x * \sin\phi_y + \cos\phi_x * \cos\phi_y * \cos(\lambda_y - \lambda_x)) * R_0 \qquad \text{(Eq.6.2)}$$

Where, ϕ is the geographical settlement latitude in radians; λ – is the geographical settlement longitude in radians; R_0 is the average Earth radius which is defined as follows $R_0 = \frac{6378{,}388\ km + 6356{,}912\ km}{2} = 6367{,}65\ km.$

The historical analysis showed that the average airfare for all considered years starts from \$140 on average. Thus, the y-axis interception is 140, or, in other words, the minimum airfare in the study was assumed to be \$140. As seen in Fig. 6.2, the slope of the linear regression is varied for every year. Tab.6.1 reflects the R^2 for the regressions and slopes for every year from 2002 to 2012 and the annual crude oil price.

	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
R^2	0.9646	0.9653	0.9772	0.9721	0.9756	0.9705	0.9791	0.9668	0.9829	0.981	0.9832
Slope, *a*	7.0705	7.3204	7.8936	7.9006	7.8767	7.9685	8.2128	6.5038	7.081	8.226	9.3243
Oil price	\$25.01	\$28.85	\$38.26	\$54.53	\$65.17	\$72.50	\$96.94	\$61.74	\$79.61	\$111.37	\$111.67

Tab.6.1. R^2 for the regressions and slopes for every year from 2002 to 2012 and the annual crude oil price

Based on the table, a correlation between the slope and oil price figures can be seen. The correlation between crude oil price and the slopes are presented in Fig.6.3.

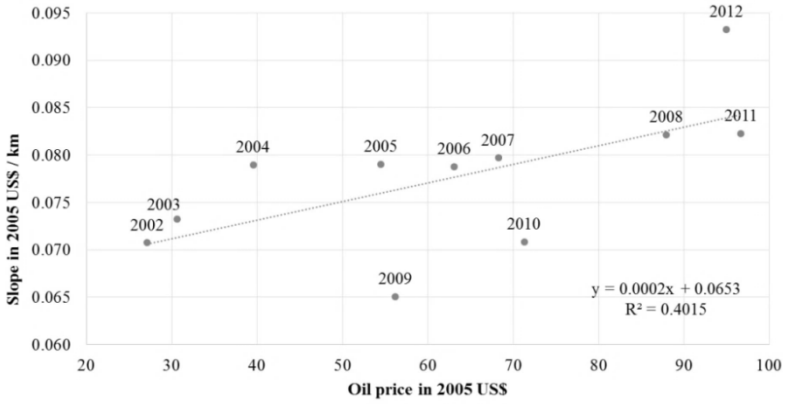


Fig.6.3. The correlation between crude oil price and the slopes

Fig.6.3 demonstrates relationships between the crude oil price and the slope from regression in Eq.6.1. The R^2 shows 0.4015. However, there was a high impact on the R^2 from the 2008 financial crisis and other external impacts. The R^2 for years 2002 to 2008 is approximately 0.76. Nevertheless, the regression obtained for the slope based on the crude oil price was used in Eq.6.1. Thus, assuming that the minimum average airfare is \$140, the next equation was retrieved:

$$t_{xy_G} = (2 * 10^{-4} * p_G + 0.0653) * d_{xy} + 140 \qquad \text{(Eq.6.3)}$$

Where p_G is the crude oil price in year G. Thus, knowing the crude oil price and geographical settlement coordinates, the average airfare can be modeled. For example, an APD connection in 2012 between Bangkok (Thailand) and Hong Kong had an average airfare of \$ 270.24 (in 2005 US\$). Using Eq. 6.3, the average airfare can be modeled. The distance between these two cities is 1,716 km, the crude oil price in 2012 was 90.27\$ (in 2005 US\$). Thus, the modeled airfare is 283.1\$. The modeled airfare is 4.75% higher than the original airfare, which is a reasonable result.

However, the SAM was validated on real dataset. This was achieved by modeling airfares on real connections from 2002 to 2012 and then comparing results with the actual data. The average airfares were assessed in 100 km intervals by distance. The average deviation in percent from the actual average airfare on 1000 km intervals are shown in Fig.6.4. In general the deviation is in interval [10%; -15%]. The vertical axis indicates the deviation in percent; the intervals in kilometers are shown in the outer contour.

The analysis demonstrates that the 5,000-6,000 km interval showed the lowest accuracy which is about 14% lower than the actual average airfare in this interval. The 16,000-17,000 km interval showed the highest result which is about 0.18% higher than the actual average airfare in this interval. The detailed annual validation results on 100 km intervals from 2002 to 2012 are shown in Appendix B. In general, the proposed model with given limitation demonstrates sufficient results. The conducted analysis and modeling allowed the model to be applied for estimating the future airfare on APD connections based on crude oil price and distance. Therefore, using an oil price scenario, the average annual airfare on a connection could be assessed. Thus, the proposed SAM was applied in this study to the topology

forecasting model and the passenger forecasting model in order to estimate the future APD between settlements worldwide.

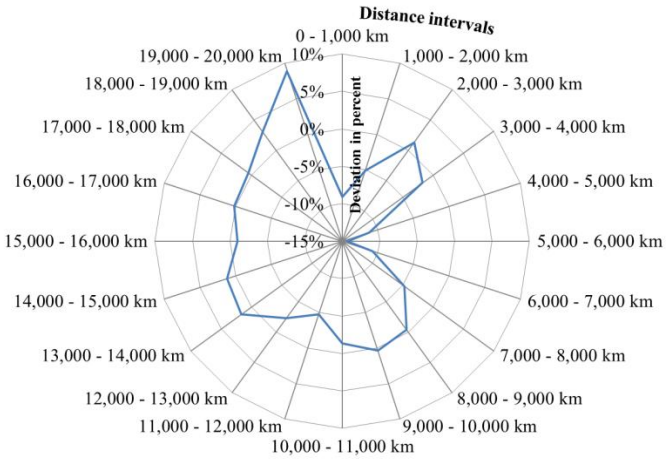


Fig.6.4. The average deviation in percent from the actual average airfare at 1000 km intervals

6.2. APD modeling for GEO-4 scenarios

The proposed and validated APD model (described in Chapter 4 and Chapter 5 respectively) was applied to four scenarios, called Markets First, Policy First, Security First and Sustainability First (Fig.6.5), obtained from the GEO-4 issued by the UNEP (GEO-4, 2007). The Outlook assessed the interaction between environmental change in the past and presented four scenarios for the future. The scenarios were built based on current socio-economic trends along divergent development paths in the future. In addition, it tried to determine their significance for the future environment, development and well-being. The scenarios have been developed up to 2050 using qualitative data to explore different policy approaches and societal choices. These scenarios included various indicators at country level e.g. country GDPs, populations and crude oil prices. These scenarios for global GDP, population and crude oil price are presented in Fig.6.6, Fig.6.7 and Fig.6.8 respectively. The provided scenarios allow the future APD to be assessed. However, they should first be disaggregated to settlement level.

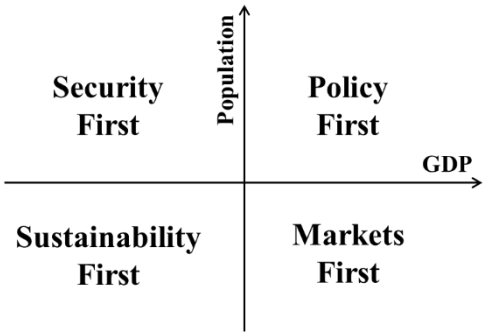


Fig.6.5. The four GEO-4 scenario positions in terms of GDP and population growth

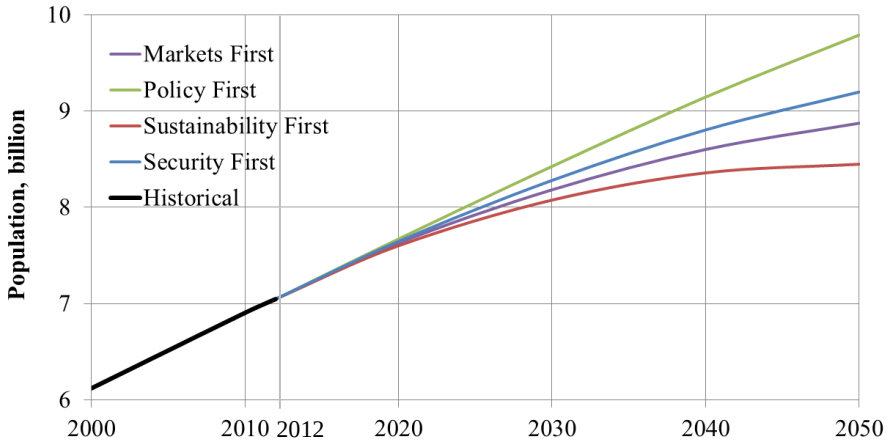


Fig.6.6. Historical and GEO-4 scenarios for world population

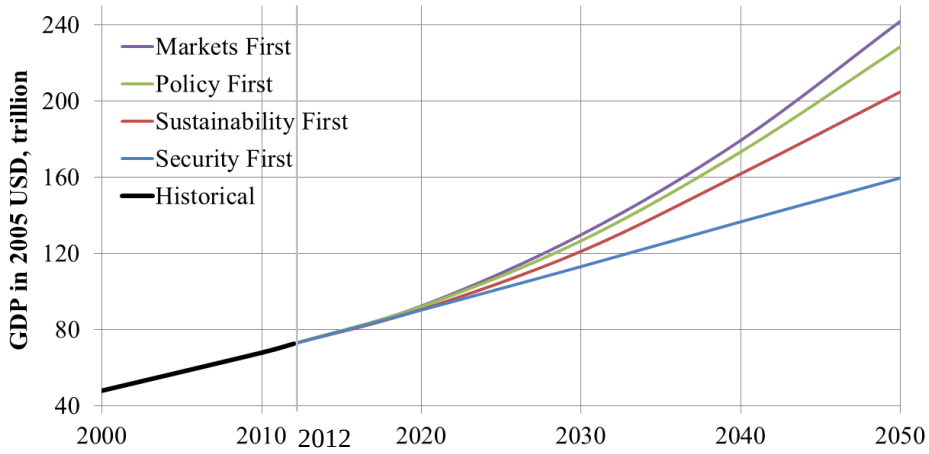


Fig.6.7. Historical and GEO-4 scenarios for world GDP

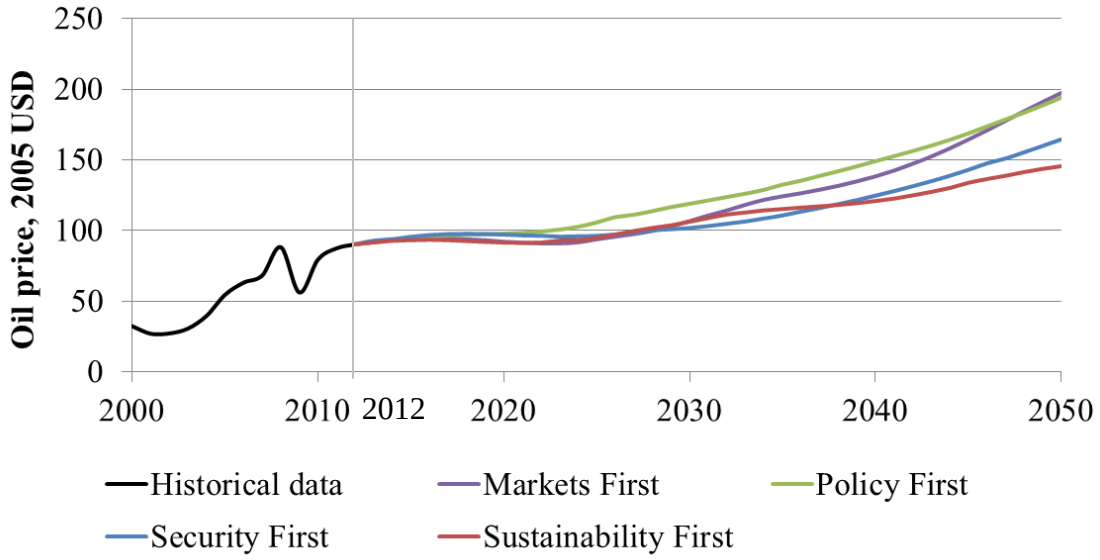


Fig.6.8. Historical and GEO-4 scenarios for the average annual crude oil price

As shown in Chapter 5, 4,435 settlements were retrieved from the 2012 APD network using the ADI database. For every settlement, the following data were obtained: settlement GDP and population and geographical coordinates. Thus, taking country population and GDP figures into account as well as the urbanization rates for every country which are known from the scenario data and obtained from the International Future model (The Frederick S. Pardee Center for International Futures, 2014), the scenarios were disaggregated to settlement level. However, the country number in the collected database for 2012 was different to the country number in the GEO-4 scenarios. The scenarios considered 183 countries, while the base year data contained 216. Nonetheless, these 33 countries (mainly islands such as St. Kitts & Nevis and Palau) contributed negligibly in terms of GDP, population and APD and were, therefore, eliminated from the collected 2012 data. Thus, the GEO-4 scenarios considered 4,120 settlements. The summarized GDPs and populations of settlements for the four scenarios are presented in Fig.6.8 and Fig.6.9.

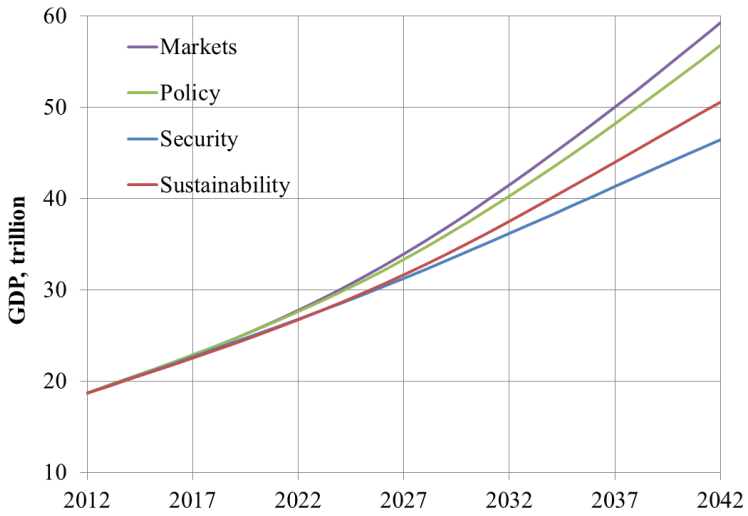


Fig.6.9. Summarized GDP for GEO-4 settlements for four scenarios

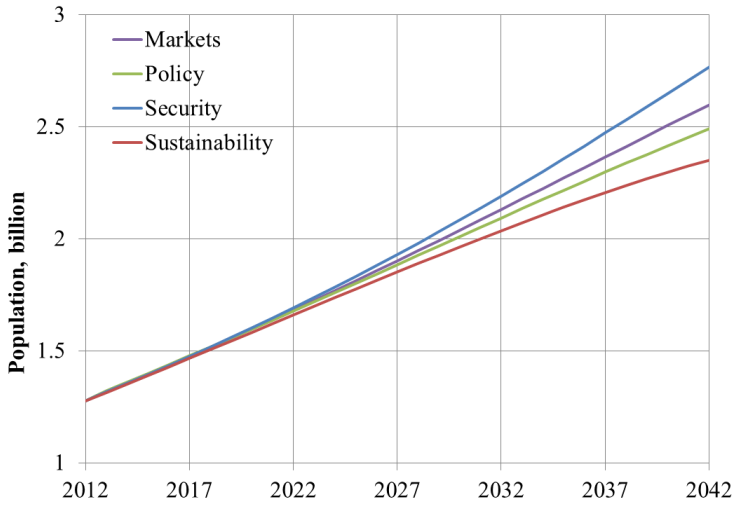


Fig.6.10. Summarized population for GEO-4 settlements for four scenarios

As shown in Chapter 5, the conditions required for the appearance of new settlements in the APD network are not clear and are difficult to predict. Thereby, the model only considered these 4,120 settlements, and did not add or eliminate them. The base year for forecasting APD was 2012. This means that the forecasting process starts from the 2012 APD network with these 4,120 settlements where the APD connections are eliminated or added. Thus, the data required for modeling were prepared for the APD forecasting based on the four socio-economic scenarios. The model was applied to the long-term time frame. Since the GEO-4 scenarios were presented on an annual basis to 2050 and based on the conducted validation,

the time interval for the APD was ten years, and, thereby the forecast horizon was 2042. Thus, four APD forecasts were made for the time points: 2022, 2032 and 2042. These four forecasts allowed future APDs to be assessed for different possible future developments. Brief general scenario descriptions are provided below. Key questions relating to the scenario assumptions are provided in Appendix C. The full description can be found in the GEO-4 (2007).

Markets First

The main characteristic of this scenario is an assumption that the market not only provides economic improvements, but also social and environmental development. The role of the private sector dominates in areas which were previously monopolized by governments, such as education, health, military, research and development. Free trade continues to expand worldwide. Ecosystem services are turned into commodities in order to increase privatization and trade. The expanding trade and economic investments negatively influence formal environment protection, slowing this process down. Thus, fossil fuels are still the main energy source. Water privatization increases water usage efficiency. However, decreasing subsidies in most regions lead to difficulties paying for water. Nevertheless, the global population in 2050 reaches about 9.2 billion people (Fig.6.6) and global GDP is about 5 times higher than in 2000 (Fig.6.7).

Policy First

The main characteristic of this scenario is a very high level of centralization in order to balance strong economic growth and decrease potential environment and social impacts. The governments are solving obvious problems (e.g. access to safe water in many parts of the world) and concealed issues (e.g. climate change) in the new century. The governments have adopted a more holistic approach to respond to environmental challenges –economic growth is no longer considered without assessing its social and environmental impacts. The scenario offers high economic and political integration. Inefficient subsidies are reduced in order to decrease over-exploitation. Investments in science and technology grow. Climate change and its associated impacts are the main concern. However, when the investments in R&D motivate efforts to increase energy efficiency, the total energy consumption continues to increase. Thus,

fossil fuels still dominate. Nevertheless, the global population in 2050 reaches about 8.2 billion (Fig.6.6) and the global GDP is about five times larger than in 2000 (Fig.6.7).

Security First

The main characteristic of this scenario is a concentration on security. Thus, the movements of people are more restricted than they used to be. Migration and the movement of goods across borders are also limited. This is due, in part, to ongoing conflicts in many regions of the world. At the same time, the world becomes more crowded as the population grows. Security costs also rise at the expense of other areas such as R&D and technology. Governments play a strong role in decision making. International institution authority decreases. These changes cause less attention to be devoted to environmental issues. Total energy consumption increases significantly. Coal usage grows rapidly, approaching natural gas and oil levels. Global population in 2050 reaches about 9.7 billion (Fig.6.6) and global GDP is about four times larger than in 2000 (Fig.6.7).

Sustainability First

The main characteristic of this scenario is that authorities at all levels (local, national, regional and international) consider environmental and social issues. Governments solve a large number of problems. At the same time, the private and civil sectors do not wait for governments to act. In the energy sector, an effort to balance the desire to reduce overall consumption with the need to address issues such as fuel is made. With the increased investments, this challenge is met in more environmentally friendly way. The mix of fuels changes significantly: oil and coal usage decrease to the point where more is produced by solar and wind. Natural gas is the dominant source of energy. For this scenario, global population in 2050 reaches about 8.5 billion (Fig.6.6) and the total GDP is about five times larger than in 2050 (Fig.6.7).

Application of the APD model allowed the future APD for the proposed scenarios to be assessed. In addition, using the scenarios from one source united by a common philosophy, the introduced APD model sensitivity could be assessed to help verify the results. Thus, the APD was calculated based on each of the aforementioned scenarios. The modeling results are presented below. For every scenario, the cluster dynamics were initially calculated. The

cluster dynamics show how settlements change clusters within the scenario. Secondly, based on defined cluster content for all three time points (2022, 2032 and 2042), the topology forecasting model and the passenger forecast model were applied for every time interval. The scenarios were considered in the following order: Markets First, Policy First, Security First and Sustainability First.

6.3. GEO-4 scenario results

This section presents the APD model results for the GEO-4 scenarios. The total settlement GDP, population and crude oil price for this scenario are shown in Fig.6.11.

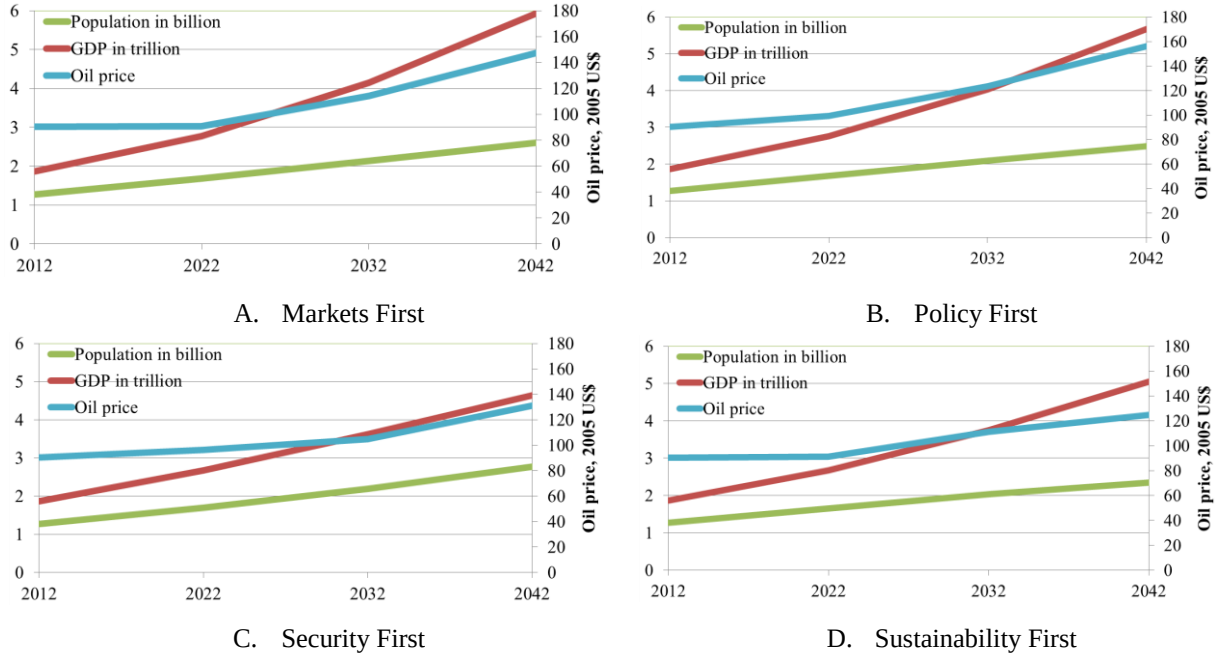
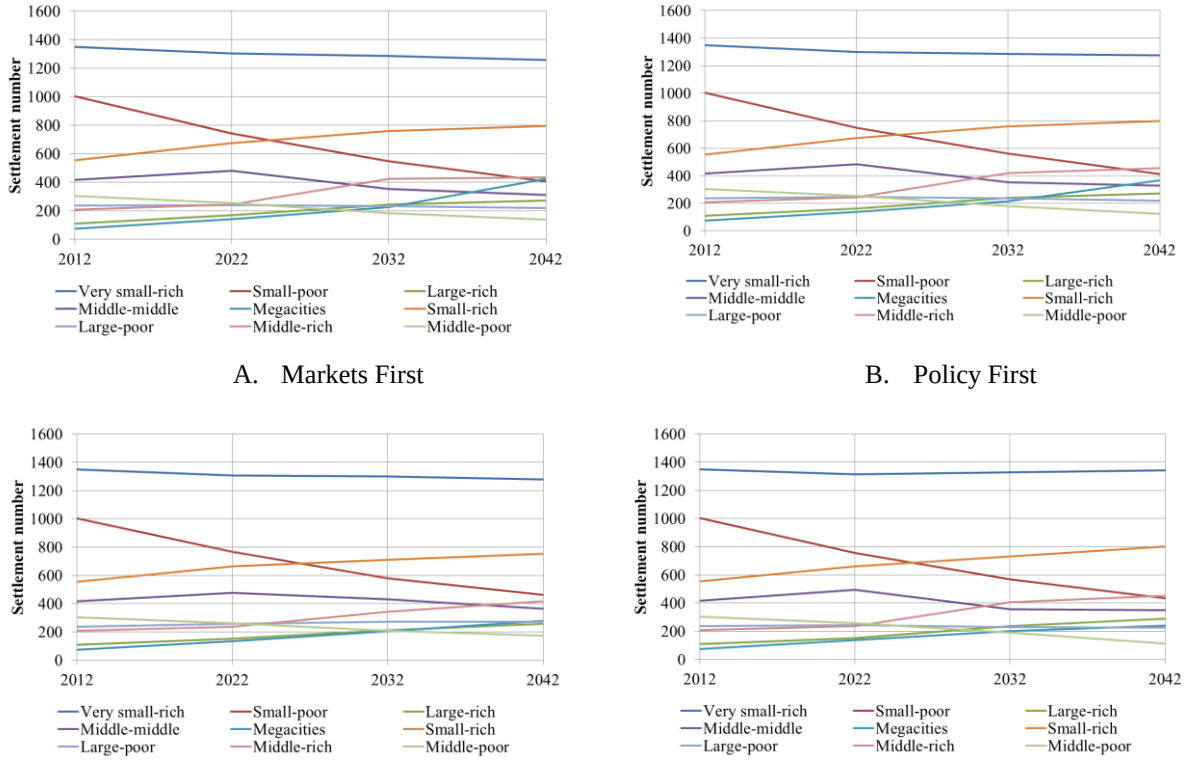


Fig.6.11. Total settlement GDP, population and oil price for the GEO-4 scenarios

The cluster dynamics (Fig.6.12) for clusters were calculated using the cluster centers obtained in Section 5.1 and the provided scenario shown in Fig.6.11.



A. Markets First

B. Policy First

C. Security First

D. Sustainability First

Fig.6.12. Cluster dynamics for the GEO-4 scenarios

Based on the GEO-4 scenarios, *small–poor* and *middle–poor* clusters showed significant decreases in the number of settlements. This is because GEO-4 scenarios were positive, where almost all settlements demonstrated population and GDP growth. In addition, the reason for this reaction was that the cluster centers remain fixed as in the base year 2012. The cluster dynamics did not take future inflation into account and, therefore, they showed cluster content from a 2012 perspective. Thus, cluster dynamics shows settlements moving to more "powerful" clusters. As a consequence, *middle-rich*, *large-rich* and *megacities* showed significant increases in the Markets and Policy First scenarios. However, the Security and Sustainability First scenarios demonstrated slower increases for those clusters.

Transition diagrams of 4,120 settlements in clusters between the base year 2012 and the last year of the scenario, 2042, are presented in Fig.6.13. The diagram shows nine clusters on three levels according to settlement population: small, middle and large. Settlements which either remain in clusters or change are indicated in percent of the total settlement number. Arrows demonstrate which cluster settlements are moved. The diagram shows transitions

between clusters for more than 1% of settlements. Based on this diagram, tendencies of moving settlements between clusters could be seen.

As seen in Fig.6.13, the settlements have a tendency to move into strong clusters from *poor* to *rich* clusters and from *small* to *middle* and *large* clusters. The Markets First scenario considered the market as the main driver, settlement development in terms of GPD and population showed a rapid growth. At the same time, the Policy First scenario showed a very high level of centralization in order to balance strong economic growth and decreased potential environment and social impacts, settlement development in terms of their GPD and population demonstrate growth. Thus, about 46% and 44% of all settlements for the Markets (Fig.6.13.A) and Policy First (Fig.6.13.B) scenarios change cluster when comparing 2012 to 2042.

Since the Security First scenario shows a very high focus on security, the economic growth is very slow due to restricted migration and movement of goods across borders. The Sustainability First scenario's main characteristic is that authorities at all levels consider environmental and social issues, and all decisions in economy and sociology are made in regards to these issues. Due to this, the economy and population show low levels of growth. Thus, growth in these two scenarios is lower than in the Markets First and Policy First and, thereby, less settlements move between clusters and less settlements move into *richer* and *larger* clusters. Thereby, about 37% and 39% of all settlements for Security (Fig.6.13.C) and Sustainability First (Fig.6.13.B) scenarios change cluster when comparing 2012 to 2042.

Cluster contents for all three time points of the forecast 2022, 2032 and 2042 were obtained. The topology forecasting and passenger forecasting models were applied using the cluster contents as input.

Using the topology forecast model, three APD networks were forecasted for 2022, 2032 and 2042 from the base year 2012. The total connection number for three forecasted APD networks as well as for the 2012 base year is depicted in Fig. 6.14.

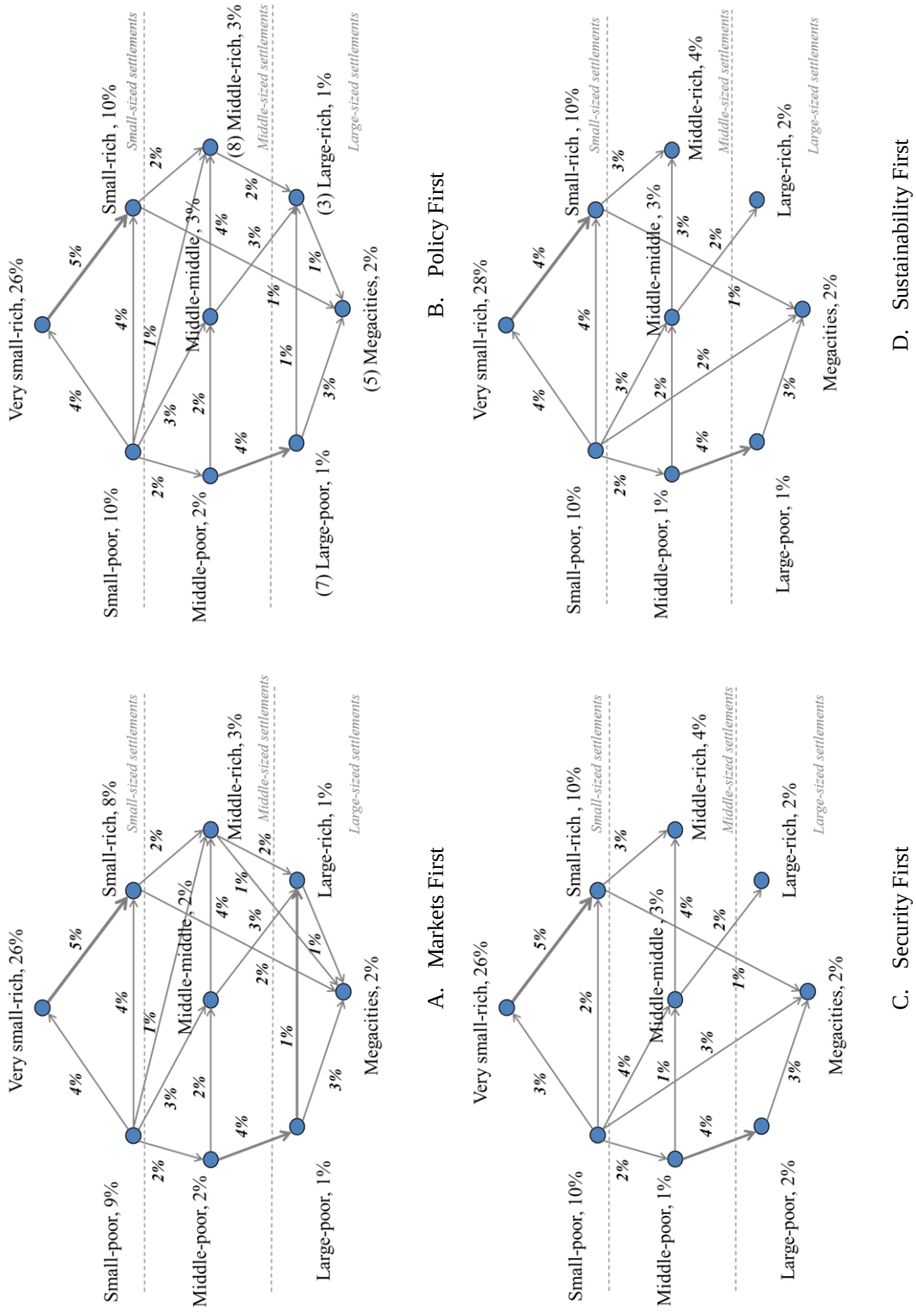
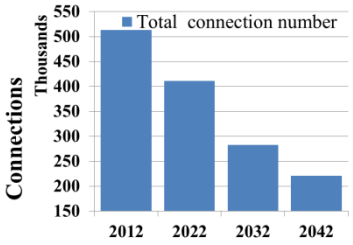
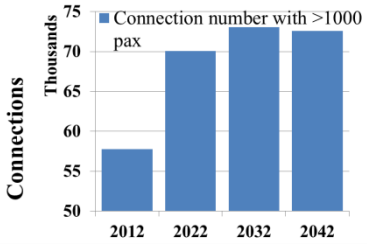


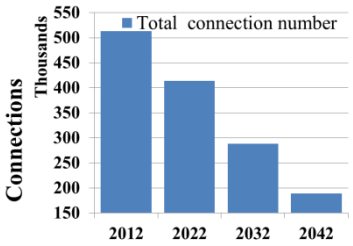
Fig.6.13.The transition diagram of settlements in clusters between the base year 2012 and last year of the scenario 2042 for the Markets First scenario



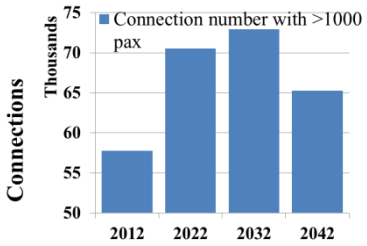
A. The forecasted total APD connection number for the **Market First** scenario 2022, 2032 and 2042 and the base year 2012



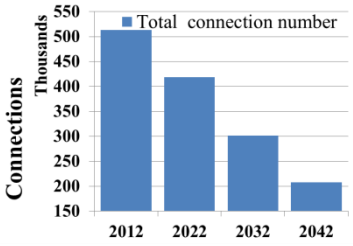
B. The forecasted APD connection number with >1,000 passengers for the **Market First** scenario 2022, 2032 and 2042 and the base year 2012



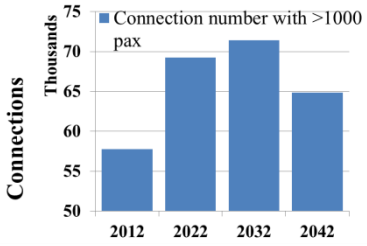
C. The forecasted total APD connection number for the **Policy First** scenario 2022, 2032 and 2042 and the base year 2012



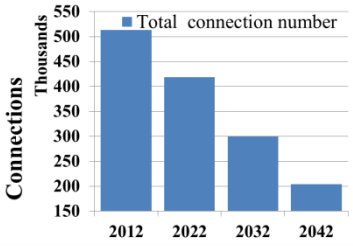
D. The forecasted APD connection number with >1,000 passengers for the **Policy First** scenario 2022, 2032 and 2042 and the base year 2012



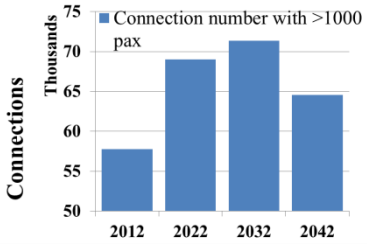
E. The forecasted total APD connection number for the **Security First** scenario 2022, 2032 and 2042 and the base year 2012



F. The forecasted APD connection number with >1,000 passengers for the **Security First** scenario 2022, 2032 and 2042 and the base year 2012



G. The forecasted total APD connection number for the **Sustainability First** scenario 2022, 2032 and 2042 and the base year 2012



H. The forecasted APD connection number with >1,000 passengers for the **Sustainability First** scenario 2022, 2032 and 2042 and the base year 2012

Fig.6.14. Forecasted APD connection number for the Market First scenario 2022, 2032 and 2042 and the base year 2012

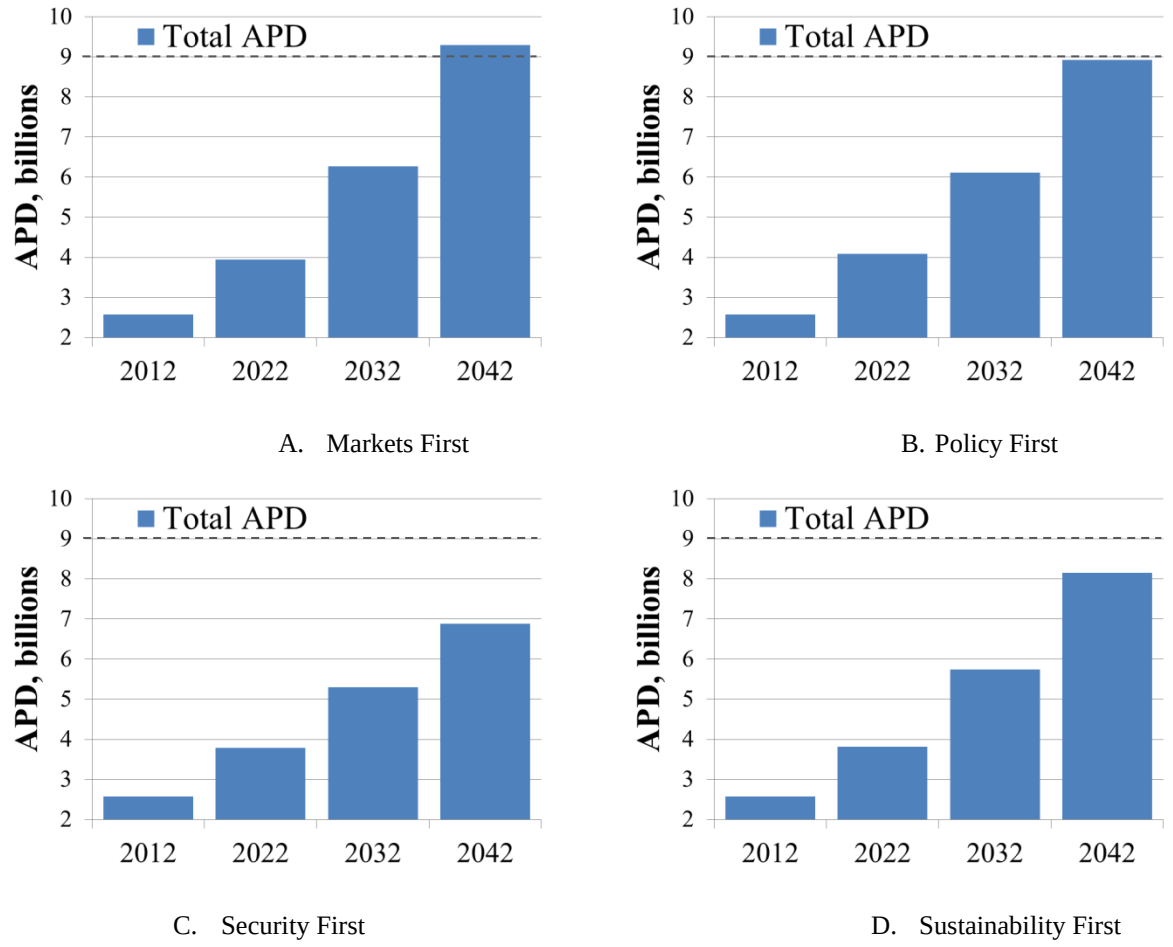
As seen in Fig.6.14, the total connection number in the APD network decreases, while connections with more than 1,000 passengers increase until 2032. This can be explained by the boundaries definition in Sub-section 5.3.2 and settlement movement between clusters. For example, settlement *A* in the base year 2012 was in the *small-poor* cluster and had ten connections with settlements from *small-poor* cluster and one connection with a settlement in *large-rich* cluster. For the *small-poor* – *small-poor* cluster pair, the boundary for link elimination is 0.0002. If the connection score calculated is lower than this boundary using the WRA index (shown in Section 5.2), – the connection is eliminated from the network. The elimination boundary for the cluster pair *small-poor* – *large-rich* is 0.09. Settlement *A* in 2022 changes cluster to the *middle-rich* due to the GDP and population growth, while connected settlements do not change their clusters. Therefore, for the new cluster pair *middle-rich* – *small-poor* the elimination boundary is 0.02 and for the *middle-rich* – *large-rich* it is 0.21. Thus, it is likely that settlement *A* loses the connections transferring from the weak cluster *small-poor* to the stronger *middle-reach*. Following this logic, it can be concluded that the more settlements move between clusters, the more previous connections could be lost. However, connection numbers with more than 1,000 passengers increased due to the rapid socio-economic growth of the settlements. Thereby, the connection number change could be explained by the scenarios, where about half of all settlements changed clusters, while the socio-economic indicators demonstrated rapid growth. In addition, the settlement set is fixed in the proposed modeling approach. In other words, as shown before, settlements are not added to or eliminated from the APD network. Since all the settlements have a positive development trend in terms of GDP and population, they tend to move to the strong clusters, and the weak clusters do not fill the formed gap with newly added settlements, which would generally be placed in weak clusters. The total connection numbers for 2012, 2022, 2032 and 2042 as well as the new and eliminated connection numbers are presented in Tab.6.2.

	2012	2022				2032				2042			
		MF*	PF*	SF*	SuF*	MF	PF	SF	SuF	MF	PF	SF	SuF
Total connection number	513,596	410,720	413,683	418,894	418,857	282,434	288,502	301,275	300,067	221,048	189,090	207,922	203,986
Connection number, >1000 pax	57,756	70,041	70,544	69,232	69,014	73,038	72,955	71,414	71,365	72,590	65,274	64,842	64,587
Added connections	x	46,064	46,950	47,657	47,335	10,170	10,349	10,240	11,286	3,344	3,369	3,595	3,461
Eliminated connections	x	148,940	146,863	142,359	142,074	138,456	135,530	127,859	130,076	64,730	102,781	96,948	99,542
Settlements, changed the cluster, %	x	17%	16%	14%	15%	19%	18%	14%	16%	16%	15%	11%	12%

Tab.6.2. The total connection numbers for 2012, 2022, 2032 and 2042 as well as the new and eliminated connection numbers and the changed settlement clusters in percent for the GEO-4 scenarios

*MF – Markets First, PF – Policy First, SF – Security First, SuF – Sustainability First

As seen, the number of added and eliminated APD connections decreases. However, the number of the changed settlement clusters is the same on average. In 2022, the largest APD connection number is eliminated and added to the APD network compared to other years. This is because many settlements changed their clusters, and, therefore, as described above, many connection scores are lower than elimination boundaries and, therefore, these connections are eliminated. Based on the obtained APD networks, the passenger demand model was applied. The total APD number for the four scenarios is depicted in Fig.6.15.



A. Markets First

B. Policy First

C. Security First

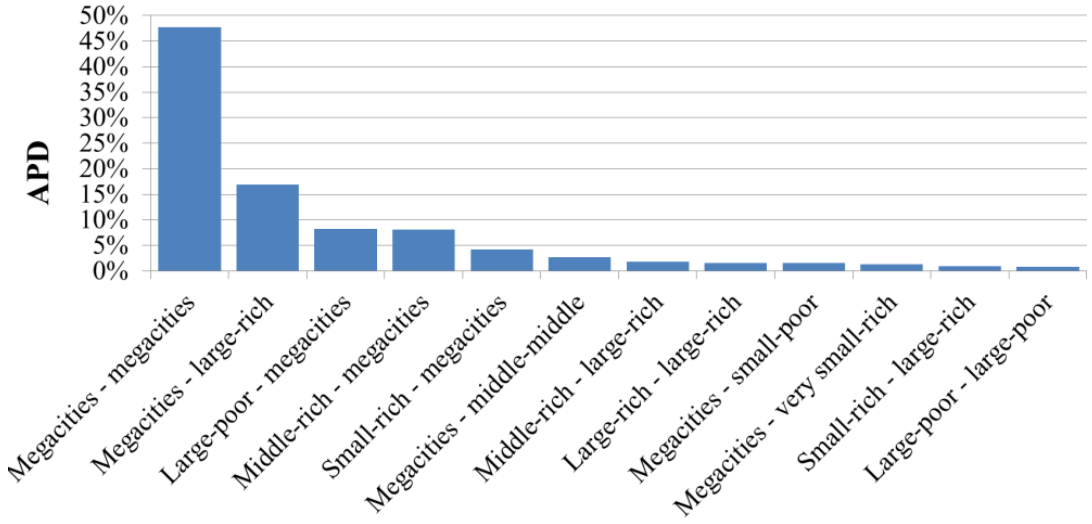
D. Sustainability First

Fig.6.15. The total APD for GEO-4 scenarios

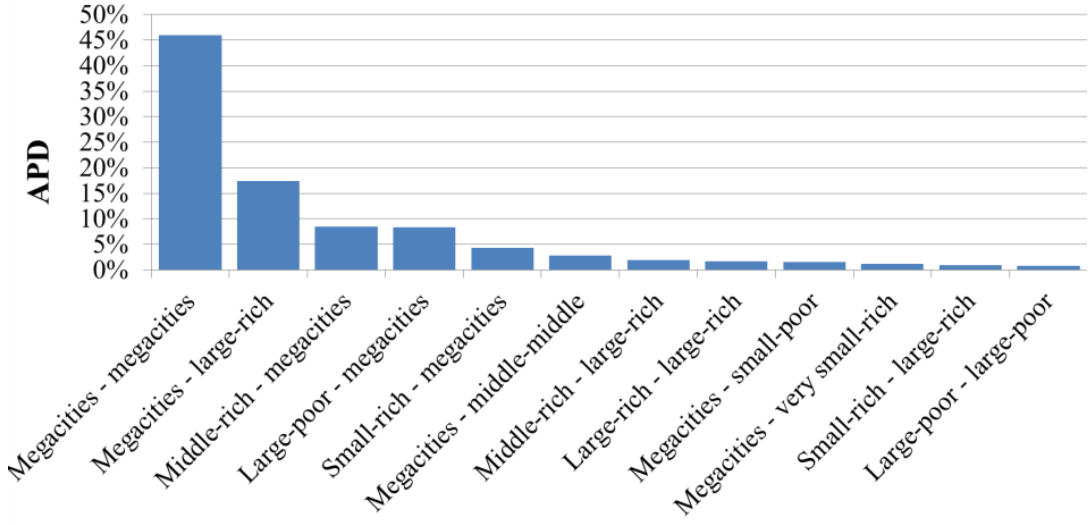
The total number of passengers increases in all scenarios (Fig.6.15). Despite a reduction in the total number of connections, the total number of passengers constantly grew. Indeed, the number almost tripled by 2032 compared to 2012 in all scenarios. This could be explained by the high level of growth in large settlements. For instance, for the Market First scenario, the

number of settlements in the *megacities* cluster changed from 73 in 2012 to 422 in 2042 and from 108 to 273 in the *large-rich* cluster for the same period; the passenger number in the *megacity* – *megacity* cluster pair is 16 times larger, in *large-rich* – *megacities* – five times larger in 2042 compared to 2012.

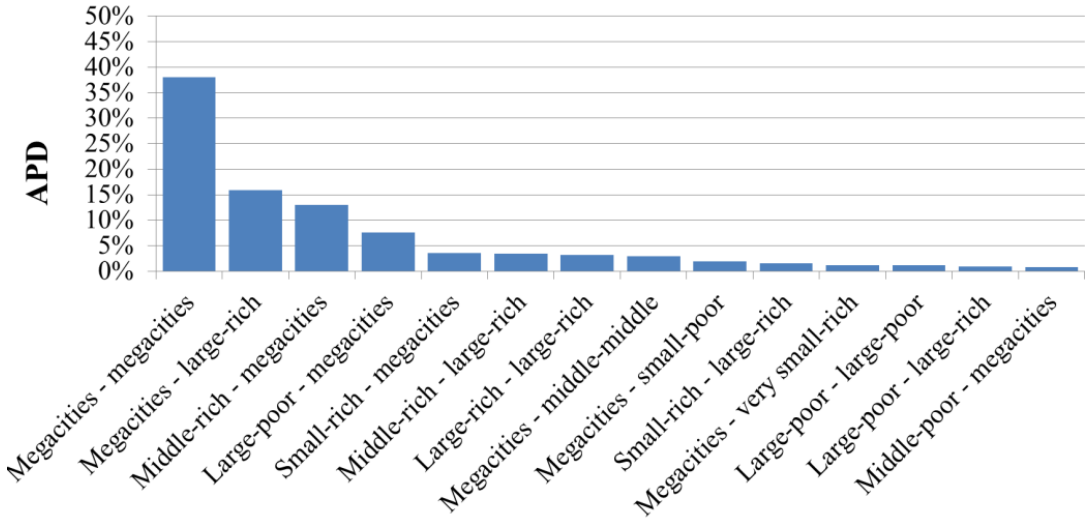
In addition, the *megacity* – *megacity* cluster pair generates about 40% of the total APD in 2042 in every scenario. Moreover, cluster pairs with *megacities* contribute more than 85% to the total APD in 2042. Fig.6.16 demonstrates cluster pairs representing 95% of the total APD in descending order by their contribution to the total APD in 2042.



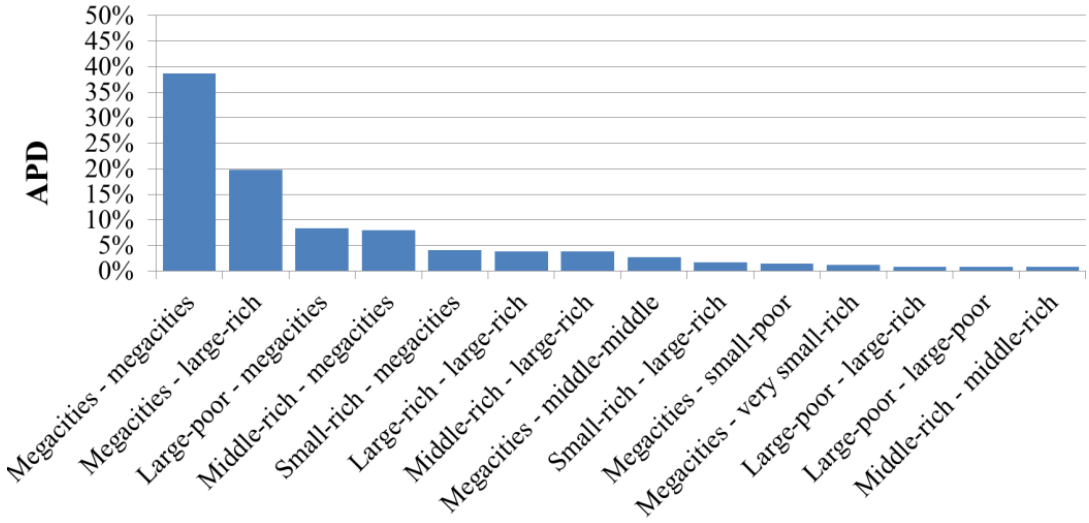
A. Markets First



B. Policy First



C. Security First



D. Sustainability First

Fig.6.16. APD share in 2042 for clusters pairs generating 95% of the total APD for the GEO-4 scenarios

The Asian region showed the largest growth. In 2012, its share in total APD was approximately 32% while in 2042 it is about 50% for all scenarios. Europe and North America both had 24% shares in 2012, while in 2042 the level is about 12% and 16% respectively. The same tendency could be traced at APD connection level. For example, the APD between Chengdu and Guangzhou in 2012 (both *large-poor* cluster) was 2,432,121 and in 2042 (both *megacity* cluster) the forecasted figure for the Market First scenario is 21,504,179, which is more than eight times larger than in 2012. For the Policy First scenario, for instance, the APD between Hanoi and Ho Chi Minh City in 2012 (*large-poor* and *megacities* clusters respectively) was 3,980,720 and in 2042 (stayed in the same clusters) the forecasted figure is

25,361,721, which is more than six times larger than in 2012. For the Security First scenario, for example, the APD between Rio De Janeiro and Sao Paulo (i.e. both in Brazil) and in the *megacities* cluster pair was 7,657,758 in 2012 and in 2042 (stayed in the same clusters) the forecasted figure is 18,726,589, which is more than two times larger than in 2012. For instance, for the Sustainability First scenario, the APD between Jeju (South Korea) and Seoul (South Korea), which are in *middle-rich* and *megacities* clusters respectively, in 2012 is 8,977,671and in 2042 (they stayed in the same clusters) the forecasted figure is 13,867,809, which is more than 1.5 times larger than in 2012.

In Appendix D, the top 15 APD connections in 2042 are presented by APD numbers for all GEO-4 scenarios, showing the dominance of the Asian region in APD.

6.4. Consolidated summary of scenario results and verification

After considering each of the four scenarios in turn, the results shall now be summarized and verified. All scenarios were found to be positive in terms of GDP and population growth. The GEO-4 scenarios could be divided into two groups: large GDP and population growth (Markets and Policy First scenarios) and slow growth of these indicators (Security and Sustainability First scenarios). In Fig.6.17, historical APD and the forecast for the four scenarios are shown.

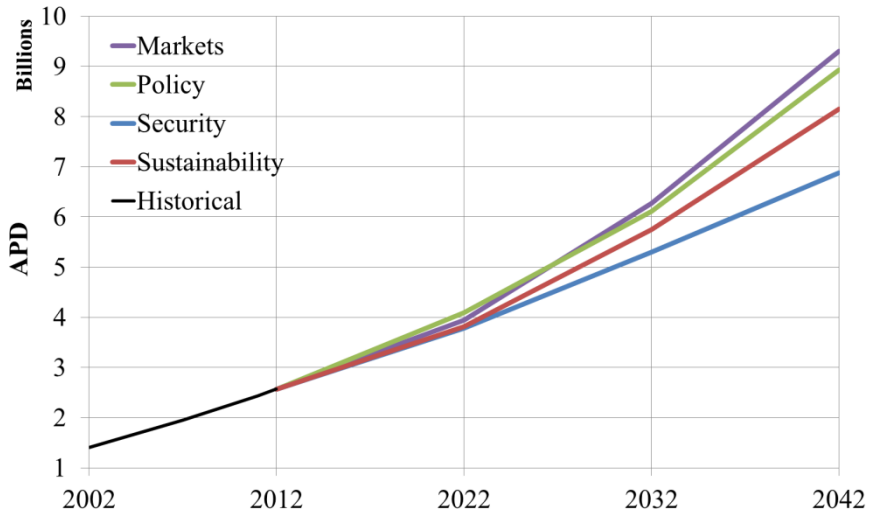


Fig.6.17. Total historical ADP and the forecast for the four scenarios

As demonstrated, all scenarios follow the historical trend with some deviations. The deviations are the reason for the scenario description. The Markets First scenario, which concentrated on free markets, demonstrated the highest future APD, since the GDP and population growth levels there are the largest of all the scenarios. The contrary can be said of the Security First scenario. Due to the imposed restriction of limited migration and trade, this scenario shows the lowest forecasted APD figures. The Policy and Sustainability First scenarios demonstrate a middle APD forecast. They show average GDP and population growth compared to the Markets and Security scenarios. Thus, verifying the expected future APD for four scenarios on the scenario descriptions, it could be seen that they match and the APD forecasts follow the scenario logics, and as well showing the APD forecasting model sensitivity. Despite the APD growth, the APD connection number decreases for each scenario, as shown before. The APD connection numbers are shown in Fig.6.18.

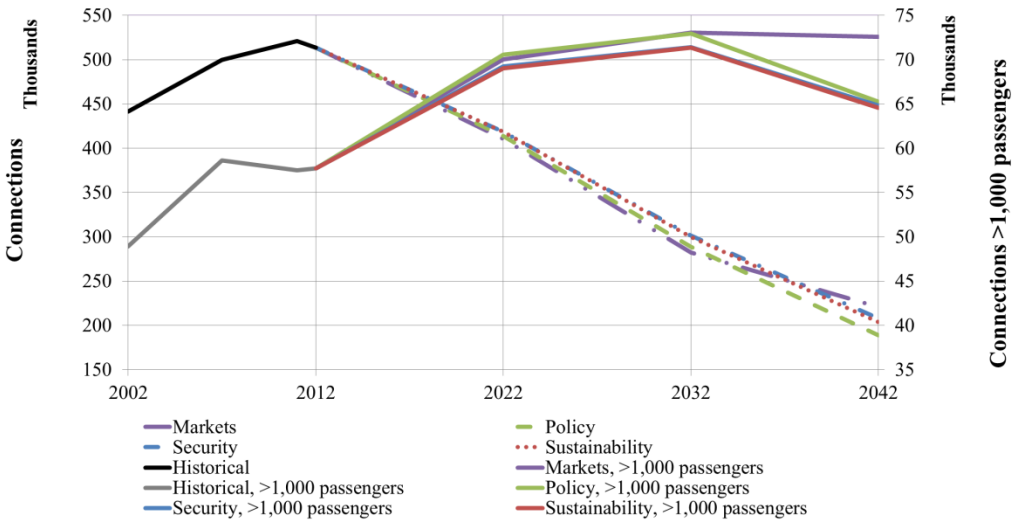


Fig.6.18. All historical and forecasted ADP connections as well as APD connections with more than 1,000 passengers for the four scenarios

The connection number decrease is explained by the cluster pair boundary condition definition for the elimination and the cluster dynamics processes. Since all scenarios are positive in terms of GDP and population growth, settlements from *poor* and *small* clusters tend to move to the *large* and *rich* clusters. Thus, when a settlement changes cluster to a more superior one, the boundaries for all its pairs change correspondingly. Thus, even with growing GDP and population rates for almost all settlements, these settlements are not able to maintain

links where the link weight is higher than the new boundary. An example of this situation and corresponding explanation are shown in Section 6.3. However, the APD connection number with more than 1,000 passengers increases. It shows that within the forecasting process, connections mainly with a few passengers are eliminated from the APD network since they possess low socio-economic indicators. Thus, despite the rapid decrease in connection number, the total passenger number on them is very low compared to the total passenger number and can be disregarded for this study. The passenger percentages on eliminated connections of the total passenger number for the four scenarios are shown in Tab.6.3.

	2022	**2032**	**2042**
Markets First	0.27%	0.60%	0.63%
Policy First	0.25%	0.59%	1.00%
Security First	0.26%	0.58%	0.98%
Sustainability First	0.26%	0.55%	0.93%

Tab.6.3. The passenger percentage on eliminated connection of the total passenger number for the four scenarios

The passenger number on eliminated connections did not exceed 1% of the total APD in a given year. Thus, the weak connections which were not able to maintain the connection score above the elimination boundary were erased from the APD network. In addition, on the other hand, the newly added connections definitely demonstrated a strong ability to establish connections based on the strong socio-economic indicators, which allowed for a connection score higher than the elimination and adding boundaries. Nevertheless, the study results of the modeled APD at global level for all four scenarios shown in this section and Section 6.1 displayed APD growth. However, in order to identify the difference and verify the obtained forecasts, they were compared to other studies. In addition, it allowed the differences between existing forecasts and forecasts from the presented study to be identified.

The four APD forecasts were compared to those provided by Airbus, Boeing and ICAO FESG. These forecasts were made in revenue passenger kilometers (RPK). RPK for one year are measured as the total number of kilometers travelled by all passengers on all routes. The APD forecasting model presented in this study does not define the real routes for passengers, yet it does provide information for the passenger number on origin – destination pairs. Thus, it is possible to obtain the demand passenger kilometers (DPK). In contrast to RPK, DPK are defined as the number of APD by great circle distance. As shown by Ghosh and Terekhov

(2015), the APD network is the passenger number on the origin – destination pairs (in other words, a point-to-point network), when the RPK is based on the real route network used by airlines (with hubs, i.e. a so-called hub and spoke network). Thus, the APD network is the ideal version for the real flight route network from a passenger perspective. In other words, in the ideal flight routes network the passenger is able to take a direct flight to any destination. Using DPK and RPK, the directness of the network could therefore be obtained. The directness shows how the real global airline route network (represented in RPK) is close to the ideal point-to-point network (represented in DPK) which is the APD network. Therefore, the directness factor DF is presented as follows:

$$DF = \frac{DPK}{RPK} \qquad \text{(Eq.6.1)}$$

For the ideal flight network, the DF is 1. Based on historical data obtained from the ADI database (Sabre Airline Solutions, 2014), the annual RPK and DPK are calculated for the period 2003-2012 in addition to the DF (Fig.6.19).

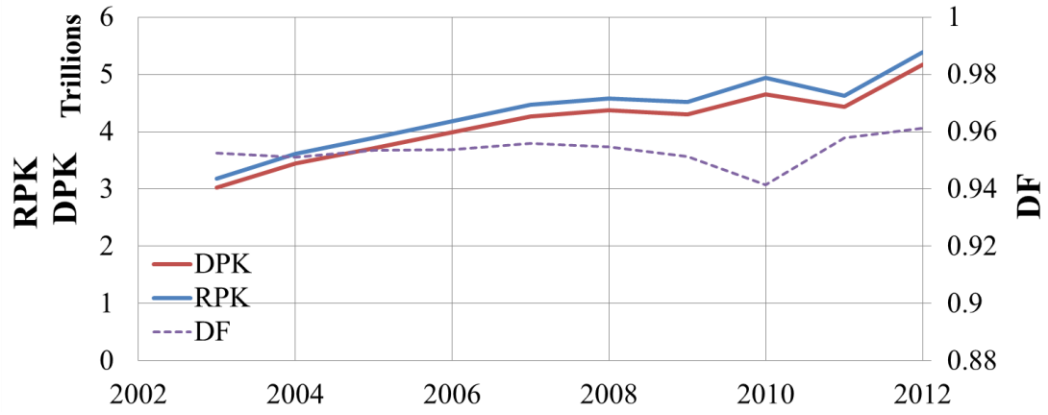


Fig.6.19. Annual historical RPK, DPK and DF

As seen, the DF does not demonstrate large deviations and remains fairly constant, except for the period 2008 – 2011. This is likely to be related to the 2008 economic crisis, when hub and spoke models for air companies were more attractive in order to enhance revenues during this time. However, the DF for 10 years is 0.953 on average. Thus, using Eq. 6.1 and applying the average DF, the DPK can be transferred to the RPK. Using this approach, the forecasted APD for the four scenarios and the calculated DPK were transferred to the RPK. Then they were compared to the RPK forecasts from Airbus (2015), Boeing (2015) and ICAO FESG (ICAO, 2013). The results are presented in Fig.6.20.

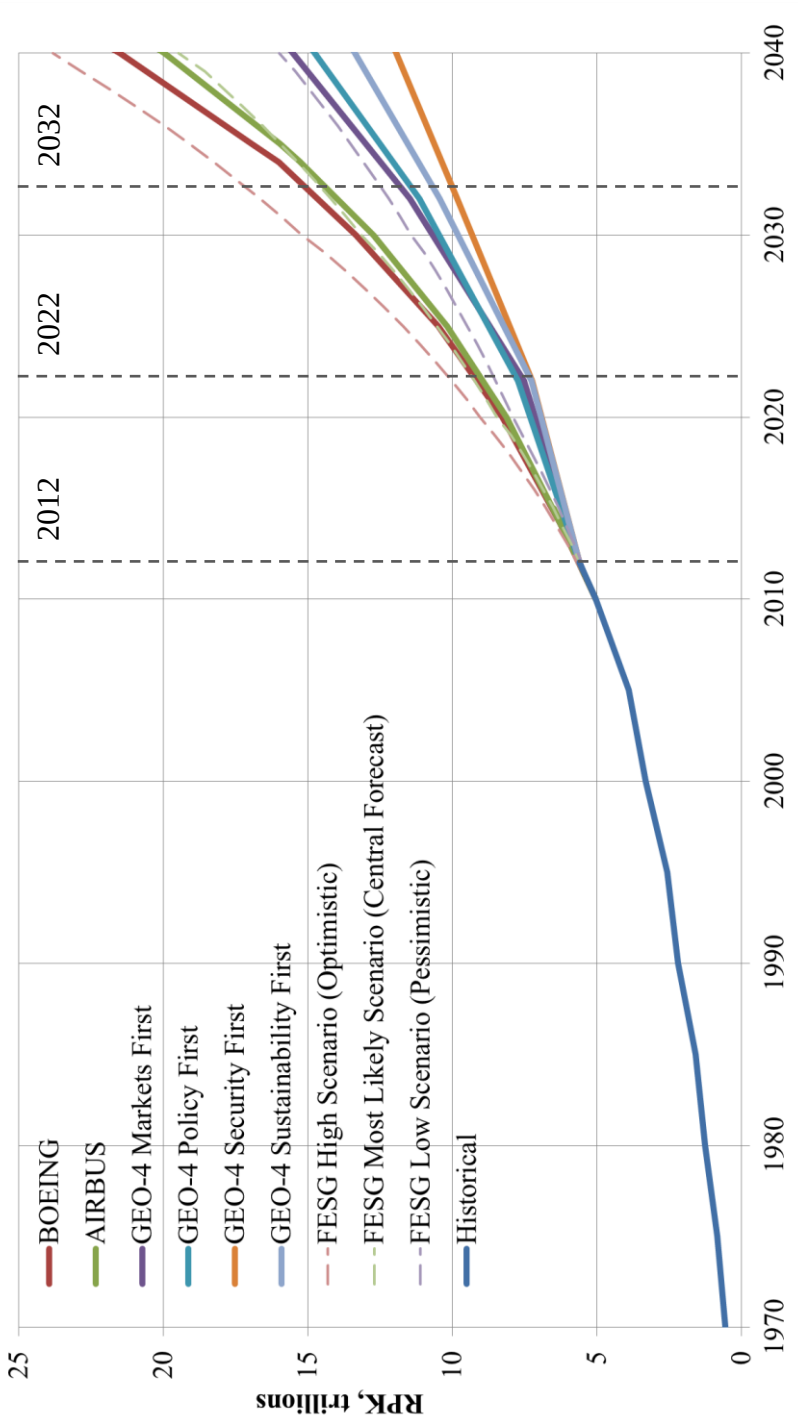
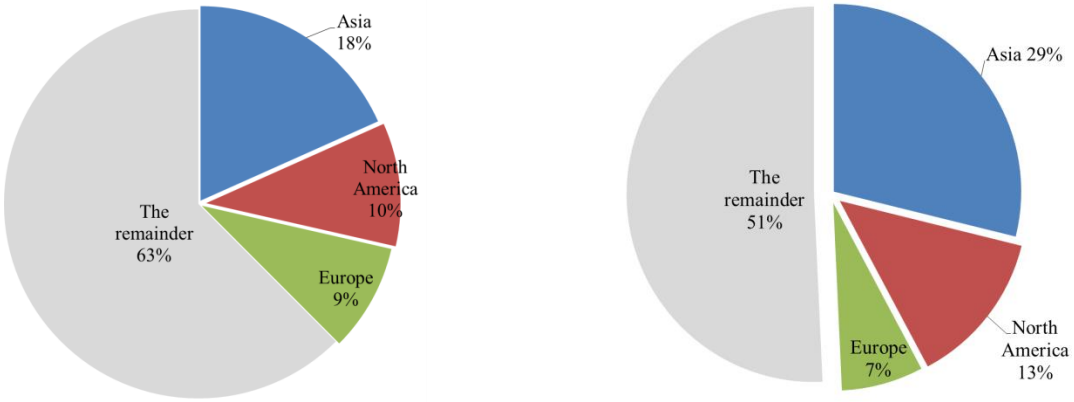


Fig.6.20. Comparison of the modeled RPK with Airbus, Boeing and FSEG RPK forecasts

The Airbus and Boeing RPK forecasts are more optimistic than the results of this study, i.e. the APD model demonstrates lower results. This relates to the fact that the existing forecasts mainly use historical data and its extrapolation into the future referring to historical socio-economic data and do not consider possible deviations. Since the ICAO FESG RPK forecast is a consensus-based forecast which is made based on inputs provided by ICAO, IATA, Airbus, Boeing, General Electric and other organizations, it follows the same logic as those of Airbus and Boeing. However, a comparison of results allows the proposed model to be verified in comparison to the existing studies. It can then be concluded that the obtained APD model results are credible and potentially useful and that the model may be placed with the existing studies. Moreover, the results can be compared at regional level, since the Airbus, Boeing and ICAO FESG forecasts are made at regional level and the APD model results could easily be grouped from settlement level at country, regional and then at continent level. However, only the Boeing RPK forecast contains information for identical regions to the regions in the APD forecasting model. In addition, the Boeing forecast is made for 2034, while the APD model contains RPK information for 2032. Nevertheless, the comparison allows modeled results to be assessed in contrast to the existing Boeing forecast. From the APD forecasting model results, the Sustainability First scenario was chosen because it is the middle scenario among those considered in GEO-4. Although all regions show growth in RPK, the internal contribution of RPK in each region or between region pairs to the global RPK number is relatively small. Three world regions, however, demonstrate the largest internally generated RPK. Thus, results for within the main regions, Europe, North America and Asia, were compared and the RPK shares for the Boeing RPK forecast and the APD forecasting model results for the Sustainability First scenario are presented in Fig.6.21A and Fig.6.21B respectively.

In the Boeing RPK scenario, European, North American and Asian traffic generates about 37% of the total RPK in 2034 collectively. The APD forecasting model shows a result of 49% in 2032 for the same regions. The common share of Europe, North America and Asia traffic in the base year 2012 (Fig.6.22) for the same regions is about 72% (Asia – 30%, Europe – 21%, North America – 21%). As seen, the share decreases in both scenarios when compared to 2012. The Boeing scenario demonstrates a larger reduction when compared to the

Sustainability First. At the same time, the total GDP share of these regions is still the same i.e. about 82%. However, the Asian share increases in both scenarios from 35% in 2012 to 46% in 2034 in the Boeing scenario and to 48% in 2032 in Sustainability First. The European GDP share decreases to 18% from 24% in the Boeing scenario and to 15% in Sustainability First. For North America, the share decreases from 23% to 20% for the Boeing and 16% for the Sustainability First scenario. Thus, the Asian region shows rapid share growth in GDP and, thus, in the total RPK share. The Asian population share in the Boeing scenario is not indicated, but it can be assumed that its share is lower than in the Sustainability scenario.



A. Boeing RPK forecast for 2034

B. APD model RPK forecast for the Sustainability First scenario for 2032

Fig.6.21. Forecasted RPK shares

North America demonstrates a share reduction in RPK. Thus, from a share of 21% in 2012 (Fig.6.22), the total share in 2034 from the Boeing scenario is 10% and 13% from the Sustainability First. Europe shows the same dynamic: the 20% share in 2012 (Fig.6.22) reduces to 9% in the Boeing scenario and 7% in the Sustainability First scenario. This analysis shows that both scenarios expect the importance of the Asian region to increase. However, while the Boeing forecast shows restrained expectations, the APD forecasting model based on the Sustainability First scenario demonstrates a higher RPK than expected in the Asian region. This is certainly related to the rapid Asian socio-economic development described in the Sustainability First scenario. However, the largest contribution to the Asian total share is made by China. In 2012, China's share of the Asian RPK is 40% or 12% of the total RPK. In 2032, China's share is approximately 60% for the Sustainability First and 58% for the Boeing scenario (Fig.6.23).

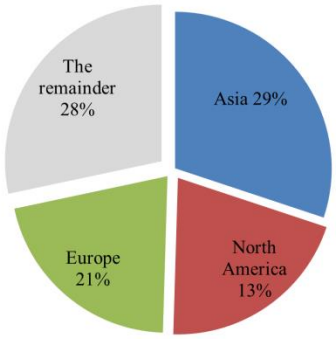
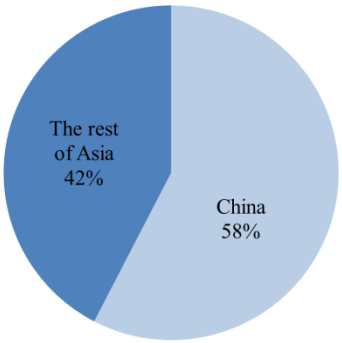
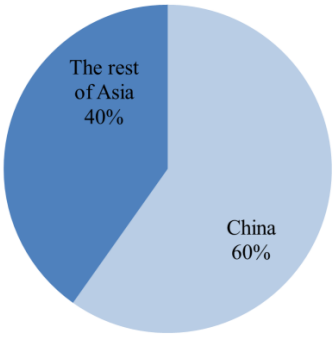


Fig.6.22. RPK shares in 2012



A. Boeing RPK forecasted RPK shares for 2034 for Asia and China



B. APD model forecasted RPK for the Sustainability First scenario for 2032 for Asia and China

Fig.6.23. China's RPK share of the total internal RPK in Asia for Boeing and Sustainability First scenarios

It can therefore be presumed that both of the aforementioned scenarios expect high RPK growth in China, when internally-generated RPK are considered. A comparison of the results obtained from the APD forecasting model with the Boeing scenario showed similar correlations and tendencies for the most part, although some deviations exist. These deviations were mainly caused by the different visions of the future GDP and population development. Nevertheless, it can be concluded that the results of the APD model can be used in order to assess future APD for the various expected scenarios at regional and country level. However, the APD model allows future APD to be analyzed in an even more precise and detailed manner, including, inter alia, settlement pairs. Therefore, APD flows in China were analyzed since the country demonstrates the largest APD growth. The modeled results are nevertheless available for more than 170 countries.

6.5. APD analysis at settlement level for China

In order to demonstrate the APD model abilities of forecasting the APD at settlement level, 154 settlements were considered for China. In 2012, approximately 687 million passengers traveled between these settlements where internal passenger volume was about 619 and external was around 67 million. The APD model based on the Sustainability First scenario for 2042 for the same settlements demonstrated that the total expected APD for China was 5.6 billion, where the internal APD was about 5.3 billion and the external was about 329 million. Thus, in China, domestic APD dominates over the international APD. The international share for China in 2012 was about 11% of the total APD and will decrease to 6% in 2042. These shares show that the domestic APD will play an even more important role in the future. In Fig. 6.24 and 6.25, the APD topology between settlements in China for the base year 2012 and forecasted 2042 based on the Sustainability First scenario for connections containing more than 1 million passengers using the network visualization and exploration tool, Gephi (Bastian et al., 2009) are shown respectively. Lines indicate the APD connection between settlements. The thicker the line, the more passengers it contains in the given year. Circles indicate the settlements which are placed according their geographical coordinates. The darker the circle, the more passengers this settlement has in the given year. As seen, the APD develops rapidly: the number of connections with more than 1 million passengers increases from 582 in 2012 to 1,220 in 2042. China's internal expected APD in 2042 is 2,637,881,794. Fig.6.25 demonstrates that almost all of the considered Chinese settlements will generate more than 1 million passengers in 2042. Nanchong shows the most rapid APD growth. In domestic China, this city generated 246,694 passengers with 13 APD connections in 2012. In contrast, this city will demonstrate the expected APD of 2,482,345 and 129 ADP connections in 2042. Such quick growth is certainly related to the rapid development of the country as a whole and, probably, to its central geographical location. On average, APD on every connection inside China in 2042 will be eight times higher than in 2012. The average distance of the APD connection in 2042 will be 1,277 km. The average passenger number on connections in China in 2042 will be 1,195,775. In 2012, these figures were 1,357 km and 100,772 passengers.

Fig.6.24. APD in China in the base year 2012 with connections of more than 1 million passengers

Fig.6.25. Forecasted APD in China for 2042 based on the Sustainability First scenario with connections of more than 1 million passengers

International China APD showed the same rapid growth as for internal China. Fig.6.27 and 6.28 demonstrate the changes in topology and APD growth between 2012 and forecasted 2042 for connections of more than 100,000 passengers. The APD connection figure increases significantly. The connection number with more than 100,000 passengers in 2012 and in forecasted 2042 are presented in Fig.6.26.

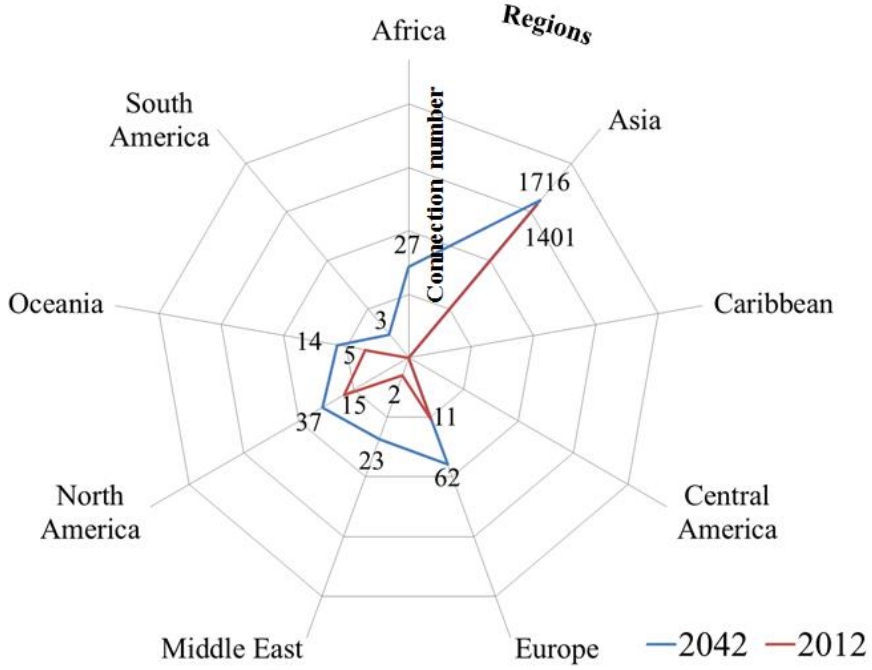


Fig.6.26. The APD connection number in 2012 and 2042 for connections containing more than 100,000 passengers between China and world regions

The Asian region shows the largest growth of connections with more than 100,000 passengers. For instance, in 2012 there were 1,401 connections between China's settlements and the settlements in other Asian countries. This figure will more than double by 2042, reaching 3,234. However, settlements connecting China to Europe, Middle East and Africa demonstrated rapid growth: from 11 APD connections in 2012 to 62, from 2 to 23 in 2042 and from 0 to 27 respectively. In contrast, the Oceania and South America connections demonstrated minimal growth. Nanchong showed the fastest growth. In 2012, this city had 40 APD connections with 542 passengers and by 2042, the expected APD connection number is 1,697 with 2,674,900 passengers. Changzhi, Tianjin, Chongqing, Shenzhen, among others, displayed the same rapid growth rates. In addition, despite some settlements losing their APD connections, passenger numbers still continued to increase significantly.

Fig.6.27. APD in China in the base year 2012 with connections of more than 100,000 passengers

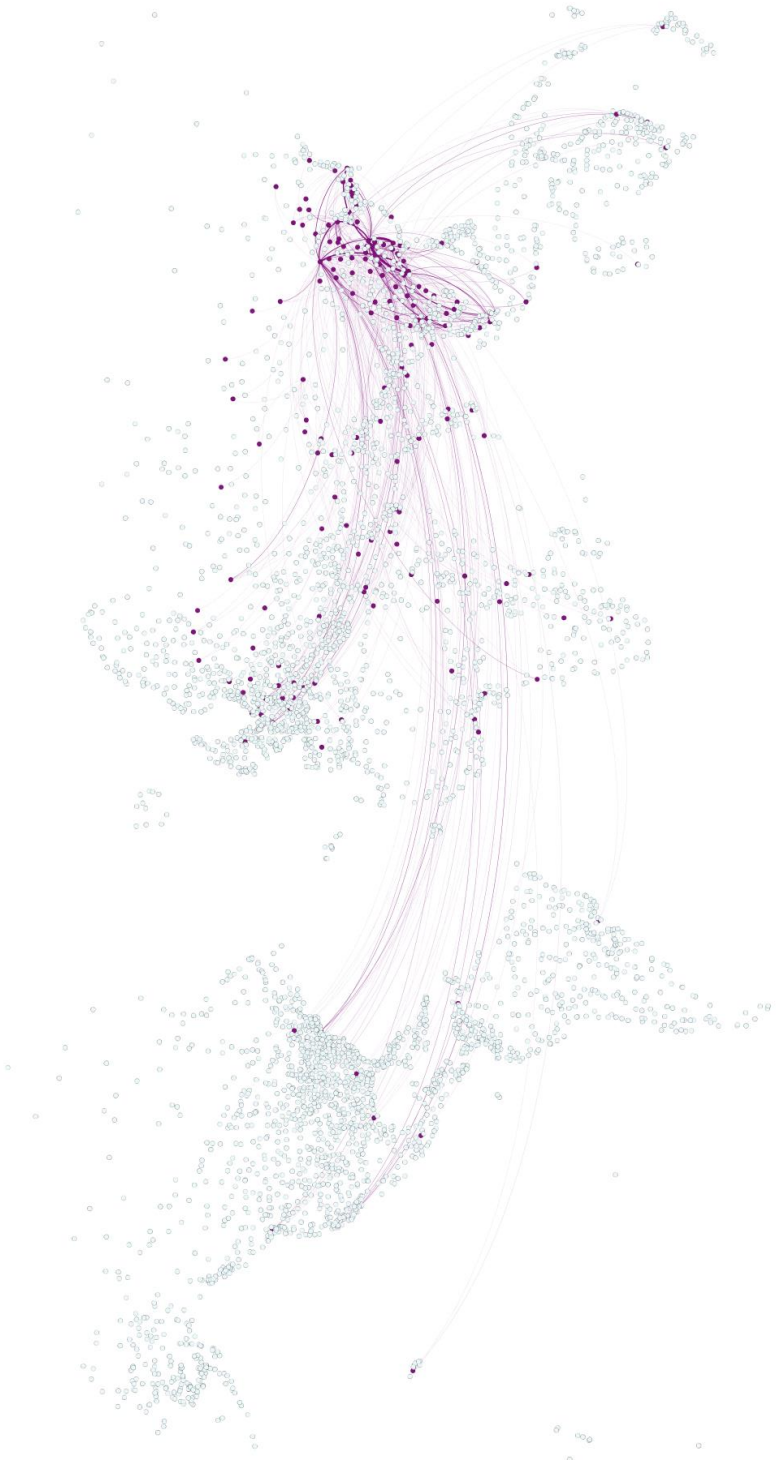


Fig.6.28. Forecasted APD in China for 2042 based on the Sustainability First scenario with connections of more than 100,000 passengers

For example, in 2012, Xiamen was connected to 758 settlements outside China with 1,741,894 passengers. In 2042, this city will have 78 APD connections with 7,859,259 passengers. Such a rapid reduction could be related to Xiamen's socio-economic growth and the connected settlements. It is most likely that the growth dynamic is different and the future expected APD is concentrated on other development settlements outside China.

Tab.6.4 provides a comparison between 2012 and forecasted 2042 for China in terms of passenger number, APD connection number, APD connection number with more than 100,000 passengers on them and average distances between regions.

Country	**Region**	**Passenger number**		**APD connection number**		**Connections with >100k passengers**		**Average distance between regions, km**	
		2012	2042	2012	2042	2012	2042	2012	2042
China	Africa	1,256,100	13,279,233	1,896	3,261	0	27	10,426	10,220
China	Asia	48,024,083	3,070,078,025	6,051	19,870	1,401	1,716	3,000	1,014
China	Caribbean	16,156	173,551	176	349	0	0	14,361	14,112
China	Central America	22,532	205,941	158	232	0	0	14,568	14,205
China	Europe	8,403,429	34,170,425	9,024	6,195	11	62	7,956	7,617
China	Middle East	1,633,455	10,815,981	1,495	2,102	2	23	6,613	6,533
China	North America	5,705,741	27,031,982	7,925	7,405	15	37	11,344	11,381
China	Oceania	2,002,655	9,045,920	1,822	1,350	5	14	8,340	8,219
China	South America	364,189	3,120,842	1,178	2,570	0	3	17,331	17,044

Tab.6.4. APD indicator comparison between 2012 and 2042 for China

The expected average distance between China's settlements and settlements in other regions hardly changed except for the Chinese connection to Asia. As seen, in 2042 the average distance is 3 times less than in 2012. This reduction is related to socio-economic settlement growth. The Asian region showed strong growth, thus the settlement populations and GDPs develop rapidly. Thus, there are more "stronger" settlements in 2042 compared to 2012 and, therefore, the settlement density in the same area is higher in 2042 which leads to a smaller average distance between China's settlements and settlements in Asian regions.

When collaborative internal and external expected APD for China is considered, the number of settlements demonstrated rapid growth. In Tab.6.5, the top 8 settlements with their corresponding APD connection growth figures for 2042 are presented below.

	2012		2042	
Settlement	**APD**	**APD connections**	**APD**	**APD connections**
Nanchong	247,236	53	5,157,245	1,826
Changzhi	631,997	101	7,071,087	1,071
Tianjin	8,690,805	573	72,453,364	1,508
Chongqing	23,118,706	693	195,581,005	1,549
Shenzhen	32,185,794	709	270,060,018	1,555
Lanzhou	4,840,602	345	42,362,845	1,016
Shanghai	75,644,194	1,618	560,843,000	2,264
Wuhan	14,072,335	768	118,519,357	1,413

Tab.6.5. Top 8 settlements with APD connection growth number in 2042

As seen, settlements with a relatively small ADP in 2012, such as Nanchong and Changzhi, are expected to have a significant passenger number desiring to travel to the various destinations as in China as well as worldwide in 2042. At the same time, large settlements in 2012 will only increase their APD in 2042. For example, Shanghai will have 2,264 connections with worldwide settlements in 2042. This is more than half of the considered settlements in the ADP forecasting model. This city, as well as China and the whole Asian region, is one of the world leaders in terms of generated APD and APD connection number in 2042. Thus, the provided forecast could be used for the future ATS planning and related activities at different aggregation levels such as settlement, country, region, and world. This is particularly true for the manufacturers – the provided results could give some insights to the general requirements of future aircraft. Air companies could also assess and plan their future route networks and the corresponding aircraft types required. Airports could make estimations of future development using the APD forecasted data. Thus, the generated results could be used to enhance the performance of the air transport system, providing insights to the state of the system to decision makers.

6.6. Conclusion

In this chapter, the proposed APD model was applied to the four GEO-4 socio-economic scenarios. The forecast was made for every ten years from the base year 2012 to 2042. The GEO-4 scenarios were disaggregated from country level to settlement level using collected settlement populations and GDPs and the GDP, population and country urbanization rates from the GEO-4 scenarios for the base year 2012. Since the airfares and their importance for the APD modeling are not available in the scenarios, a simple airfare model was developed. The model provides a correlation between the average annual airfare and the distance between settlements, and the annual average crude oil price. This correlation was retrieved by analyzing the historical annual data from 2002 to 2012. The obtained simple airfare model was then applied for the APD forecasting.

As input for the model, the four socio-economic scenarios from GEO-4 were used: Markets First, Policy First, Security First and Sustainability First. The scenarios, united by a common philosophy from one source, are built based on the current socio-economic trends along divergent development paths in the future. Using the APD forecasting model, the APD results for all scenarios were obtained. The results are available at city level which could be presented at country, regional or world level of aggregation. The obtained APD values for the four scenarios relative to each other matched the scenario descriptions, and also showed the APD forecasting model sensitivity. Then, the obtained results were compared to the existing studies: Boeing forecast, Airbus forecast and ICAO FESG forecast. Since these forecasts are made in RPK and the APD forecasting model results are in DPK, historical data were analyzed and the ratio, so-called Directness Factor (DF), between DPK and RPK for every year from 2002 to 2012 was obtained. DP is approximately constant through these years, with some deviations most likely relating to the financial crisis of 2008. Thus, the average DP is 0.953. Using this coefficient, the total DPKs were transferred to RPKs and then compared with other forecasts.

It can be concluded that the APD model results are comparable to other forecasts such as Boeing, Airbus and ICAO FESG forecasts and the model shows adequate results. However, the APD forecasting model results for GEO-4 scenarios show more pessimistic results

compared to other forecasts. One of the main reasons for the difference is that the APD model operates with socio-economic scenarios, whereas other studies project socio-economic historic trends into the future.

China was chosen in order to demonstrate the APD model abilities of forecasting the APD at settlement level, since it shows the fastest growth. The expected internal and external APD are analyzed. The analysis showed the considerable importance of some settlements in China which, although they did not have a large APD in 2012, they do become more relevant in 2042. According to the Sustainability First scenario, in 2042, this country will not only dominate in the Asian region, but worldwide. Thus, particular attention should be dedicated to this country in order to correctly estimate the future impact on air transport system stakeholders such as air carriers, airports, manufacturers, authorities. The future APD data between settlements can be used as a basis for further studies, using the detailed settlement level results proposed by the APD model as a reference.

The APD forecasting model application covers the third step of the proposed research methodology indicated in Section 3.2 – *Model Application*. Thus, following the research methodology this chapter, along with Chapters 4 and 5, fulfills the proposed key research objective. The next chapter offers conclusions for the main outputs obtained within this thesis.

7. Conclusion

This study presents an APD forecasting model. This model, based on an external socio-economic scenario, is able to forecast the APD between settlements worldwide. Using the developed framework and external socio-economic scenarios, a two-step modelling approach is applied: the topology forecasting model and the passenger forecasting model. The APD forecasting model is validated on real world data and then applied to the four socio-economic scenarios. Based on the study, the main conclusion is that it is possible to model the APD at settlement level. This is done by collecting available data from various sources and applying the proposed models.

As presented in Section 5.4, the overall model accuracy (35% for the APD connection forecast, covering more than 70% of the actual passenger number) and the error propagation analysis showed that the developed topology and the passenger forecasting models demonstrate sufficient accuracy for modeling. In addition, the analysis revealed that settlement clustering according to their socio-economic indicators improves the overall forecast from 0.68 to 0.78 based on the precision metric shown in Section 5.2.

Based on the modeled results, it could be concluded that the global APD will show speedy growth in all of the GEO-4 scenarios considered in this study. On average, the global APD in 2042 is expected to be between 7 – 9 billion passengers. This is approximately three times larger than in 2012. China demonstrates the highest APD growth in all of the scenarios considered. By analyzing the Sustainability First scenario, which is the middle scenario of the group, China shows expected growth from 687 million passengers in 2012 to more than 5.6

billion passengers in 2042. It can be concluded that China will not only dominate in the Asian region, but worldwide. Thus, particular attention should be dedicated to this country in order to correctly estimate the future impact on ATS. Using the APD model results at settlement level, the future APD data between these settlements could be used as the basis for further studies.

One of the important conclusions in this study is that the connection emergence and disappearance forecast is important and should be taken into account for forecasting APD at settlement level. A good example of adding connections is the city Nanchong, China showed in Section 6.5. By including the processes of adding and eliminating connections, a more realistic understanding of the global mobility can be obtained.

It can be concluded that the obtained results from the proposed APD forecasting are well comparable according to other studies. The APD forecasting model based on the GEO-4 scenarios demonstrated more pessimistic expectations on the future total APD growth in contrast to Boeing, Airbus and ICAO FESG forecasts. This relates to the fact that the existing forecasts mainly use historical data and its extrapolation into the future, taking the historical socio-economic data into account, and do not consider possible deviations. This leads to more optimistic GDP growth than is made in the GEO-4 scenarios.

The results were also compared to the Boeing forecast at regional level since the APD forecasting model and Boeing forecast consider the similar regions. The three main regions with internal APD generation were analyzed: Europe, North America and Asia. Both forecasts showed the significant importance of the Asian region. However, the APD forecasting model demonstrated a larger Asian total ADP share than in the Boeing forecast. The greatest contribution to Asian growth is China. Thus, the expected Chinese APD from the Boeing forecast and the APD forecasting model were compared. The comparison showed that forecasted APD is less than in the Boeing forecast but the Chinese share in total APD in the Asian region was approximately the same. Thus, the comparison analysis confirms that the APD forecasting model generates adequate results based on the considered scenarios. Results are valid and comparable to other studies and they follow the scenario logics.

Therefore, based on the conclusions, the study provides valuable results for the expected future APD. The APD model itself can be used for APD estimation for various socio-economic scenarios, assessing different desecrate impacts or long term effects. For example, element exclusion (a settlement, a country or a region) due to different socio-economic impacts can be assessed on a worldwide scale. The influence on the whole APD network and expected APD for a newly introduced settlement could be assessed. The APD forecasting model is a useful tool to provide an expected global APD for the future at settlement level based on GPD, population and crude oil price scenarios. The APD model usage could give a decision maker valuable outputs for strategy planning: the ability to assess a scenario's impact on the future APD and therefore, assessing possible decision impacts on a scenario in order to achieve planned strategy aims. Thus, in practice, the APD forecasting model at settlement level would be useful for all ATS stakeholders in various ways for example: air companies: for assessing a future route network, aircraft manufacturers: for assessing future aircraft demand, airports: for assessing future capacities, air transport management: for assessing future traffic. In addition, the model results can be used as a basis for assessing future environmental impacts associated with ATS, for example analyzing changes in non-CO_2 emissions volumes over time at global, regional, country or single connection level. Using the APD forecasting model results, decision makers would have an understanding of the expected future and be able to create alternatives. The proposed model in this thesis provides decision makers with an opportunity to see the consequences of their decisions in the future, generating awareness that there is not just one possible future outcome, but several, depending on the current decisions made. Thus, the proposed APD model is a valuable and unique tool among other models and tools. The model makes APD forecasts with the investigated accuracy − something which is not presented in other studies. The forecasting model is a reliable instrument for retrieving the APD values at settlement level which are able to be used for assessing different visions of future world development.

8. Recommendations for future research

This chapter presents recommendations for future research. The APD forecasting model includes a number of sub-models, as shown in Chapter 4. The recommendations are given in order to improve the presented sub-models and, in turn, the APD forecasting model performance.

For clustering settlements to groups it is recommended that additional studies in community investigating are conducted. Since the AIC and BIC metrics do not demonstrate the best values, clustering with additional socio-economic parameters could be considered. Identifying capital settlements and pairing particular regions or countries would almost certainly enhance the clustering metric and, therefore, clustering quality.

There are a few recommendations for the APD topology forecasting model. In this study, weighted similarity algorithms are used to predict APD connections between settlements. However, a number of link prediction methods exist in network science. Notwithstanding the large expected calculation time costs, these methods could provide higher accuracy, although such accuracy benefits could be insignificant.

The APD network in this study is considered as a weighted network, where the weights are defined as the attractiveness between settlements (see Eq.5.1). The main assumption for the weight is that the attractiveness between settlements could be presented as variation of Newton's based on GDPs, populations, average airfare and distance. This equation with the same coefficients is applied to every cluster pair. It is recommended that further studies on weight equation are carried out. It is believed that every cluster pair could have its own coefficients in the weight equation and the attractiveness for different settlement types is different. Thus, this additional study on weights is likely to improve the topology forecast

model accuracy. In addition, the main network metrics (e.g. average weighted degree, average path length, modularity, etc.) could be analyzed and compared with the metrics obtained from historical data described by Ghosh and Terekhov (2015). This may help to understand latent processes for the APD connections generation.

The passenger forecasting model (Section 5.3) is aimed at defining the passenger number on newly appeared (Sub-section 5.3.1) and remaining (Sub-section 5.3.2) APD connections. This thesis for the newly appeared APD connections considers the QA approach using GDPs, populations, distance and average airfare. It is likely that some variables have a larger impact on the similarity between connections. An additional study could be conducted to define the appropriate variable combination for every cluster pair to forecast the passenger number on these connections with higher accuracy. For remaining connections, passenger growth is defined as the average settlement GDP growth. A further study could be performed for this ADP connection type. Improvements to accuracy could be achieved by taking airfares, distances between settlements and populations into account. In addition, gravity models could be considered for application on the remaining APD connections.

For the average airfare model (Section 6.1) additional analysis could be performed and this model could be adopted for every cluster pair. It is likely that average airfare varies for different cluster pairs.

In addition, integrating the APD forecasting model into the modular environment in AIRCAST would allow the model to be calibrated (Chapter 1). The back loops from other layers (Fig.1.3) are able to provide additional inputs to the APD forecasting model, such as travel time, aircraft frequency on a route and so on. The iterative approach is likely to further improve the accuracy of the APD forecast.

Thus, applying the aforementioned recommendations could enhance the overall accuracy of the model and it would be able to generate more accurate APD forecasts at settlement level worldwide.

9. References

Airbus, Airbus Global Market Forecast 2014-2033, Blagnac Cedex, France, 2014

Akaike, H., 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716-723

Alam, M.J.B., Karim, D.M., 1998. Air travel demand model for domestic air transportation in Bangladesh, Journal of Civil Engineering, Institution of Engineers, Bangladesh, ISSN 0379-4318, Vol. 26, No.1. pp. 1-13

Alderighi, M., Cento, A., Nijkamp, P., & Rietveld, P., 2007. Assessment of New Hub-and-Spoke and Point-to-Point Airline Network Configurations. Transport Reviews, 27(5), 529-549

Anker, R., 2000. Comparison of Airbus, Boeing, Rolls-Royce and AVITAS Market Forecasts. Air & Space Europe, 2(3), 4-9

Apffelstaedt, A., Langhans, S., & Gollnick, V., 2008. Identifying carbon dioxide reducing aircraft technologies and estimating their impact on global CO2 emissions, Deutscher Luft- und Raumfahrt Kongress (DLRK), Aachen, Germany

Armstrong, J. S., Green, K. C., 2005. Demand forecasting: Evidence-based methods (No. 24/05). Monash University, Department of Econometrics and Business Statistics

Bania, N., Bauer, P. W., & Zlatoper, T. J., 1998. US air passenger service: A taxonomy of route networks, hub locations, and competition. Transportation Research Part E: Logistics and Transportation Review, 34(1), 53-74

Barabasi A.-L., Albert R., 1999. Emergence of scaling in random networks, Science 286, 509

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A., 2004. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 101(11), 3747-3752

Bastian, M., Heymann, S., & Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. ICWSM, 8, 361-362

Belobaba, P., & Odoni, A. R.,Barnhart, C. (Eds.), 2009. The Global Airline Industry. Wiley

Berkhin, P., 2006. A survey of clustering data mining techniques. Grouping multidimensional data, Springer Berlin Heidelberg, 25-71

Boeing, Current Market Outlook 2013-2032, USA, 2013

Boeing, Current Market Outlook 2015-2034, USA, 2015

Brown, S. L., & Watkins, W. S., 1968. The demand for air travel: a regression study of time-series and cross-sectional data in the US domestic market. Highway Research Record, (213)

Delta Air Lines, https://www.deltamuseum.org/exhibits/delta-history/first-in-the-air, [cited 31.03.2016]

Doganis, R., 2002. Flying off course: The economics of international airlines. Psychology Press

Doucet R., Margaretic, P., Thomas-Agnan, C., Villota, Q., 2014. Spatial dependence in (origin-destination) air passenger flows, ITEA Annual Conference and Summer School on Transportation Economics (Kuhmo Nectar), Toulouse, France

Dray L., Evans A.D., Reynolds T., Schäfer A., 2010. Mitigation of Aviation Emissions of Carbon Dioxide: Analysis for Europe, Transportation Research Record, 2177, pp. 17-26

Dray L., Evans A.D., Reynolds R., Schäfer A., Vera-Morales M., Bosbach W., 2014. Airline Fleet Replacement Funded by a Carbon Tax: an Integrated Assessment, Transport Policy, Special Issue on Aviation and the Environment, 34, pp. 75-84

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R., 2012. Sensitivity and specificity of information criteria. The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University

Embraer, 2015. Market Outlook 2015-2034

European Commission, Flightpath 2050, Luxembourg, 2011

Flightglobal, ASCEND, cited [08.04.2015]

Fotheringham, A. S., 1983. Some theoretical aspects of destination choice and their relevance to production-constrained gravity models. Environment and Planning A, 15(8), 1121-1132.

Ghosh, R., Terekhov, I., 2015. Future Passenger Air Traffic Modelling: Trend Analysis of the Global Passenger Air Travel Demand Network, 53rd AIAA Aerospace Science Meeting, Kissimmee, Florida

Ghosh, R., Schilling, T., & Wicke, K., 2014. Theoretical framework of systems design for the air transportation system including an inherently quantitative philosophy of scenario development. In 29th Congress of the International Council of the Aeronautical Sciences (ICAS) (Vol. 7, p. 12)

Gössling, S., & Upham, P., 2009. Climate change and aviation: Issues, challenges and solutions. Earthscan

Grewe, V., & Stenke, A., 2008. AirClim: an efficient climate impact assessment tool. Atmospheric Chemistry and Physics, 8, 4621-4639

Grosche, T., Rothlauf, F., Heinzl, A., 2007. Gravity models for airline passenger volume estimation, Journal of Air Transport Management, Vol. 13, pp. 175-183

Guimera, R., Mossa, S., Turtschi, A., & Amaral, L. A., 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

Han, D. D., Qian, J. H., & Liu, J. G., 2007. Network topology of the Austrian airline flights. arXiv preprint physics/0703193

Hanely J.A., McNeil B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology, 143: 29-39

Herlocker J.L., Konstann J.A., Terveen L.G., Riedl J.T., 2004. "Evaluating Collaborative Filtering Recommender Systems", ACM Transactions on Information Systems, Vol. 22, No. 1, pp. 5–53

Hruschka, R.E., Freitas, A.A., de Carvalho A.L., 2009. "A Survey of Evolutionary Algorithms for Clustering", IEEE Transactions on Systems, Man, and Cybernetics—part C: Applications and Reviews, vol. 39, no. 2

Hummels, D., 2007. Transportation costs and international trade in the second era of globalization. The Journal of Economic Perspectives, 21(3), 131-154

IATA, IATA Technology roadmap 2013

International Civil Aviation Organization (ICAO), 2013. FESG CAEP/9 traffic and fleet forecasts- methodological paper. CAEP/9-IP/11

Ishutkina, M. A., & Hansman, R. J., 2008. Analysis of interaction between air transportation and economic activity. Retrieved November, 17, 2014

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. "Data clustering: a review", ACM computing surveys (CSUR), 31(3), 264-323

Japan Aircraft Development Corporation, 2010. Worldwide Market Forecast for Commercial Air Transport 2010-2029, Chiyodaku, Tokyo, Japan

Jorge-Calderon, J. D., 1997. A demand model for scheduled airline services on international European routes. Journal of Air Transport Management, 3(1), 23-35

Keeling, D. J. ,1995. Transport and the world city paradigm. World cities in a world-system, 115-131

Krishnan, T. and McLachlan, G., 1997. The EM algorithm and extensions. Wiley, 1(997), 58-60

Koch, A., Lührs, B., Dahlmann, K., Linke, F., Grewe, V., Litz, M., Plohr, B.,Nagel, B., Gollnick, V., Schumann, U. 2011. Climate impact assessment of varying cruise flight altitudes applying the CATS simulation approach. In Third International Conference of the European Aerospace Societies, Venice, Italy

Kollmuss, A., Crimmins, M.A., 2009. Carbon Offsetting & Air Travel, Part 2: Non-CO2 Emissions Calculations, Stockholm Environment Institute (SEI) discussion paper, 2009.

Lakshmanan, T. R., 2011. The broader economic consequences of transport infrastructure investments. Journal of transport geography, 19(1), 1-12

Lee, D. S., Fahey, D. W., Forster, P. M., Newton, P. J., Wit, R. C., Lim, L. L., Owen, B. & Sausen, R., 2009. Aviation and global climate change in the 21st century. Atmospheric Environment, 43(22), 3520-3537

Leicht E.A., Holme P., Newman M.E.J., 2006. Vertex similarity in networks, Phys. Rev. E 73, 026120

Li, W., & Cai, X., 2004. Statistical analysis of airport network of China. Physical Review E, 69(4), 046106

Lillo, F., Micciche, S., Mantegna, R. N., Beato, V., & Pozzi, S., 2011. Elsa project: Toward a complex network approach to atm delays analysis. Proceedings of the SESAR Innovation Days EUROCONTROL. ISBN, 978-2

Lü L., Zhou T., 2012. Link prediction in weighted networks: The role of weak ties, EPL, A letters Journal Exploring the Frontiers of Physics, 89 18001, January 2010.

Lü, L., Zhou, T., 2011. Link prediction in complex networks: A survey, Physica A, Vol. 390, pp. 1150-1170

MaxMind, Free World Cities Database [cited 19.11.2014]

Morrison, S.A., Winston, C., 1995. The Evolution of the Airline Industry. Brookings Institution, Washington, DC

Murata T., Moriyasu S., 2007. Link prediction of social networks based on weighted proximity measures, IEEE/WIC/ACM International conference on Web Intelligence, Fremont, California

Newman, M. E., 2003. The structure and function of complex networks. SIAM review, 45(2), 167-256

Newton, I., Bernoulli, D., MacLaurin, C., & Euler, L., 1833. Philosophiae naturalis principia mathematica (Vol. 1). excudit G. Brookman; impensis TT et J. Tegg, Londini

Nolte, P., Apffelstaedt, A., & Gollnick, V., 2012. Quantitative Assessment of Technology Impact on Aviation Fuel Efficiency. Air Transport and Operations, 514

O'Kelly, M. E., Song, W., & Shen, G., 1995. New estimates of gravitational attraction by linear programming. Geographical Analysis, 27(4), 271-285

OpenFlights, Airport database [cited 19.11.2014].

Our Airports [cited 19.11.2014]

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., ... & Dubash, N. K., 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Ravasz E., Somera A.L., Mongru D.A., Oltvai Z.N., Barabási A.-L., 2002. Hierarchical organization of modularity in metabolic networks, Science 297, 1551

Rengaraju, V. R., & Arasan, V. T., 1992. Modeling for air travel demand. Journal of Transportation Engineering, 118(3), 371-380

Reynolds, T. G., Barrett, S., Dray, L., Evans, A., Kohler, M. O., Morales, M. V., ... & Hunsley, R., 2007. Modelling environmental & economic impacts of aviation: introducing the aviation integrated modelling project. In 7th AIAA aviation Technology, integration and operations Conference.

Rolls-Royce, 2012. Market outlook 2012-31

Russon, M. G., & Riley, N. F., 1993. Airport substitution in a short haul model of air transportation. International Journal of Transport Economics/Rivista internazionale di economia dei trasporti, 157-174

Sabre Airline Solutions, Aviation Data Intelligence [cited 19.11.2014]

Salton G., McGill M.J., 1983. Introduction to Modern Information Retrieval, McGraw-Hill, Auckland

SAS Institute Inc., 2014. JMP® 11 Multivariate Methods, Cary, NC: SAS Institute Inc.

Schwarz, G., 1978. "Estimating the dimension of a model". The annals of statistics, 6(2), 461-464

Shen, G., 2004. Reverse-fitting the gravity model to inter-city airline passenger flows by an algebraic simplification. Journal of Transport Geography, 12(3), 219-234

Sorensen T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 1

Suryani, E., Chou, S.-Y., Chen C.-H., 2010. Air passenger demand forecasting and passenger terminal capacity expansion: A system dynamics framework, Expert Systems with Applications, Vol. 37, pp. 2324-2339

Svensson, F., 2005. Potential of reducing the environmental impact of civil subsonic aviation by using liquid hydrogen. Cranfield University

Terekhov, I., Gollnick, V., 2015. Clustering of airport cities and cluster dynamics for the air passenger demand forecasting model based on a socio-economic scenario, CEAS conference, Delft, Netherlands

Terekhov, I., Ghosh, R., Gollnick, V., 2015. A concept of forecasting origin-destination air passenger demand between global city pairs using future socio-economic development scenarios, 53rd AIAA Aerospace Sciences Meeting, Kissimmee, Florida, USA

Terekhov, I., Evans A.D., Gollnick V., 2015. Forecasting global air passenger demand network using weighted similarity-based algorithms, ATRS conference, Singapore, 2015

Terekhov, I., Evans, A., & Gollnick, V., 2016. Forecasting a Global Air Passenger Demand Network Using Weighted Similarity-Based Algorithms. In Complex Networks VII . Springer International Publishing, pp. 335-347

Thacker, B. H., Doebling, S. W., Hemez, F. M., Anderson, M. C., Pepin, J. E., & Rodriguez, E. A., 2004. Concepts of model verification and validation (No. LA-14167). Los Alamos National Lab., Los Alamos, NM (US)

The Frederick S. Pardee Center for International Futures, 2014. http://pardee.du.edu/ cited [19.11.2014]

The World Bank, World Bank Open Data [cited 19.11.2014]

U.S. Energy Information Administration. Petroleum & Other Liquids. Spot Prices, http://www.eia.gov/dnav/pet/pet_pri_spt_s1_a.htm [cited 20.08.2014]

UK Department for Transport, 2013. UK aviation forecasts 2013, https://www.gov.uk/ government/uploads/system/uploads/attachment_data/file/223839/aviation-forecasts.pdf, [cited 19.11.2014]

UN, 2014-1. World population Prospects: The 2012 Revision, [cited 19.11.2014]

UN, 2014-2. National Accounts Main Aggregates Database, [cited 19.11.2014]

United Nations Environment Programme, 2007. Global environment outlook 4: environment for development. United Nations Environment Program

Verleger Jr, P. K., 1972. Models of the demand for air transportation. The Bell Journal of Economics and Management Science, 437-457

Wu, Z., Braunstein, L. A., Colizza, V., Cohen, R., Havlin, S., & Stanley, H. E., 2006. Optimal paths in complex networks with correlated weights: The worldwide airport network. Physical Review E, 74(5), 056104

Xu, Z., & Harriss, R., 2008. Exploring the structure of the US intercity passenger air transportation network: a weighted complex network approach. GeoJournal, 73(2), 87-102

Xu, R., Wunsch, D, 2005. Survey of clustering algorithms, Neural Networks, IEEE Transactions on, 16(3), 645-678

Zanin, M., & Lillo, F., 2013. Modelling the air transport with complex networks: A short review. The European Physical Journal Special Topics, 215(1), 5-21

Zheleva, E., Golbeck, J., Kuter, U., 2012. “Using Friendship Ties and Family Circles for Link Prediction”, Advances in Social Network Mining and Analysis Lecture Notes in Computer Science, Vol. 5498, pp. 97-113

Appendix A

Detailed Model Validation Results

This appendix lists the detailed model validation results in the form of tables.

List of Tables

A.14. Passenger numbers covered by connections from Tab. A.13 for 2012 from 2011 for every cluster pair at specified percentage intervals

	2002 con	2002 pas	2007 con	2007 pas	2011 con	2011 pas	2012 con	2012 pas
Very small-rich – very small-rich	14,678	6,335,841	12,928	6,870,237	12,205	7,114,701	12,220	6,777,836
Small-poor – very small-rich	5,267	1,389,203	4,592	1,753,221	4,776	2,455,500	4,915	2,192,576
Large-rich – very small-rich	19,066	30,009,742	20,390	33,900,853	19,785	39,971,542	19,738	39,722,874
Middle-middle – very small-rich	4,504	1,201,722	4,838	1,576,175	5,387	2,618,091	5,640	2,085,459
Megacities – very small-rich	13,240	23,505,449	14,498	30,385,202	14,426	37,783,433	14,789	35,711,812
Small-rich – very small-rich	31,334	10,851,568	29,795	11,986,712	28,063	19,220,313	27,889	12,225,544
Large-poor – very small-rich	8,576	1,934,284	8,903	2,643,095	9,892	5,465,645	9,943	2,988,132
Middle-rich – very small-rich	24,336	22,069,514	24,388	26,460,053	23,272	29,852,250	23,036	27,994,851
Middle-poor – very small-rich	2,694	519,740	2,891	658,331	3,473	1,360,269	3,552	825,981
Small-poor – small-poor	2,442	3,809,421	2,648	4,198,369	2,803	5,420,493	3,230	4,548,417
Large-rich – small-poor	8,617	7,711,935	10,608	13,705,048	11,390	16,318,953	11,866	16,376,026
Middle-middle – small-poor	3,930	3,963,788	4,629	4,993,110	5,907	6,290,819	6,729	6,218,860
Megacities – small-poor	7,486	25,401,243	9,745	51,771,712	10,699	71,389,074	11,430	76,591,762
Small-rich – small-poor	9,291	1,403,181	9,841	2,401,931	10,019	5,359,739	10,268	2,459,236
Large-poor – small-poor	6,535	13,075,701	8,613	23,354,230	10,350	40,613,450	11,167	43,554,623
Middle-rich – small-poor	9,952	4,103,144	11,128	8,523,521	11,693	11,032,940	11,882	10,983,548
Middle-poor – small-poor	2,285	2,897,634	3,057	3,964,237	4,243	6,707,383	4,535	8,168,326
Large-rich – large-rich	2,806	95,023,471	4,206	123,987,355	4,070	128,618,891	4,193	134,255,703
Middle-middle – large-rich	9,004	8,449,385	11,971	13,945,373	13,227	18,782,904	13,681	18,955,789
Megacities – large-rich	4,262	187,832,636	6,381	242,954,928	6,323	263,509,560	6,548	288,493,322
Small-rich – large-rich	18,921	61,875,366	22,302	74,895,178	21,358	81,702,652	21,571	87,019,611
Large-poor – large-rich	10,932	15,408,250	14,411	28,158,668	14,966	33,173,545	15,120	37,089,637
Middle-rich – large-rich	10,010	145,461,409	13,317	179,421,111	12,970	179,198,857	12,975	186,107,657
Middle-poor – large-rich	5,846	4,496,376	7,789	7,631,284	8,726	12,500,169	8,846	12,683,918
Middle-middle – middle-middle	3,252	5,232,395	4,352	6,853,564	5,741	9,656,462	6,408	10,645,839
Megacities – middle-middle	6,811	38,213,403	9,663	59,194,585	10,822	90,990,793	11,430	100,371,130
Small-rich – middle-middle	9,136	1,279,869	11,512	2,001,832	12,268	3,908,553	12,558	3,285,321
Large-poor – middle-middle	7,706	19,166,333	11,047	31,868,729	13,763	49,743,614	14,829	55,056,943
Middle-rich – middle-middle	10,703	5,918,899	13,257	8,965,865	14,320	12,568,492	14,763	12,494,518
Middle-poor – middle-middle	1,755	716,745	2,732	1,169,997	3,991	1,671,988	4,343	1,492,697
Megacities – megacities	1,408	129,616,129	2,359	194,751,657	2,435	244,182,276	2,512	268,766,800
Small-rich – megacities	14,524	54,970,474	17,341	67,976,136	17,127	73,697,195	17,539	83,588,367
Large-poor – megacities	8,230	109,039,677	11,427	174,175,196	12,291	267,847,087	12,648	292,623,248
Middle-rich – megacities	8,060	149,434,205	10,258	176,457,456	10,232	184,910,841	10,291	206,384,256
Middle-poor – megacities	5,356	20,948,603	7,387	48,898,228	8,435	77,323,982	8,743	83,870,970
Small-rich – small-rich	20,555	11,212,993	21,378	12,954,050	19,796	13,471,926	19,872	13,431,502
Large-poor – small-rich	17,075	2,145,762	20,531	3,489,575	21,864	6,778,877	22,169	4,413,014
Middle-rich – small-rich	27,693	46,772,635	30,375	57,356,144	28,659	53,305,924	28,726	58,318,254
Middle-poor – small-rich	5,831	647,413	7,460	1,228,117	8,328	2,499,062	8,748	1,906,900
Large-poor – large-poor	5,817	55,363,267	8,920	99,337,549	10,137	190,772,648	10,424	206,036,231
Middle-rich – large-poor	15,354	8,820,375	18,741	13,604,503	19,641	17,310,285	19,621	16,947,207
Middle-poor – large-poor	4,661	7,839,170	7,104	14,570,235	9,348	22,766,893	9,712	25,580,386
Middle-rich – middle-rich	8,067	56,354,023	9,723	68,956,011	9,426	65,440,463	9,350	70,662,813
Middle-poor – middle-rich	6,745	1,764,152	8,385	3,132,879	9,400	4,461,468	9,440	3,794,987
Middle-poor – middle-poor	1,397	3,079,450	2,199	5,902,073	3,172	7,779,286	3,281	8,990,225

A.1. Total passenger and connection numbers for every cluster pair for 2002, 2007, 2011 and the base year 2012

	0	5	10	25	50	100	150	200	300	400	500
Very small-rich – very small-rich	0.085956	0.459443	0.58293	0.737288	0.81931	0.887712	0.916768	0.925545	0.94431	0.95339	0.960654
Small-poor – very small-rich	0.134467	0.564529	0.694745	0.818779	0.879057	0.921175	0.945904	0.954405	0.967156	0.973725	0.979521
Large-rich – very small-rich	0.055991	0.358653	0.493735	0.64213	0.743931	0.813234	0.857478	0.877839	0.898982	0.915818	0.929522
Middle-middle – very small-rich	0.114629	0.579694	0.714338	0.840611	0.906114	0.936317	0.949782	0.959607	0.971616	0.979985	0.981441
Megacities – very small-rich	0.080596	0.427426	0.564171	0.720779	0.796028	0.870512	0.896486	0.917112	0.944232	0.954545	0.957983
Small-rich – very small-rich	0.081302	0.440313	0.578367	0.738659	0.824942	0.891123	0.919054	0.938089	0.954101	0.963885	0.970646
Large-poor – very small-rich	0.115445	0.546066	0.678323	0.821116	0.884331	0.943735	0.953598	0.96234	0.973773	0.984308	0.988119
Middle-rich – very small-rich	0.064342	0.378447	0.507132	0.671632	0.776862	0.84691	0.879873	0.896038	0.918542	0.929635	0.936926
Middle-poor – very small-rich	0.134574	0.621809	0.76117	0.88883	0.930851	0.968617	0.981383	0.984574	0.989894	0.989894	0.990957
Small-poor – small-poor	0.117145	0.485052	0.592434	0.701647	0.76388	0.835876	0.85845	0.882855	0.910921	0.924954	0.942038
Large-rich – small-poor	0.063263	0.362982	0.49264	0.646101	0.750705	0.82399	0.864078	0.88788	0.913561	0.92327	0.93016
Middle-middle – small-poor	0.06544	0.393546	0.500603	0.64415	0.733414	0.803679	0.844391	0.870627	0.893245	0.911339	0.933655
Megacities – small-poor	0.060392	0.324034	0.441447	0.58553	0.692826	0.772532	0.822808	0.854384	0.888719	0.907725	0.918761
Small-rich – small-poor	0.101545	0.521952	0.64778	0.776551	0.85455	0.903606	0.922001	0.939907	0.957812	0.966397	0.972774
Large-poor – small-poor	0.074375	0.36103	0.463103	0.606624	0.700368	0.776099	0.81658	0.842146	0.879527	0.899671	0.915359
Middle-rich – small-poor	0.073675	0.433314	0.569016	0.735585	0.806931	0.870122	0.900699	0.915259	0.934479	0.948165	0.957484
Middle-poor – small-poor	0.109453	0.4801	0.628939	0.785655	0.849502	0.888474	0.91335	0.922886	0.935738	0.946517	0.955638
Large-rich – large-rich	0.018248	0.105839	0.145985	0.248175	0.343066	0.419708	0.481752	0.532847	0.60219	0.653285	0.693431
Middle-middle – large-rich	0.042086	0.24775	0.357861	0.516411	0.622022	0.733192	0.777131	0.808629	0.852832	0.875331	0.892536
Megacities – large-rich	0.01004	0.058233	0.104418	0.184739	0.228916	0.309237	0.363454	0.427711	0.5	0.554217	0.594378
Small-rich – large-rich	0.029731	0.202973	0.290452	0.422527	0.534591	0.639794	0.696398	0.734706	0.769583	0.813608	0.832476
Large-poor – large-rich	0.030162	0.197989	0.282676	0.429234	0.541377	0.624903	0.671694	0.704176	0.769915	0.809745	0.834107
Middle-rich – large-rich	0.041433	0.198208	0.286674	0.43897	0.56551	0.640538	0.690929	0.721165	0.758119	0.784994	0.81523
Middle-poor – large-rich	0.052702	0.322465	0.447075	0.605628	0.711925	0.796338	0.837874	0.861992	0.893703	0.910228	0.919607
Middle-middle – middle-middle	0.062704	0.326205	0.42008	0.547662	0.633563	0.706415	0.750272	0.783617	0.822399	0.851758	0.873505
Megacities – middle-middle	0.04742	0.266329	0.382941	0.526394	0.625708	0.724128	0.77632	0.804056	0.844915	0.864897	0.881002
Small-rich – middle-middle	0.110157	0.528168	0.680419	0.817801	0.89801	0.941571	0.954974	0.964188	0.975497	0.982827	0.985759
Large-poor – middle-middle	0.070277	0.358272	0.470966	0.602129	0.684145	0.75583	0.792612	0.817029	0.849742	0.867428	0.88668
Middle-rich – middle-middle	0.065792	0.380737	0.517748	0.693873	0.797875	0.868415	0.906172	0.923129	0.938955	0.948677	0.95953
Middle-poor – middle-middle	0.109819	0.462963	0.608958	0.774332	0.861757	0.912575	0.927649	0.938846	0.949182	0.955211	0.963394
Megacities – megacities	0.004739	0.037915	0.085308	0.151659	0.279621	0.322275	0.369668	0.450237	0.516588	0.545024	0.56872
Small-rich – megacities	0.037019	0.245311	0.345015	0.485686	0.612043	0.705331	0.757651	0.786772	0.827739	0.851431	0.880553
Large-poor – megacities	0.028421	0.183922	0.267154	0.398701	0.499391	0.593179	0.641088	0.682907	0.735688	0.777101	0.808364
Middle-rich – megacities	0.031111	0.162222	0.233333	0.355556	0.458889	0.562222	0.621111	0.656667	0.71	0.748889	0.776667
Middle-poor – megacities	0.037412	0.268046	0.371479	0.53213	0.62456	0.709507	0.761004	0.790053	0.827465	0.854754	0.868398
Small-rich – small-rich	0.06142	0.348944	0.467179	0.625336	0.715163	0.785797	0.830326	0.862188	0.898656	0.915931	0.927447
Large-poor – small-rich	0.078285	0.419053	0.559361	0.726148	0.824579	0.903871	0.925457	0.940999	0.958411	0.966038	0.973809
Middle-rich – small-rich	0.046102	0.303598	0.430285	0.565967	0.667541	0.742129	0.781859	0.814468	0.8497	0.875562	0.886432
Middle-poor – small-rich	0.130838	0.604584	0.742095	0.874093	0.939368	0.965767	0.979692	0.986365	0.993037	0.995358	0.996229
Large-poor – large-poor	0.058871	0.323118	0.438172	0.568011	0.667742	0.754301	0.791935	0.823387	0.858065	0.880914	0.894355
Middle-rich – large-poor	0.043972	0.29669	0.414032	0.599555	0.712451	0.795455	0.833992	0.86413	0.892045	0.91749	0.934042
Middle-poor – large-poor	0.073061	0.396235	0.521515	0.660018	0.750784	0.82407	0.863066	0.889063	0.915733	0.934334	0.945092
Middle-rich – middle-rich	0.036988	0.26288	0.363276	0.458388	0.544254	0.63144	0.675033	0.717305	0.758256	0.795244	0.81638
Middle-poor – middle-rich	0.078823	0.430803	0.571377	0.749728	0.840538	0.903015	0.931348	0.942245	0.958954	0.970578	0.976026
Middle-poor – middle-poor	0.066914	0.406444	0.526022	0.651797	0.747831	0.80855	0.840768	0.856258	0.884758	0.909542	0.92627

A.2. QA approach accuracy for connection numbers with the correctly predicted passenger numbers for 2012 from 2002 for every cluster pair at specified intervals

	0	5	10	25	50	100	150	200	300	400	500
Very small-rich – very small-rich	0.003942	0.030803	0.047638	0.078862	0.104146	0.131718	0.149665	0.157247	0.167977	0.180028	0.189822
Small-poor – very small-rich	0.007551	0.054944	0.080338	0.115641	0.150063	0.194681	0.219908	0.237059	0.269194	0.281712	0.298745
Large-rich – very small-rich	0.002091	0.017514	0.028503	0.045081	0.060174	0.078785	0.089305	0.10715	0.116285	0.12332	0.134556
Middle-middle – very small-rich	0.004533	0.034994	0.050873	0.079206	0.095396	0.110847	0.125517	0.138847	0.146551	0.155232	0.155363
Megacities – very small-rich	0.001746	0.013978	0.022204	0.036493	0.043851	0.053487	0.061472	0.067916	0.075747	0.080827	0.083582
Small-rich – very small-rich	0.002273	0.020427	0.031951	0.051139	0.065488	0.079283	0.090028	0.096534	0.10572	0.11888	0.12416
Large-poor – very small-rich	0.008244	0.059935	0.086763	0.129218	0.158645	0.186341	0.195668	0.20139	0.213596	0.222214	0.237742
Middle-rich – very small-rich	0.000426	0.003879	0.006182	0.01083	0.014295	0.019278	0.021489	0.023289	0.025317	0.026897	0.027516
Middle-poor – very small-rich	0.008264	0.061613	0.092255	0.132203	0.160304	0.18314	0.192866	0.196797	0.225739	0.225739	0.225848
Small-poor – small-poor	0.001241	0.007905	0.012895	0.019125	0.027412	0.039582	0.05201	0.064421	0.082641	0.096454	0.109508
Large-rich – small-poor	0.001158	0.012271	0.020133	0.034843	0.048347	0.066758	0.080621	0.088344	0.104258	0.114888	0.125832
Middle-middle – small-poor	0.001205	0.010261	0.017409	0.032621	0.048611	0.072776	0.092501	0.11153	0.140693	0.171208	0.206494
Megacities – small-poor	0.000605	0.004983	0.008429	0.015822	0.024067	0.03523	0.043611	0.050876	0.060949	0.069336	0.074408
Small-rich – small-poor	0.002653	0.022581	0.035371	0.055009	0.074318	0.098649	0.114432	0.131268	0.149991	0.166418	0.182979
Large-poor – small-poor	0.000552	0.003916	0.006113	0.011507	0.017794	0.02655	0.034437	0.040208	0.051239	0.057939	0.064354
Middle-rich – small-poor	0.000902	0.009362	0.014415	0.024671	0.031579	0.041301	0.047891	0.052301	0.0601	0.06839	0.075265
Middle-poor – small-poor	0.000891	0.005942	0.009253	0.015415	0.020436	0.026667	0.031374	0.033668	0.039358	0.047862	0.052123
Large-rich – large-rich	0.000534	0.004064	0.006005	0.009184	0.016176	0.032523	0.04291	0.043988	0.063297	0.075807	0.091836
Middle-middle – large-rich	0.001052	0.010097	0.018861	0.03468	0.050845	0.078388	0.093035	0.107412	0.126936	0.147364	0.165691
Megacities – large-rich	0.000227	0.001166	0.003602	0.007625	0.011651	0.023278	0.036162	0.058508	0.068758	0.083208	0.091525
Small-rich – large-rich	0.000223	0.002899	0.004386	0.006947	0.010917	0.014838	0.018567	0.0209	0.024343	0.029679	0.033268
Large-poor – large-rich	0.001199	0.011401	0.017906	0.034767	0.055266	0.075036	0.088079	0.100906	0.124007	0.136016	0.146817
Middle-rich – large-rich	0.000259	0.001915	0.003275	0.006835	0.010396	0.013979	0.016913	0.020361	0.023086	0.025938	0.037154
Middle-poor – large-rich	0.001611	0.016181	0.029342	0.04882	0.072989	0.09755	0.11878	0.128695	0.146124	0.153487	0.166134
Middle-middle – middle-middle	0.000969	0.009176	0.014426	0.027436	0.046424	0.068553	0.087293	0.106826	0.138165	0.165337	0.196151
Megacities – middle-middle	0.000432	0.004489	0.008686	0.016807	0.025269	0.037491	0.04973	0.05951	0.068488	0.075652	0.085425
Small-rich – middle-middle	0.004231	0.033711	0.051599	0.079047	0.10138	0.125474	0.137324	0.145266	0.157323	0.160848	0.170182
Large-poor – middle-middle	0.000474	0.003788	0.005873	0.010485	0.015317	0.023028	0.029905	0.036061	0.04623	0.054232	0.061164
Middle-rich – middle-middle	0.000903	0.00826	0.013427	0.022834	0.03229	0.046655	0.054903	0.060294	0.06491	0.071213	0.07438
Middle-poor – middle-middle	0.000948	0.007549	0.011664	0.020556	0.028406	0.035409	0.041403	0.048106	0.053512	0.059436	0.067987
Megacities – megacities	1.04E-05	0.000239	0.01653	0.017674	0.024688	0.027044	0.04165	0.067014	0.074429	0.083175	0.091993
Small-rich – megacities	0.000381	0.005779	0.00978	0.018178	0.028237	0.03589	0.040488	0.042525	0.048607	0.057204	0.063755
Large-poor – megacities	0.000282	0.002843	0.005166	0.011231	0.018696	0.030851	0.037641	0.045639	0.058838	0.06972	0.083942
Middle-rich – megacities	0.001535	0.005447	0.008928	0.019289	0.037079	0.057452	0.070327	0.077942	0.118309	0.134097	0.143684
Middle-poor – megacities	0.000157	0.001567	0.002533	0.00453	0.007143	0.009953	0.012761	0.014249	0.017858	0.022042	0.024051
Small-rich – small-rich	0.001101	0.011938	0.018882	0.041233	0.05401	0.065002	0.08233	0.095591	0.115463	0.127441	0.136118
Large-poor – small-rich	0.003916	0.032113	0.050842	0.082459	0.105123	0.127376	0.137388	0.149275	0.1685	0.178713	0.188809
Middle-rich – small-rich	0.000187	0.001714	0.003082	0.004794	0.006556	0.010181	0.011767	0.012995	0.016197	0.019606	0.022351
Middle-poor – small-rich	0.006024	0.04608	0.067213	0.100805	0.129512	0.149412	0.16138	0.171495	0.17787	0.183025	0.189359
Large-poor – large-poor	0.000197	0.001931	0.003128	0.006212	0.00981	0.01365	0.016068	0.018415	0.022116	0.024788	0.027177
Middle-rich – large-poor	0.003666	0.03347	0.05378	0.097779	0.140804	0.179167	0.205317	0.219715	0.245507	0.274161	0.312811
Middle-poor – large-poor	0.000509	0.004439	0.00724	0.012205	0.018195	0.026557	0.032137	0.037622	0.045086	0.051412	0.056621
Middle-rich – middle-rich	0.000157	0.001475	0.002226	0.003256	0.005356	0.00657	0.007209	0.007542	0.008761	0.009997	0.013397
Middle-poor – middle-rich	0.015106	0.137763	0.219988	0.379995	0.498356	0.628625	0.70146	0.774	0.84023	0.89633	0.900625
Middle-poor – middle-poor	0.000262	0.002675	0.004166	0.007138	0.011527	0.019209	0.02541	0.027173	0.033387	0.041662	0.045111

A.3. Passenger numbers covered by connections from Tab. A.2 for 2012 from 2002 for every cluster pair at specified intervals

	0	5	10	25	50	100	150	200	300	400	500
Very small-rich – very small-rich	0.083619	0.47464	0.596984	0.747772	0.82111	0.886223	0.910898	0.92769	0.942769	0.955449	0.959561
Small-poor – very small-rich	0.116105	0.558052	0.682183	0.795078	0.859283	0.904227	0.919208	0.936865	0.947566	0.956126	0.964687
Large-rich – very small-rich	0.056738	0.346099	0.478487	0.626478	0.72766	0.805674	0.847281	0.868085	0.900236	0.921986	0.940898
Middle-middle – very small-rich	0.138862	0.581485	0.705882	0.840405	0.898746	0.927676	0.93973	0.952266	0.963356	0.967695	0.976374
Megacities – very small-rich	0.089893	0.415452	0.547619	0.695335	0.77551	0.835277	0.875607	0.896987	0.920797	0.931973	0.943149
Small-rich – very small-rich	0.077844	0.414671	0.55432	0.712789	0.791061	0.8642	0.894354	0.912532	0.939478	0.948674	0.960009
Large-poor – very small-rich	0.112605	0.56601	0.694444	0.8181	0.885603	0.934588	0.953405	0.959976	0.973118	0.9773	0.98178
Middle-rich – very small-rich	0.070337	0.367673	0.493783	0.646536	0.742096	0.825577	0.85897	0.878863	0.900888	0.918295	0.927531
Middle-poor – very small-rich	0.143047	0.62342	0.76181	0.878244	0.926813	0.964072	0.974717	0.982036	0.986693	0.990685	0.992681
Small-poor – small-poor	0.086596	0.408133	0.509789	0.618223	0.703313	0.779367	0.82003	0.839608	0.870482	0.888554	0.904367
Large-rich – small-poor	0.053688	0.371615	0.485061	0.643324	0.746499	0.816993	0.853408	0.880486	0.907096	0.921102	0.932773
Middle-middle – small-poor	0.078036	0.374952	0.480396	0.601827	0.690902	0.773887	0.813095	0.849638	0.87933	0.898363	0.91397
Megacities – small-poor	0.06174	0.32402	0.420009	0.5516	0.644885	0.742677	0.785038	0.811627	0.845877	0.878774	0.899504
Small-rich – small-poor	0.097456	0.534934	0.646005	0.775708	0.8445	0.893587	0.916517	0.927625	0.941598	0.954497	0.961304
Large-poor – small-poor	0.081245	0.381713	0.486367	0.620766	0.699807	0.765078	0.802809	0.831727	0.862848	0.883779	0.901129
Middle-rich – small-poor	0.092094	0.440286	0.551724	0.738856	0.832632	0.900757	0.91968	0.930614	0.939445	0.946173	0.953743
Middle-poor – small-poor	0.10087	0.482847	0.606247	0.734255	0.822325	0.882744	0.907322	0.916027	0.926267	0.940092	0.949821
Large-rich – large-rich	0.008	0.088	0.144	0.248	0.368	0.496	0.52	0.576	0.672	0.712	0.72
Middle-middle – large-rich	0.044795	0.262547	0.369556	0.50477	0.623393	0.720863	0.777271	0.807134	0.855247	0.881377	0.904604
Megacities – large-rich	0.019608	0.098039	0.151961	0.22549	0.269608	0.343137	0.377451	0.392157	0.480392	0.534314	0.568627
Small-rich – large-rich	0.03218	0.226066	0.310539	0.452936	0.541432	0.624296	0.670957	0.70716	0.765084	0.790829	0.81255
Large-poor – large-rich	0.026515	0.183333	0.265909	0.387879	0.506818	0.6	0.663636	0.688636	0.752273	0.790909	0.815152
Middle-rich – large-rich	0.046847	0.273874	0.369369	0.513514	0.605405	0.693694	0.727928	0.771171	0.789189	0.81982	0.830631
Middle-poor – large-rich	0.05189	0.310058	0.42729	0.595131	0.68738	0.772582	0.830878	0.858424	0.887892	0.905189	0.919283
Middle-middle – middle-middle	0.066205	0.323669	0.420597	0.544353	0.615318	0.696668	0.744267	0.774989	0.814799	0.840329	0.856772
Megacities – middle-middle	0.05654	0.287015	0.382846	0.521802	0.616195	0.711548	0.759943	0.797796	0.829899	0.853378	0.877815
Small-rich – middle-middle	0.096988	0.536747	0.683735	0.826506	0.89006	0.931928	0.946687	0.957229	0.968675	0.973193	0.976807
Large-poor – middle-middle	0.069103	0.362791	0.471096	0.62392	0.709635	0.788704	0.821041	0.843189	0.873976	0.89103	0.903433
Middle-rich – middle-middle	0.077426	0.401176	0.544593	0.689644	0.793205	0.864097	0.899379	0.920614	0.937275	0.949363	0.960144
Middle-poor – middle-middle	0.095883	0.488438	0.619289	0.766497	0.834743	0.882121	0.908629	0.926114	0.944726	0.956007	0.958827
Megacities – megacities	0	0.050505	0.080808	0.161616	0.242424	0.343434	0.393939	0.424242	0.444444	0.464646	0.484848
Small-rich – megacities	0.047401	0.233945	0.32263	0.453364	0.548165	0.657492	0.701835	0.74159	0.788991	0.825688	0.853211
Large-poor – megacities	0.028731	0.207502	0.286512	0.422187	0.513966	0.592179	0.651237	0.691939	0.736632	0.777334	0.802873
Middle-rich – megacities	0.023555	0.190578	0.267666	0.415418	0.498929	0.571734	0.599572	0.635974	0.69379	0.760171	0.777302
Middle-poor – megacities	0.0475	0.28625	0.391875	0.53125	0.6375	0.714375	0.760625	0.7925	0.831875	0.85625	0.87
Small-rich – small-rich	0.072076	0.369928	0.473031	0.608592	0.705967	0.781384	0.815274	0.844391	0.873986	0.900716	0.919809
Large-poor – small-rich	0.090664	0.443568	0.572199	0.726349	0.830498	0.902075	0.93029	0.947925	0.959544	0.965145	0.970124
Middle-rich – small-rich	0.046392	0.292877	0.412371	0.559513	0.664011	0.75492	0.803187	0.833646	0.867385	0.888941	0.904405
Middle-poor – small-rich	0.118911	0.561181	0.706559	0.846951	0.920215	0.956272	0.969697	0.979287	0.98926	0.991945	0.99463
Large-poor – large-poor	0.070776	0.352968	0.457078	0.594977	0.687671	0.7621	0.808676	0.82968	0.864384	0.883105	0.896804
Middle-rich – large-poor	0.058199	0.312623	0.422336	0.570193	0.672041	0.754621	0.817538	0.843099	0.872985	0.89186	0.914274
Middle-poor – large-poor	0.0867	0.418272	0.534183	0.674953	0.765693	0.828465	0.85954	0.882225	0.905842	0.923244	0.93381
Middle-rich – middle-rich	0.034672	0.262774	0.34854	0.467153	0.560219	0.645985	0.70438	0.740876	0.782847	0.821168	0.84854
Middle-poor – middle-rich	0.077935	0.426693	0.547004	0.705796	0.80565	0.872869	0.899172	0.917194	0.933755	0.948855	0.957136
Middle-poor – middle-poor	0.076923	0.441736	0.557502	0.683168	0.777609	0.836253	0.860625	0.881188	0.90556	0.923077	0.93374

A.4. QA approach accuracy for connection numbers with correctly predicted passenger numbers for 2012 from 2007 for every cluster pair at specified intervals

	0	5	10	25	50	100	150	200	300	400	500
Very small-rich – very small-rich	0.005051	0.042826	0.059469	0.090834	0.110495	0.134503	0.150116	0.161951	0.182896	0.198443	0.203569
Small-poor – very small-rich	0.033312	0.228547	0.321737	0.448099	0.512756	0.597934	0.622742	0.631035	0.679176	0.73554	0.739897
Large-rich – very small-rich	0.002722	0.020447	0.033041	0.046741	0.061973	0.076291	0.084425	0.090585	0.107396	0.127721	0.148564
Middle-middle – very small-rich	0.01107	0.0645	0.090331	0.135515	0.154985	0.171648	0.186927	0.200035	0.202245	0.202841	0.213469
Megacities – very small-rich	0.002072	0.015537	0.024207	0.035404	0.057731	0.062982	0.067788	0.070556	0.071531	0.072808	0.073833
Small-rich – very small-rich	0.003494	0.026602	0.040197	0.060923	0.07428	0.088809	0.099617	0.106979	0.125758	0.139055	0.141778
Large-poor – very small-rich	0.023551	0.176725	0.247377	0.339205	0.394359	0.431803	0.453845	0.470562	0.47576	0.476947	0.491289
Middle-rich – very small-rich	0.001202	0.008866	0.013524	0.019311	0.025971	0.030719	0.032992	0.035525	0.04223	0.049181	0.05407
Middle-poor – very small-rich	0.020042	0.141649	0.20428	0.271816	0.300418	0.331524	0.347129	0.368946	0.399426	0.420825	0.421086
Small-poor – small-poor	0.000968	0.008278	0.012556	0.020322	0.030871	0.041222	0.054108	0.065722	0.086825	0.099477	0.109398
Large-rich – small-poor	0.003738	0.043013	0.068128	0.10578	0.137143	0.160389	0.177373	0.192707	0.212242	0.23602	0.251149
Middle-middle – small-poor	0.001577	0.012952	0.020065	0.034493	0.047658	0.071612	0.088216	0.110639	0.131391	0.157606	0.17825
Megacities – small-poor	0.002628	0.023201	0.035835	0.065538	0.087033	0.121155	0.138146	0.148513	0.178393	0.202785	0.210393
Small-rich – small-poor	0.00315	0.025947	0.036102	0.054452	0.067848	0.082155	0.091001	0.096477	0.10534	0.117535	0.123496
Large-poor – small-poor	0.000765	0.005987	0.009248	0.015022	0.021834	0.032343	0.03925	0.04589	0.056742	0.067837	0.077876
Middle-rich – small-poor	0.002233	0.016843	0.024238	0.036718	0.046327	0.058335	0.070853	0.07388	0.074862	0.081742	0.088792
Middle-poor – small-poor	0.001025	0.007827	0.011984	0.017262	0.023825	0.029551	0.034858	0.037397	0.043461	0.052284	0.060629
Large-rich – large-rich	4.17E-05	0.001148	0.001607	0.002733	0.0058	0.010203	0.010474	0.0111	0.016087	0.016609	0.01665
Middle-middle – large-rich	0.002448	0.021175	0.036522	0.062819	0.097776	0.139821	0.15805	0.177298	0.204261	0.231239	0.252069
Megacities – large-rich	7.79E-05	0.000655	0.001146	0.003164	0.00505	0.008043	0.010996	0.012555	0.018899	0.019499	0.020395
Small-rich – large-rich	0.000191	0.002169	0.003247	0.006797	0.008917	0.012383	0.013567	0.01535	0.023954	0.025777	0.025957
Large-poor – large-rich	0.000453	0.006473	0.011726	0.020179	0.034254	0.046679	0.07328	0.083752	0.102208	0.122683	0.130459
Middle-rich – large-rich	0.000516	0.004817	0.007252	0.012476	0.015186	0.020229	0.022561	0.025271	0.025882	0.029566	0.033983
Middle-poor – large-rich	0.003866	0.02934	0.04498	0.077242	0.107771	0.136646	0.164962	0.184385	0.199943	0.226988	0.237806
Middle-middle – middle-middle	0.00077	0.00714	0.010954	0.022181	0.031564	0.049579	0.063161	0.073977	0.099182	0.12751	0.146961
Megacities – middle-middle	0.000395	0.003334	0.005546	0.009291	0.013021	0.018443	0.020743	0.022989	0.025506	0.027851	0.032428
Small-rich – middle-middle	0.004566	0.038277	0.055386	0.07908	0.090673	0.10548	0.117159	0.124951	0.136307	0.140701	0.147445
Large-poor – middle-middle	0.000704	0.005946	0.009179	0.016573	0.02359	0.032807	0.043665	0.049168	0.057762	0.065739	0.075989
Middle-rich – middle-middle	0.001233	0.009321	0.015358	0.023066	0.030931	0.039407	0.045832	0.048585	0.053563	0.055692	0.060575
Middle-poor – middle-middle	0.003593	0.033343	0.04798	0.077121	0.101763	0.133039	0.16459	0.180001	0.19508	0.221812	0.233752
Megacities – megacities	0	0.000956	0.001824	0.012639	0.018806	0.031185	0.038699	0.049644	0.049731	0.050513	0.050903
Small-rich – megacities	0.000586	0.004964	0.007719	0.0128	0.016235	0.019046	0.023684	0.027881	0.029325	0.032598	0.044011
Large-poor – megacities	0.000434	0.005822	0.012455	0.02168	0.033019	0.043764	0.058066	0.067624	0.082467	0.094311	0.11148
Middle-rich – megacities	0.000893	0.013717	0.025811	0.051175	0.075078	0.097317	0.10584	0.12256	0.174019	0.259851	0.263829
Middle-poor – megacities	0.001722	0.020978	0.032458	0.056314	0.082583	0.118455	0.153841	0.179729	0.209794	0.24087	0.25632
Small-rich – small-rich	0.004415	0.033322	0.047347	0.090119	0.108262	0.130708	0.141025	0.154195	0.169907	0.184814	0.197959
Large-poor – small-rich	0.003969	0.02903	0.043246	0.064504	0.081561	0.092899	0.097979	0.104462	0.112128	0.116861	0.120932
Middle-rich – small-rich	0.000489	0.0042	0.007621	0.011314	0.014103	0.017328	0.020662	0.027354	0.033424	0.037747	0.042938
Middle-poor – small-rich	0.032312	0.246648	0.367817	0.525015	0.658301	0.730949	0.750983	0.798912	0.808875	0.831063	0.863644
Large-poor – large-poor	0.000424	0.003427	0.005313	0.009047	0.012528	0.017418	0.024033	0.026781	0.032268	0.035393	0.03829
Middle-rich – large-poor	0.004015	0.030948	0.044166	0.066921	0.085779	0.106211	0.122034	0.128461	0.146637	0.160315	0.189861
Middle-poor – large-poor	0.001939	0.015468	0.023812	0.039858	0.055446	0.076448	0.090174	0.101864	0.116322	0.129897	0.151231
Middle-rich – middle-rich	0.000154	0.001871	0.003094	0.00487	0.005564	0.007501	0.009342	0.011959	0.01247	0.014337	0.016898
Middle-poor – middle-rich	0.008619	0.061155	0.087487	0.140902	0.178982	0.210291	0.226358	0.252141	0.261126	0.276625	0.279351
Middle-poor – middle-poor	0.0009	0.011084	0.01536	0.024666	0.037118	0.053721	0.067548	0.074663	0.089368	0.102051	0.107183

A.5. Passenger numbers covered by connections from Tab. A.4 for 2012 from 2007 for every cluster pair at specified intervals

	0	5	10	25	50	100	150	200	300	400	500
Very small-rich – very small-rich	0.068241	0.543307	0.643045	0.737533	0.7979	0.868766	0.913386	0.918635	0.944882	0.958005	0.958005
Small-poor – very small-rich	0.088123	0.643678	0.735632	0.789272	0.869732	0.89272	0.961686	0.961686	0.969349	0.980843	0.980843
Large-rich – very small-rich	0.075064	0.300254	0.42112	0.559796	0.675573	0.746819	0.78117	0.826972	0.863868	0.886768	0.902036
Middle-middle – very small-rich	0.111386	0.487624	0.569307	0.658416	0.784653	0.856436	0.930693	0.94802	0.95297	0.985149	0.985149
Megacities – very small-rich	0.044248	0.256637	0.370417	0.546144	0.639697	0.726928	0.785082	0.817952	0.845765	0.876106	0.922882
Small-rich – very small-rich	0.076982	0.431871	0.488068	0.625096	0.787529	0.823711	0.847575	0.856043	0.888376	0.924557	0.925327
Large-poor – very small-rich	0.046875	0.44401	0.621094	0.645833	0.885417	0.898438	0.914063	0.915365	0.915365	0.915365	0.915365
Middle-rich – very small-rich	0.144467	0.560451	0.689549	0.723361	0.759221	0.765369	0.770492	0.770492	0.772541	0.773566	0.773566
Middle-poor – very small-rich	0.097122	0.730216	0.78777	0.809353	0.938849	0.94964	0.956835	0.956835	0.956835	0.956835	0.956835
Small-poor – small-poor	0.139535	0.503876	0.620155	0.821705	0.852713	0.860465	0.906977	0.914729	0.945736	0.968992	0.968992
Large-rich – small-poor	0.060383	0.328424	0.443299	0.603829	0.696613	0.776141	0.808542	0.840943	0.877761	0.895434	0.94109
Middle-middle – small-poor	0.094987	0.366755	0.416887	0.527704	0.604222	0.773087	0.878628	0.883905	0.897098	0.92876	0.931398
Megacities – small-poor	0.056385	0.270315	0.389718	0.538972	0.600332	0.661692	0.704809	0.769486	0.776119	0.802653	0.832504
Small-rich – small-poor	0.10034	0.491497	0.596939	0.717687	0.870748	0.892857	0.908163	0.930272	0.940476	0.942177	0.97619
Large-poor – small-poor	0.053492	0.410104	0.526003	0.607727	0.769688	0.793462	0.888559	0.89896	0.907875	0.915305	0.979198
Middle-rich – small-poor	0.041237	0.539028	0.777614	0.799705	0.815906	0.829161	0.833579	0.836524	0.845361	0.991163	0.991163
Middle-poor – small-poor	0.054054	0.401544	0.440154	0.505792	0.57529	0.590734	0.594595	0.594595	0.598456	0.621622	0.687259
Large-rich – large-rich	0.020619	0.154639	0.185567	0.237113	0.278351	0.360825	0.391753	0.463918	0.649485	0.649485	0.670103
Middle-middle – large-rich	0.035613	0.210826	0.262108	0.324786	0.367521	0.531339	0.615385	0.703704	0.860399	0.900285	0.903134
Megacities – large-rich	0.04	0.2	0.232	0.336	0.464	0.568	0.592	0.6	0.664	0.768	0.784
Small-rich – large-rich	0.064286	0.285714	0.408929	0.544643	0.628571	0.741071	0.7625	0.8125	0.864286	0.878571	0.889286
Large-poor – large-rich	0.025701	0.186916	0.21729	0.406542	0.434579	0.471963	0.635514	0.663551	0.67757	0.890187	0.897196
Middle-rich – large-rich	0.015385	0.203846	0.303846	0.457692	0.507692	0.553846	0.630769	0.7	0.738462	0.773077	0.788462
Middle-poor – large-rich	0.012407	0.101737	0.116625	0.205955	0.255583	0.543424	0.841191	0.885856	0.908189	0.952854	0.955335
Middle-middle – middle-middle	0.0625	0.298077	0.348558	0.420673	0.584135	0.721154	0.882212	0.911058	0.935096	0.956731	0.963942
Megacities – middle-middle	0.029772	0.231173	0.309982	0.499124	0.577933	0.642732	0.707531	0.770578	0.782837	0.82662	0.828371
Small-rich – middle-middle	0.056229	0.573319	0.707828	0.791621	0.864388	0.941566	0.949283	0.954796	0.965821	0.971334	0.972437
Large-poor – middle-middle	0.068783	0.343915	0.483598	0.689947	0.733333	0.771429	0.848677	0.851852	0.967196	0.971429	0.975661
Middle-rich – middle-middle	0.042194	0.28692	0.345992	0.422996	0.468354	0.89557	0.910338	0.915612	0.92827	0.953586	0.96097
Middle-poor – middle-middle	0.106618	0.334559	0.404412	0.922794	0.930147	0.963235	0.963235	0.966912	0.985294	0.992647	0.992647
Megacities – megacities	0.028571	0.171429	0.314286	0.4	0.485714	0.6	0.6	0.685714	0.714286	0.8	0.8
Small-rich – megacities	0.061753	0.428287	0.539841	0.681275	0.727092	0.778884	0.788845	0.828685	0.850598	0.866534	0.876494
Large-poor – megacities	0.029326	0.196481	0.319648	0.457478	0.510264	0.615836	0.656891	0.686217	0.841642	0.856305	0.870968
Middle-rich – megacities	0.065789	0.368421	0.480263	0.585526	0.697368	0.809211	0.835526	0.842105	0.855263	0.861842	0.875
Middle-poor – megacities	0.053221	0.29972	0.386555	0.588235	0.641457	0.689076	0.711485	0.722689	0.784314	0.803922	0.834734
Small-rich – small-rich	0.069913	0.496879	0.560549	0.726592	0.782772	0.833958	0.842697	0.895131	0.898876	0.911361	0.911361
Large-poor – small-rich	0.132239	0.511668	0.733794	0.830596	0.840104	0.880726	0.966292	0.971478	0.977528	0.977528	0.977528
Middle-rich – small-rich	0.033932	0.382236	0.505988	0.862275	0.886228	0.907186	0.922156	0.924152	0.93014	0.937126	0.944112
Middle-poor – small-rich	0.142105	0.764912	0.921053	0.950877	0.957895	0.964912	0.984211	0.991228	0.992982	0.992982	0.992982
Large-poor – large-poor	0.023766	0.197441	0.54479	0.564899	0.605119	0.817185	0.837294	0.840951	0.85192	0.855576	0.86106
Middle-rich – large-poor	0.086301	0.430137	0.630137	0.739726	0.791781	0.861644	0.886301	0.90274	0.912329	0.936986	0.939726
Middle-poor – large-poor	0.089347	0.396907	0.501718	0.604811	0.627148	0.792096	0.82646	0.831615	0.840206	0.860825	0.876289
Middle-rich – middle-rich	0.056604	0.6	0.633962	0.701887	0.74717	0.788679	0.818868	0.849057	0.867925	0.898113	0.901887
Middle-poor – middle-rich	0.19084	0.641221	0.721374	0.748092	0.791985	0.902672	0.931298	0.935115	0.944656	0.948473	0.948473
Middle-poor – middle-poor	0.228155	0.76699	0.854369	0.898058	0.932039	0.985437	0.985437	0.985437	0.985437	0.985437	0.985437

A.6. QA approach accuracy for connection numbers with correctly predicted passenger numbers for 2012 from 2011 for every cluster pair at specified intervals

	0	5	10	25	50	100	150	200	300	400	500
Very small-rich – very small-rich	0.03352	0.341798	0.466734	0.590147	0.635856	0.691722	0.803454	0.808024	0.94261	0.947689	0.947689
Small-poor – very small-rich	0.089536	0.458468	0.640777	0.7411	0.817691	0.830636	0.909385	0.909385	0.919094	0.943905	0.943905
Large-rich – very small-rich	0.013052	0.07259	0.120281	0.170582	0.222088	0.249398	0.262048	0.276506	0.32249	0.326908	0.384538
Middle-middle – very small-rich	0.04994	0.3213	0.465704	0.608303	0.759326	0.842359	0.942238	0.968712	0.969916	0.994585	0.994585
Megacities – very small-rich	0.023672	0.233687	0.372686	0.578452	0.661305	0.756601	0.801821	0.826404	0.861608	0.889226	0.930804
Small-rich – very small-rich	0.028939	0.203714	0.256238	0.391022	0.480916	0.518172	0.54301	0.573317	0.622878	0.643614	0.643956
Large-poor – very small-rich	0.03263	0.341011	0.46897	0.567818	0.78183	0.837172	0.892834	0.894114	0.894114	0.894114	0.894114
Middle-rich – very small-rich	0.008252	0.072607	0.099497	0.124253	0.139808	0.144835	0.145167	0.145167	0.157498	0.157688	0.157688
Middle-poor – very small-rich	0.084084	0.552553	0.676677	0.735736	0.855856	0.863864	0.886887	0.886887	0.886887	0.886887	0.886887
Small-poor – small-poor	0.048733	0.327485	0.483431	0.71345	0.732943	0.873294	0.900585	0.910331	0.925926	0.959064	0.959064
Large-rich – small-poor	0.00343	0.033813	0.055841	0.086048	0.107857	0.125929	0.134811	0.140395	0.178033	0.198435	0.203491
Middle-middle – small-poor	0.006826	0.034318	0.046568	0.085562	0.106882	0.163269	0.204601	0.222742	0.256031	0.260426	0.261174
Megacities – small-poor	0.008929	0.081101	0.139137	0.198413	0.250496	0.295015	0.310392	0.37438	0.376488	0.382688	0.390625
Small-rich – small-poor	0.02399	0.227662	0.306977	0.438433	0.522399	0.555936	0.680539	0.695226	0.696695	0.699633	0.714076
Large-poor – small-poor	0.000988	0.00798	0.012036	0.018436	0.02236	0.025626	0.028722	0.036664	0.037003	0.043148	0.046386
Middle-rich – small-poor	0.003576	0.044054	0.057471	0.076037	0.081942	0.083418	0.088207	0.088831	0.10392	0.114811	0.114811
Middle-poor – small-poor	0.003689	0.04829	0.056338	0.078974	0.083333	0.085681	0.086016	0.086016	0.096412	0.100268	0.117706
Large-rich – large-rich	0.002317	0.011068	0.014929	0.029086	0.04453	0.106049	0.107336	0.114286	0.146203	0.146203	0.148005
Middle-middle – large-rich	0.007222	0.031582	0.045379	0.060577	0.074266	0.112584	0.137968	0.166316	0.195635	0.200162	0.222851
Megacities – large-rich	0.000603	0.002392	0.003032	0.005142	0.007232	0.009078	0.01162	0.014935	0.015406	0.018099	0.029154
Small-rich – large-rich	0.00226	0.017617	0.033927	0.041333	0.049583	0.059304	0.06042	0.075532	0.087676	0.13745	0.137995
Large-poor – large-rich	0.001193	0.009299	0.013159	0.026949	0.030107	0.035546	0.046319	0.048284	0.052319	0.067935	0.069198
Middle-rich – large-rich	0.000125	0.005144	0.007855	0.012322	0.019699	0.023313	0.026475	0.028558	0.049161	0.0795	0.079651
Middle-poor – large-rich	0.00251	0.026538	0.028869	0.048951	0.103102	0.202259	0.32114	0.334947	0.34714	0.363098	0.363457
Middle-middle – middle-middle	0.00432	0.032793	0.056554	0.079136	0.166814	0.197447	0.249386	0.295042	0.383702	0.385567	0.439666
Megacities – middle-middle	0.002347	0.02532	0.036977	0.060291	0.072515	0.092537	0.125956	0.133752	0.13432	0.139921	0.140111
Small-rich – middle-middle	0.017988	0.177187	0.226196	0.30224	0.330569	0.369884	0.397243	0.399505	0.402413	0.404028	0.404459
Large-poor – middle-middle	0.003176	0.022886	0.035291	0.049567	0.058248	0.072065	0.084558	0.088052	0.104057	0.123626	0.132502
Middle-rich – middle-middle	0.006027	0.060273	0.08753	0.126842	0.16843	0.304313	0.319381	0.332708	0.392982	0.399411	0.463702
Middle-poor – middle-middle	0.014645	0.102891	0.161097	0.350732	0.353736	0.410815	0.410815	0.41119	0.418701	0.419452	0.419452
Megacities – megacities	0.000556	0.00584	0.015851	0.027253	0.042547	0.062291	0.062291	0.064516	0.064794	0.066463	0.066463
Small-rich – megacities	0.003801	0.033399	0.046864	0.066192	0.082851	0.090858	0.092152	0.146415	0.148356	0.149246	0.14965
Large-poor – megacities	0.013173	0.116297	0.195709	0.26195	0.302221	0.391795	0.441475	0.560783	0.727512	0.731276	0.73805
Middle-rich – megacities	0.000341	0.0022	0.003009	0.003843	0.005828	0.009014	0.011352	0.011365	0.011403	0.011428	0.011732
Middle-poor – megacities	0.014127	0.125126	0.193071	0.302725	0.386478	0.452069	0.462159	0.467878	0.491423	0.505214	0.528422
Small-rich – small-rich	0.018091	0.187552	0.258091	0.368299	0.42722	0.481992	0.525809	0.554191	0.607469	0.710041	0.710041
Large-poor – small-rich	0.060228	0.333456	0.530297	0.620455	0.640837	0.690415	0.747888	0.754131	0.816012	0.816012	0.816012
Middle-rich – small-rich	0.001504	0.017172	0.027808	0.039595	0.044029	0.051924	0.058669	0.058702	0.064175	0.072225	0.07249
Middle-poor – small-rich	0.085938	0.498766	0.6875	0.773438	0.80551	0.863487	0.872533	0.875822	0.997944	0.997944	0.997944
Large-poor – large-poor	0.001354	0.02434	0.050353	0.055691	0.0631	0.088993	0.09222	0.092419	0.093096	0.093256	0.093614
Middle-rich – large-poor	0.035345	0.239871	0.360776	0.483405	0.529741	0.59181	0.606466	0.619828	0.688147	0.793966	0.795474
Middle-poor – large-poor	0.008945	0.060515	0.084611	0.125137	0.138919	0.188299	0.200986	0.201533	0.203724	0.207101	0.211847
Middle-rich – middle-rich	0.009076	0.104649	0.121504	0.148916	0.189294	0.318763	0.440822	0.446379	0.47231	0.516762	0.517318
Middle-poor – middle-rich	0.058849	0.377577	0.487543	0.564003	0.637027	0.728093	0.809278	0.812715	0.927835	0.929124	0.929124
Middle-poor – middle-poor	0.040482	0.347115	0.464255	0.596038	0.723514	0.980189	0.980189	0.980189	0.980189	0.980189	0.980189

A.7. Passenger numbers covered by connections from Tab. A.6 for 2012 from 2011 for every cluster pair at specified intervals

	Eliminated connection number			Added connection number			Remaining connection number			Passenger number on eliminated connections			Passenger number on added connections			Passenger number on remaining connections		
	2002	**2007**	**2011**	**2002**	**2007**	**2011**	**2002**	**2007**	**2011**	**2002**	**2007**	**2011**	**2002**	**2007**	**2011**	**2002**	**2007**	**2011**
Very small-rich – very small-rich	5,858	5,488	4,127	5,310	4,502	3,974	8,820	7,440	8,078	464,998	495,793	44,245	159,075	106,708	1,965	5,870,843	6,374,444	7,070,456
Small-poor – very small-rich	2,206	2,664	2,405	3,797	2,986	2,469	3,061	1,928	2,371	29,260	101,044	38,021	83,958	14,231	930	1,359,943	1,652,177	2,417,479
Large-rich – very small-rich	4,838	4,435	3,326	4,858	3,607	3,099	14,228	15,955	16,459	252,532	410,425	63,414	218,050	131,169	9,960	29,757,210	33,490,428	39,908,128
Middle-middle – very small-rich	2,017	2,460	2,243	3,877	3,159	2,495	2,487	2,378	3,144	37,398	37,637	24,933	138,109	52,031	1,662	1,164,324	1,538,538	2,593,158
Megacities – very small-rich	3,317	3,395	2,751	4,408	3,299	2,697	9,923	11,103	11,675	97,352	392,514	19,0304	306,335	182,466	3,295	23,408,097	29,992,688	37,593,129
Small-rich – very small-rich	10,815	10,069	7,467	9,491	7,820	6,791	20,519	19,726	20,596	361,708	793,974	156,797	446,527	251,540	8,777	10,489,860	11,192,738	19,063,516
Large-poor – very small-rich	3,046	3,744	3,733	6,187	5,001	3,855	5,530	5,159	6,159	17,062	23,104	13,055	131,243	31,188	3,124	1,917,222	2,619,991	5,452,590
Middle-rich – very small-rich	7,024	6,527	4,798	6,027	5,054	4,302	17,312	17,861	18,474	307,413	630,283	103,240	1,166,903	433,439	21,088	21,762,101	25,829,770	29,749,010
Middle-poor – very small-rich	1,108	1,454	1,595	2,558	2,219	1,680	1,586	1,437	1,878	7,377	8,857	5,152	54,696	19,162	995	512,363	649,474	1,355,117
Small-poor – small-poor	974	1,207	961	2,153	1,765	1,302	1,468	1,441	1,842	285,089	352,163	45,100	306,251	197,267	517	3,524,332	3,846,206	5,375,393
Large-rich – small-poor	2,143	2,473	2,195	4,641	3,343	2,406	6,474	8,135	9,195	38,535	84,073	11,203	412,908	73,305	22,741	7,673,400	13,620,975	16,307,750
Middle-middle – small-poor	1,475	1,708	1,775	4,326	3,523	2,358	2,455	2,921	4,132	179,229	331,235	76,673	455,510	286,672	10,651	3,784,559	4,661,875	6,214,146
Megacities – small-poor	1,734	2,087	1,908	4,605	3,165	2,120	5,752	7,658	8,791	118,592	239,166	46,434	908,648	130,116	8109	25,282,651	51,532,546	71,342,640
Small-rich – small-poor	3,352	4,157	3,511	6,012	4,506	3,634	5,939	5,684	6,508	63,076	283,385	104,339	334,399	185,720	4039	1,340,105	2,118,546	5,255,400
Large-poor – small-poor	1,742	2,590	2,756	6,512	4,896	3,263	4,793	6,023	7,594	217,006	469,988	286,422	1,927,140	879,528	106,307	12,858,695	22,884,242	40,327,028
Middle-rich – small-poor	2,910	3,388	3,012	5,169	3,917	3,096	7,042	7,740	8,681	61,354	140,951	75,019	633,298	211,771	30,476	4,041,790	8,382,570	10,957,921
Middle-poor – small-poor	739	1,079	1,488	2,971	2,496	1,673	1,546	1,978	2,755	119,323	330,985	37,334	650,916	376,506	5968	2,778,311	3,633,252	6,670,049
Large-rich – large-rich	275	332	120	1,108	225	247	2,531	3,874	3,950	26,715	13,473	793	88,092	47,925	3889	94,996,756	123,973,882	128,618,098
Middle-middle – large-rich	1,617	1,978	1,825	4,912	3,436	2,255	7,387	9,993	11,402	123,007	271,373	47,592	505,476	124,582	18,548	8,326,378	13,674,000	18,735,312
Megacities – large-rich	251	304	159	1,747	307	289	4,011	6,077	6,164	23,606	26,513	799	242,633	128,315	53,102	187,809,030	242,928,415	263,508,761
Small-rich – large-rich	3,155	2,988	1,879	3,770	2,071	1,898	15,766	19,314	19,479	206,415	974,911	273,970	1,026,153	553,761	36,729	61,668,951	73,920,267	81,428,682
Large-poor – large-rich	818	1,255	1,051	3,452	1,817	1,182	10,114	13,156	13,915	32,224	152,604	6757	347,085	178,755	28,491	15,376,026	28,006,064	33,166,788
Middle-rich – large-rich	1,252	1,289	763	2,968	876	763	8,758	12,028	12,207	153,273	572,977	382,976	525,216	137,626	39,856	145,308,136	178,848,134	178,815,881
Middle-poor – large-rich	937	1,217	1,324	2,851	2,165	1,399	4,909	6,572	7,402	25,082	19,523	6537	212,292	73,189	5,576	4,471,294	7,611,761	12,493,632
Middle-middle – middle-middle	891	1,008	1,155	3,384	2,845	1,784	2,361	3,344	4,586	105,417	237,873	163,453	386,871	363,440	10,185	5,126,978	6,615,691	9,493,009
Megacities – middle-middle	1,092	1,411	1,338	4,406	2,795	1,736	5,719	8,252	9,484	105,115	353,305	9,723	933,830	752,653	26,422	38,108,288	58,841,280	90,981,070
Small-rich – middle-middle	3,558	4,467	3,926	6,768	5,250	4,113	5,578	7,045	8,342	71,082	119,299	44,793	257,871	151,540	9,285	1,208,787	1,882,533	3,863,760
Large-poor – middle-middle	1,571	2,307	2,623	7,860	5,791	3,654	6,135	8,740	11,140	149,535	362,756	128,271	2,619,430	1,025,723	56,671	19,016,798	31,505,973	49,615,343
Middle-rich – middle-middle	2,723	3,183	2,873	6,050	4,564	3,311	7,980	10,074	11,447	149,574	211,497	34,706	862,716	476,152	14,930	5,769,325	8,754,368	12,533,786
Middle-poor – middle-middle	556	838	1,089	2,757	2,292	1,399	1,199	1,894	2,902	6,348	43,274	14,445	547,517	90,457	2,665	710,397	1,126,723	1,657,543
Megacities – megacities	27	40	46	873	117	80	1,381	2,319	2,389	470	991	31,900	192,318	22,224	2,796	129,615,659	194,750,666	244,150,376
Small-rich – megacities	1,980	2,037	1,492	3,519	1,889	1,477	12,544	15,304	15,635	109,392	540,899	142,411	443,884	266,891	2,5530	54,861,082	67,435,237	73,554,784
Large-poor – megacities	363	606	728	3,551	1,515	861	7,867	10,821	11,563	33,699	58,288	86,476	997,347	225,690	2,658	109,005,978	174,116,908	267,760,611
Middle-rich – megacities	651	702	492	2,081	637	428	7,409	9,556	9,740	111,386	264,038	23,417	197,364	23,616	79,101	149,322,819	176,193,418	184,887,424
Middle-poor – megacities	589	913	1,008	2,904	1,979	1,126	4,767	6,474	7,427	50,467	127,345	15,885	2,275,956	97,978	2,970	20,898,136	48,770,883	77,308,097
Small-rich – small-rich	5,551	5,042	3,336	4,447	3,536	3,044	15,004	16,336	16,460	220,930	453,551	138,227	324,330	80,639	6,027	10,992,063	12,500,499	13,333,699
Large-poor – small-rich	4,238	5,289	4,770	9,160	6,845	4,850	12,837	15,242	17,094	32,843	75,475	38,696	348,866	253,735	5,448	2,112,919	3,414,100	6,740,181
Middle-rich – small-rich	5,868	5,459	3,569	5,489	3,851	3,440	21,825	24,916	25,090	31,2057	819,073	119,119	2,019,010	543,356	89,514	46,460,578	56,537,071	53,186,805
Middle-poor – small-rich	1,961	2,553	2,473	4,607	3,785	2,788	3,870	4,907	5,855	14,711	71,163	33,851	148,395	19,485	3,348	632,702	1,156,954	2,465,211
Large-poor – large-poor	565	1,239	1,425	4,746	2,712	1,698	5,252	7,681	8,712	57,253	81,331	55,716	2,991,831	820,831	25,113	55,306,014	99,256,218	190,716,932
Middle-rich – large-poor	1,905	2,529	2,400	5,224	3,428	2,370	13,449	16,212	17,241	34,959	126,992	15,831	181,685	134,728	4,639	8,785,416	13,477,511	17,294,454
Middle-poor – large-poor	780	1,417	2,054	5,147	3,934	2,371	3,881	5,687	7,294	29,650	101,915	15,491	1,516,312	309,936	10,955	7,809,520	14,468,320	22,751,402
Middle-rich – middle-rich	1,363	1,301	910	2,080	929	817	6,704	8,422	8,516	126,844	541,920	42,570	776,149	272,096	5,401	56,227,179	68,414,091	65,397,893
Middle-poor – middle-rich	1,522	1,901	1,967	3,638	2,930	2,005	5,223	6,484	7,433	12,050	110,567	7,838	33,761	54,645	2,327	1,752,102	3,022,312	4,453,630
Middle-poor – middle-poor	307	520	762	1,863	1,561	851	1,090	1,679	2,410	102,598	263,140	19,892	918,864	224,480	1,161	2,976,852	5,638,933	7,759,394

A.8. Connection and passenger numbers for eliminated, added and remaining connections in 2012 from 2002, 2007 and 2011 for every cluster pair

	0%	5%	10%	15%	20%	30%	50%	75%	100%	125%	150%
Very small-rich – very small-rich	0.038149	0.05153	0.06847	0.092527	0.120427	0.165409	0.305907	0.468897	0.679858	0.69694	0.719146
Small-poor – very small-rich	0.047861	0.061276	0.076867	0.097534	0.126178	0.176215	0.3314	0.518129	0.791516	0.806019	0.826686
Large-rich – very small-rich	0.026316	0.048893	0.079387	0.110541	0.145873	0.211992	0.365782	0.547647	0.713972	0.736476	0.762718
Middle-middle – very small-rich	0.053357	0.061805	0.076034	0.095598	0.11783	0.166741	0.306803	0.493108	0.755892	0.769675	0.787461
Megacities – very small-rich	0.032206	0.051829	0.079236	0.107817	0.142796	0.204863	0.360456	0.538019	0.706196	0.725072	0.748107
Small-rich – very small-rich	0.035965	0.05188	0.076858	0.101665	0.136101	0.192173	0.343623	0.513423	0.704406	0.725589	0.752209
Large-poor – very small-rich	0.055336	0.066403	0.085771	0.107312	0.138538	0.194664	0.351383	0.533794	0.760277	0.773913	0.794862
Middle-rich – very small-rich	0.029498	0.049163	0.076545	0.105732	0.139648	0.201755	0.351795	0.51758	0.687722	0.714917	0.741801
Middle-poor – very small-rich	0.047912	0.058179	0.0705	0.090349	0.110198	0.157426	0.325804	0.505818	0.774812	0.788501	0.814511
Small-poor – small-poor	0.034143	0.047041	0.062215	0.079666	0.10698	0.157056	0.286798	0.450683	0.707891	0.724583	0.742033
Large-rich – small-poor	0.026838	0.044369	0.067794	0.095098	0.127676	0.185541	0.327025	0.532578	0.783897	0.800962	0.818492
Middle-middle – small-poor	0.030985	0.041596	0.056876	0.078947	0.103565	0.151528	0.264431	0.435908	0.714771	0.735993	0.758065
Megacities – small-poor	0.022334	0.039127	0.060076	0.08241	0.111669	0.16759	0.297265	0.481821	0.764197	0.779778	0.79259
Small-rich – small-poor	0.041758	0.053611	0.074398	0.094639	0.118162	0.16849	0.324581	0.521882	0.791758	0.803428	0.824398
Large-poor – small-poor	0.025534	0.037882	0.0563	0.077438	0.100251	0.146923	0.266848	0.44391	0.76936	0.780452	0.796568
Middle-rich – small-poor	0.034143	0.050625	0.074761	0.098308	0.130096	0.188079	0.356439	0.542016	0.773804	0.788079	0.809566
Middle-poor – small-poor	0.037185	0.049137	0.070385	0.091633	0.116202	0.159363	0.297477	0.466135	0.770252	0.781541	0.799469
Large-rich – large-rich	0.005359	0.058008	0.111602	0.166456	0.223518	0.337011	0.521122	0.701765	0.854666	0.87232	0.887768
Middle-middle – large-rich	0.027523	0.04327	0.070656	0.099137	0.127071	0.182391	0.326852	0.509927	0.778173	0.797754	0.819252
Megacities – large-rich	0.003358	0.049575	0.097768	0.148924	0.205807	0.313253	0.498914	0.702943	0.861742	0.879913	0.892159
Small-rich – large-rich	0.018377	0.050935	0.093921	0.133728	0.17735	0.258743	0.424329	0.604349	0.757345	0.783225	0.806054
Large-poor – large-rich	0.012524	0.038058	0.070583	0.105534	0.139029	0.205437	0.37301	0.601748	0.861748	0.874854	0.889515
Middle-rich – large-rich	0.009124	0.062262	0.114197	0.164628	0.217265	0.314919	0.490977	0.661119	0.800281	0.824845	0.845699
Middle-poor – large-rich	0.028701	0.047285	0.075573	0.105513	0.139376	0.205038	0.364237	0.560809	0.818501	0.832748	0.850093
Middle-middle – middle-middle	0.020355	0.04071	0.061932	0.080987	0.10524	0.157211	0.28194	0.422261	0.7055	0.726288	0.744911
Megacities – middle-middle	0.017763	0.037062	0.065756	0.094962	0.122801	0.181042	0.323997	0.515969	0.792485	0.81076	0.827327
Small-rich – middle-middle	0.05309	0.063788	0.081418	0.103407	0.129358	0.178487	0.329834	0.50832	0.742472	0.755349	0.779319
Large-poor – middle-middle	0.026337	0.04014	0.056798	0.076313	0.103919	0.155323	0.287799	0.464382	0.771696	0.785023	0.79835
Middle-rich – middle-middle	0.033163	0.044305	0.06685	0.092935	0.121117	0.176825	0.323765	0.494953	0.73509	0.755145	0.779788
Middle-poor – middle-middle	0.035684	0.04503	0.06627	0.086661	0.124044	0.164826	0.320306	0.503823	0.765506	0.783347	0.804588
Megacities – megacities	0.001488	0.042163	0.085813	0.129464	0.182044	0.275298	0.459821	0.688988	0.897321	0.904762	0.91121
Small-rich – megacities	0.01612	0.047393	0.083179	0.124446	0.16402	0.240348	0.409849	0.59144	0.75401	0.777223	0.798662
Large-poor – megacities	0.008793	0.034016	0.06109	0.088627	0.119866	0.186741	0.34097	0.56728	0.865093	0.876432	0.885919
Middle-rich – megacities	0.008807	0.053585	0.105557	0.157529	0.206773	0.302158	0.485487	0.66708	0.803151	0.829695	0.851278
Middle-poor – megacities	0.015529	0.032489	0.05517	0.078872	0.104618	0.154271	0.293625	0.502043	0.807928	0.823662	0.835309
Small-rich – small-rich	0.028015	0.054062	0.083743	0.116756	0.153631	0.217309	0.368441	0.543803	0.713561	0.739153	0.765503
Large-poor – small-rich	0.042779	0.057578	0.078755	0.103759	0.136928	0.19238	0.354312	0.55341	0.76314	0.779554	0.803028
Middle-rich – small-rich	0.022511	0.050529	0.087839	0.125629	0.167106	0.241295	0.402941	0.565592	0.713348	0.741032	0.769721
Middle-poor – small-rich	0.044482	0.055602	0.071726	0.093689	0.127884	0.175702	0.338338	0.530998	0.760634	0.77787	0.801223
Large-poor – large-poor	0.015414	0.035158	0.055248	0.076377	0.101489	0.15362	0.286283	0.497402	0.838067	0.849151	0.860755
Middle-rich – large-poor	0.02693	0.051029	0.084385	0.116747	0.153393	0.222095	0.390177	0.589473	0.805065	0.824573	0.843317
Middle-poor – large-poor	0.027699	0.038984	0.057964	0.080021	0.106438	0.149013	0.286997	0.469351	0.801744	0.816363	0.833547
Middle-rich – middle-rich	0.014961	0.054905	0.099365	0.144954	0.192096	0.280875	0.444884	0.60127	0.730134	0.764432	0.793649
Middle-poor – middle-rich	0.034731	0.047705	0.071058	0.097605	0.13493	0.198802	0.356886	0.549501	0.75988	0.777645	0.799601
Middle-poor – middle-poor	0.027498	0.038497	0.055912	0.067828	0.087076	0.122823	0.234647	0.406966	0.759853	0.770852	0.786434

A.9. Correlation approach accuracy for remaining connections with correctly predicted passenger numbers for 2012 from 2002 for every cluster pair at specified percentage intervals

	0%	5%	10%	15%	20%	30%	50%	75%	100%	125%	150%
Very small-rich – very small-rich	0.047705	0.063906	0.089237	0.114697	0.146329	0.198534	0.351164	0.51588	0.687412	0.710428	0.737945
Small-poor – very small-rich	0.055756	0.072634	0.099458	0.125377	0.155214	0.211875	0.389391	0.553948	0.742315	0.761001	0.787824
Large-rich – very small-rich	0.032672	0.057367	0.091565	0.126665	0.168008	0.24563	0.413776	0.574917	0.700749	0.732519	0.765191
Middle-middle – very small-rich	0.058633	0.069065	0.086331	0.103957	0.131295	0.18705	0.355755	0.521223	0.727698	0.745683	0.766906
Megacities – very small-rich	0.034538	0.060341	0.091108	0.128126	0.163061	0.228166	0.386463	0.545355	0.674375	0.702858	0.733128
Small-rich – very small-rich	0.041074	0.06135	0.090406	0.121865	0.16064	0.224916	0.393917	0.553878	0.70579	0.729045	0.757107
Large-poor – very small-rich	0.065221	0.080298	0.10249	0.128579	0.157716	0.212604	0.382179	0.564289	0.741996	0.760461	0.785702
Middle-rich – very small-rich	0.032078	0.056895	0.09409	0.128787	0.168006	0.239303	0.404511	0.564066	0.697554	0.725347	0.753437
Middle-poor – very small-rich	0.068908	0.081793	0.103081	0.129412	0.164146	0.215686	0.398319	0.558543	0.755182	0.77479	0.8
Small-poor – small-poor	0.044347	0.066209	0.089944	0.1193	0.153654	0.202998	0.353529	0.507808	0.717052	0.738289	0.755153
Large-rich – small-poor	0.033002	0.060372	0.095862	0.134494	0.175877	0.250917	0.418937	0.589314	0.740833	0.759429	0.785228
Middle-middle – small-poor	0.035088	0.051332	0.071475	0.094217	0.120533	0.179012	0.31514	0.491878	0.740741	0.760234	0.778103
Megacities – small-poor	0.025233	0.05147	0.081147	0.113692	0.151541	0.224086	0.378781	0.549964	0.725448	0.745663	0.769462
Small-rich – small-poor	0.041812	0.060666	0.089707	0.118899	0.150525	0.213927	0.389843	0.57549	0.748366	0.767675	0.78957
Large-poor – small-poor	0.032378	0.050834	0.074146	0.099401	0.129027	0.184879	0.332848	0.512708	0.75603	0.771248	0.789704
Middle-rich – small-poor	0.036587	0.058846	0.089037	0.124216	0.162083	0.239094	0.410132	0.575029	0.729308	0.754126	0.779327
Middle-poor – small-poor	0.055416	0.069521	0.090176	0.112846	0.144584	0.197985	0.356675	0.532494	0.745592	0.764232	0.782872
Large-rich – large-rich	0.006849	0.080405	0.160512	0.240322	0.312388	0.435974	0.60274	0.731983	0.82698	0.851995	0.872841
Middle-middle – large-rich	0.03049	0.051159	0.084732	0.121046	0.158273	0.230102	0.38198	0.566861	0.754825	0.77892	0.803814
Megacities – large-rich	0.004641	0.082421	0.154446	0.220902	0.294041	0.423612	0.608873	0.751253	0.844812	0.868758	0.884722
Small-rich – large-rich	0.021848	0.060293	0.110266	0.157644	0.205685	0.297302	0.470155	0.628402	0.742531	0.774579	0.802885
Large-poor – large-rich	0.020399	0.062322	0.109517	0.157403	0.205636	0.302706	0.490016	0.678451	0.815887	0.837497	0.855217
Middle-rich – large-rich	0.01252	0.072446	0.131893	0.196789	0.255758	0.357354	0.532161	0.678582	0.781325	0.815732	0.842015
Middle-poor – large-rich	0.031618	0.062533	0.103987	0.145442	0.187072	0.265414	0.436852	0.615317	0.788688	0.807483	0.830494
Middle-middle – middle-middle	0.02765	0.043956	0.063098	0.081886	0.113435	0.160936	0.28536	0.444878	0.726338	0.749025	0.771712
Megacities – middle-middle	0.024695	0.04939	0.081437	0.11404	0.152469	0.212819	0.380827	0.565899	0.762209	0.784545	0.802719
Small-rich – middle-middle	0.056829	0.068768	0.087711	0.111429	0.144381	0.195002	0.350207	0.51369	0.694365	0.718083	0.743075
Large-poor – middle-middle	0.028705	0.046692	0.070346	0.094739	0.124923	0.180855	0.322533	0.502156	0.767895	0.783171	0.801897
Middle-rich – middle-middle	0.046304	0.064422	0.090706	0.119673	0.156135	0.217873	0.381054	0.549267	0.725422	0.748238	0.776647
Middle-poor – middle-middle	0.052907	0.064535	0.085465	0.109302	0.137209	0.196512	0.363953	0.526744	0.772674	0.785465	0.80814
Megacities – megacities	0.003742	0.079981	0.158559	0.230589	0.307764	0.430309	0.616464	0.750702	0.847989	0.862488	0.869504
Small-rich – megacities	0.019997	0.05787	0.100136	0.141418	0.185881	0.272459	0.44516	0.605969	0.717467	0.754658	0.783669
Large-poor – megacities	0.012948	0.052701	0.098017	0.142424	0.192393	0.280801	0.464799	0.64212	0.795873	0.816812	0.832592
Middle-rich – megacities	0.008382	0.064337	0.126667	0.18404	0.24212	0.348955	0.527447	0.673238	0.777712	0.81112	0.839334
Middle-poor – megacities	0.02431	0.054871	0.088557	0.12589	0.164265	0.240493	0.404931	0.578052	0.773398	0.791283	0.810557
Small-rich – small-rich	0.030493	0.057644	0.093397	0.133343	0.175066	0.251972	0.423058	0.583197	0.710854	0.742555	0.771839
Large-poor – small-rich	0.05276	0.07057	0.097358	0.125705	0.161027	0.226031	0.393292	0.568789	0.728183	0.74911	0.777085
Middle-rich – small-rich	0.024134	0.060176	0.104369	0.149524	0.19619	0.280866	0.448434	0.599606	0.7146	0.746794	0.778348
Middle-poor – small-rich	0.060864	0.074066	0.097113	0.121056	0.153726	0.20631	0.375923	0.556053	0.737078	0.754307	0.780264
Large-poor – large-poor	0.025228	0.052024	0.083523	0.118729	0.156357	0.230046	0.407355	0.58951	0.783922	0.800171	0.820838
Middle-rich – large-poor	0.028362	0.058384	0.100996	0.139043	0.18539	0.265634	0.445974	0.62334	0.770753	0.798008	0.825609
Middle-poor – large-poor	0.03687	0.053736	0.082761	0.108257	0.142577	0.206903	0.36164	0.53285	0.753481	0.771524	0.795646
Middle-rich – middle-rich	0.020367	0.064473	0.115187	0.162395	0.214459	0.304424	0.477745	0.618425	0.727677	0.761802	0.793229
Middle-poor – middle-rich	0.041166	0.061921	0.099314	0.129503	0.173413	0.249914	0.426587	0.591938	0.751801	0.771184	0.794683
Middle-poor – middle-poor	0.040816	0.056616	0.076366	0.096774	0.130349	0.181698	0.312047	0.493088	0.743252	0.752469	0.76761

A.10. Passenger numbers covered by connections from Tab. A.9 for 2012 from 2002 for every cluster pair at specified percentage intervals

	0%	5%	10%	15%	20%	30%	50%	75%	100%	125%	150%
Very small-rich – very small-rich	0.047705	0.063906	0.089237	0.114697	0.146329	0.198534	0.351164	0.51588	0.687412	0.710428	0.737945
Small-poor – very small-rich	0.055756	0.072634	0.099458	0.125377	0.155214	0.211875	0.389391	0.553948	0.742315	0.761001	0.787824
Large-rich – very small-rich	0.032672	0.057367	0.091565	0.126665	0.168008	0.24563	0.413776	0.574917	0.700749	0.732519	0.765191
Middle-middle – very small-rich	0.058633	0.069065	0.086331	0.103957	0.131295	0.18705	0.355755	0.521223	0.727698	0.745683	0.766906
Megacities – very small-rich	0.034538	0.060341	0.091108	0.128126	0.163061	0.228166	0.386463	0.545355	0.674375	0.702858	0.733128
Small-rich – very small-rich	0.041074	0.06135	0.090406	0.121865	0.16064	0.224916	0.393917	0.553878	0.70579	0.729045	0.757107
Large-poor – very small-rich	0.065221	0.080298	0.10249	0.128579	0.157716	0.212604	0.382179	0.564289	0.741996	0.760461	0.785702
Middle-rich – very small-rich	0.032078	0.056895	0.09409	0.128787	0.168006	0.239303	0.404511	0.564066	0.697554	0.725347	0.753437
Middle-poor – very small-rich	0.068908	0.081793	0.103081	0.129412	0.164146	0.215686	0.398319	0.558543	0.755182	0.77479	0.8
Small-poor – small-poor	0.044347	0.066209	0.089944	0.1193	0.153654	0.202998	0.353529	0.507808	0.717052	0.738289	0.755153
Large-rich – small-poor	0.033002	0.060372	0.095862	0.134494	0.175877	0.250917	0.418937	0.589314	0.740833	0.759429	0.785228
Middle-middle – small-poor	0.035088	0.051332	0.071475	0.094217	0.120533	0.179012	0.31514	0.491878	0.740741	0.760234	0.778103
Megacities – small-poor	0.025233	0.05147	0.081147	0.113692	0.151541	0.224086	0.378781	0.549964	0.725448	0.745663	0.769462
Small-rich – small-poor	0.041812	0.060666	0.089707	0.118899	0.150525	0.213927	0.389843	0.57549	0.748366	0.767675	0.78957
Large-poor – small-poor	0.032378	0.050834	0.074146	0.099401	0.129027	0.184879	0.332848	0.512708	0.75603	0.771248	0.789704
Middle-rich – small-poor	0.036587	0.058846	0.089037	0.124216	0.162083	0.239094	0.410132	0.575029	0.729308	0.754126	0.779327
Middle-poor – small-poor	0.055416	0.069521	0.090176	0.112846	0.144584	0.197985	0.356675	0.532494	0.745592	0.764232	0.782872
Large-rich – large-rich	0.006849	0.080405	0.160512	0.240322	0.312388	0.435974	0.60274	0.731983	0.82698	0.851995	0.872841
Middle-middle – large-rich	0.03049	0.051159	0.084732	0.121046	0.158273	0.230102	0.38198	0.566861	0.754825	0.77892	0.803814
Megacities – large-rich	0.004641	0.082421	0.154446	0.220902	0.294041	0.423612	0.608873	0.751253	0.844812	0.868758	0.884722
Small-rich – large-rich	0.021848	0.060293	0.110266	0.157644	0.205685	0.297302	0.470155	0.628402	0.742531	0.774579	0.802885
Large-poor – large-rich	0.020399	0.062322	0.109517	0.157403	0.205636	0.302706	0.490016	0.678451	0.815887	0.837497	0.855217
Middle-rich – large-rich	0.01252	0.072446	0.131893	0.196789	0.255758	0.357354	0.532161	0.678582	0.781325	0.815732	0.842015
Middle-poor – large-rich	0.031618	0.062533	0.103987	0.145442	0.187072	0.265414	0.436852	0.615317	0.788688	0.807483	0.830494
Middle-middle – middle-middle	0.02765	0.043956	0.063098	0.081886	0.113435	0.160936	0.28536	0.444878	0.726338	0.749025	0.771712
Megacities – middle-middle	0.024695	0.04939	0.081437	0.11404	0.152469	0.212819	0.380827	0.565899	0.762209	0.784545	0.802719
Small-rich – middle-middle	0.056829	0.068768	0.087711	0.111429	0.144381	0.195002	0.350207	0.51369	0.694365	0.718083	0.743075
Large-poor – middle-middle	0.028705	0.046692	0.070346	0.094739	0.124923	0.180855	0.322533	0.502156	0.767895	0.783171	0.801897
Middle-rich – middle-middle	0.046304	0.064422	0.090706	0.119673	0.156135	0.217873	0.381054	0.549267	0.725422	0.748238	0.776647
Middle-poor – middle-middle	0.052907	0.064535	0.085465	0.109302	0.137209	0.196512	0.363953	0.526744	0.772674	0.785465	0.80814
Megacities – megacities	0.003742	0.079981	0.158559	0.230589	0.307764	0.430309	0.616464	0.750702	0.847989	0.862488	0.869504
Small-rich – megacities	0.019997	0.05787	0.100136	0.141418	0.185881	0.272459	0.44516	0.605969	0.717467	0.754658	0.783669
Large-poor – megacities	0.012948	0.052701	0.098017	0.142424	0.192393	0.280801	0.464799	0.64212	0.795873	0.816812	0.832592
Middle-rich – megacities	0.008382	0.064337	0.126667	0.18404	0.24212	0.348955	0.527447	0.673238	0.777712	0.81112	0.839334
Middle-poor – megacities	0.02431	0.054871	0.088557	0.12589	0.164265	0.240493	0.404931	0.578052	0.773398	0.791283	0.810557
Small-rich – small-rich	0.030493	0.057644	0.093397	0.133343	0.175066	0.251972	0.423058	0.583197	0.710854	0.742555	0.771839
Large-poor – small-rich	0.05276	0.07057	0.097358	0.125705	0.161027	0.226031	0.393292	0.568789	0.728183	0.74911	0.777085
Middle-rich – small-rich	0.024134	0.060176	0.104369	0.149524	0.19619	0.280866	0.448434	0.599606	0.7146	0.746794	0.778348
Middle-poor – small-rich	0.060864	0.074066	0.097113	0.121056	0.153726	0.20631	0.375923	0.556053	0.737078	0.754307	0.780264
Large-poor – large-poor	0.025228	0.052024	0.083523	0.118729	0.156357	0.230046	0.407355	0.58951	0.783922	0.800171	0.820838
Middle-rich – large-poor	0.028362	0.058384	0.100996	0.139043	0.18539	0.265634	0.445974	0.62334	0.770753	0.798008	0.825609
Middle-poor – large-poor	0.03687	0.053736	0.082761	0.108257	0.142577	0.206903	0.36164	0.53285	0.753481	0.771524	0.795646
Middle-rich – middle-rich	0.020367	0.064473	0.115187	0.162395	0.214459	0.304424	0.477745	0.618425	0.727677	0.761802	0.793229
Middle-poor – middle-rich	0.041166	0.061921	0.099314	0.129503	0.173413	0.249914	0.426587	0.591938	0.751801	0.771184	0.794683
Middle-poor – middle-poor	0.040816	0.056616	0.076366	0.096774	0.130349	0.181698	0.312047	0.493088	0.743252	0.752469	0.76761

A.11. Correlation approach accuracy for remaining connections with the correctly predicted passenger numbers for 2012 from 2007 for every cluster pair at specified percentage intervals

	0%	5%	10%	15%	20%	30%	50%	75%	100%	125%	150%
Very small-rich – very small-rich	0.000336	0.069756	0.164916	0.248858	0.289332	0.407079	0.542068	0.722664	0.937622	0.943642	0.949742
Small-poor – very small-rich	0.000387	0.019894	0.043066	0.067193	0.179599	0.278708	0.475244	0.856183	0.969476	0.973652	0.980519
Large-rich – very small-rich	0.000135	0.064246	0.104901	0.187853	0.252891	0.358679	0.627451	0.901093	0.970761	0.979014	0.984058
Middle-middle – very small-rich	0.000278	0.032404	0.052882	0.069683	0.122558	0.156457	0.530942	0.657794	0.962257	0.984987	0.987578
Megacities – very small-rich	8.63E-05	0.077954	0.14377	0.22812	0.307803	0.381522	0.672428	0.919039	0.97445	0.981247	0.98442
Small-rich – very small-rich	0.000452	0.052988	0.119755	0.178247	0.267262	0.367098	0.594239	0.797972	0.952726	0.957882	0.967289
Large-poor – very small-rich	0.000534	0.164027	0.246174	0.298854	0.402069	0.521743	0.697892	0.905323	0.962048	0.968491	0.971071
Middle-rich – very small-rich	0.000243	0.057104	0.123032	0.163006	0.222746	0.350805	0.585776	0.864849	0.968031	0.973613	0.978727
Middle-poor – very small-rich	0.000643	0.087246	0.111482	0.194966	0.253265	0.392899	0.68756	0.884926	0.973574	0.975388	0.97763
Small-poor – small-poor	4.22E-05	0.062234	0.084043	0.108679	0.242276	0.411657	0.719354	0.796085	0.950419	0.966193	0.974485
Large-rich – small-poor	0.000243	0.071795	0.122837	0.190788	0.23514	0.379568	0.612499	0.825619	0.970251	0.976588	0.98326
Middle-middle – small-poor	6.25E-05	0.061836	0.081291	0.097863	0.162181	0.256086	0.398143	0.65441	0.953688	0.961315	0.965822
Megacities – small-poor	4.94E-05	0.051207	0.173422	0.228678	0.291458	0.42765	0.594681	0.806694	0.969615	0.983837	0.985858
Small-rich – small-poor	0.000517	0.042923	0.077288	0.10484	0.243799	0.356652	0.522723	0.723455	0.984358	0.987493	0.990207
Large-poor – small-poor	3.58E-05	0.033967	0.071898	0.133161	0.17292	0.258782	0.505383	0.74841	0.97453	0.984178	0.986508
Middle-rich – small-poor	9.62E-05	0.029412	0.076553	0.197289	0.253682	0.35434	0.644459	0.805412	0.981673	0.984947	0.987826
Middle-poor – small-poor	6.62E-05	0.063264	0.124127	0.177195	0.191035	0.208794	0.476679	0.807577	0.957053	0.963709	0.967702
Large-rich – large-rich	2.69E-05	0.1276	0.259526	0.406155	0.519933	0.680733	0.887149	0.943436	0.971138	0.976386	0.981912
Middle-middle – large-rich	0.000199	0.041112	0.096136	0.141691	0.201658	0.269846	0.569302	0.821546	0.968536	0.975261	0.980394
Megacities – large-rich	1.51E-05	0.110088	0.211812	0.330895	0.464243	0.640969	0.852637	0.949186	0.971944	0.988653	0.991527
Small-rich – large-rich	0.000134	0.068483	0.137525	0.203494	0.253661	0.423419	0.638599	0.886409	0.966461	0.973934	0.980076
Large-poor – large-rich	0.000152	0.086831	0.160807	0.242529	0.314754	0.451949	0.683911	0.855324	0.952115	0.966943	0.978541
Middle-rich – large-rich	2.42E-05	0.13232	0.243036	0.348958	0.452081	0.616771	0.786084	0.908843	0.964192	0.972869	0.977344
Middle-poor – large-rich	0.000196	0.047077	0.121047	0.232699	0.297735	0.409249	0.594829	0.854175	0.9751	0.979671	0.982316
Middle-middle – middle-middle	3.82E-05	0.053944	0.103216	0.138262	0.299722	0.441951	0.628208	0.81049	0.961535	0.97018	0.97322
Megacities – middle-middle	3.27E-05	0.050902	0.104896	0.182826	0.229897	0.31176	0.57887	0.803236	0.979574	0.984458	0.98623
Small-rich – middle-middle	0.000222	0.027816	0.044436	0.063431	0.075844	0.091633	0.235636	0.686085	0.988327	0.990959	0.992094
Large-poor – middle-middle	3.43E-05	0.028597	0.059985	0.091347	0.117458	0.243465	0.520994	0.763027	0.971465	0.980302	0.984807
Middle-rich – middle-middle	0.000334	0.025915	0.071398	0.145615	0.184787	0.308168	0.437236	0.663241	0.961356	0.968671	0.973571
Middle-poor – middle-middle	0.000615	0.002528	0.075072	0.083378	0.197574	0.356013	0.413793	0.600181	0.957802	0.959422	0.960745
Megacities – megacities	5.62E-06	0.11746	0.257058	0.386961	0.537988	0.71859	0.900597	0.953896	0.988152	0.992881	0.99365
Small-rich – megacities	3.4E-05	0.064186	0.121882	0.186459	0.23883	0.360794	0.567548	0.904982	0.967689	0.974632	0.983267
Large-poor – megacities	3.68E-05	0.055733	0.095494	0.183902	0.238318	0.459809	0.797245	0.94323	0.983812	0.987605	0.990771
Middle-rich – megacities	9.99E-06	0.107757	0.225713	0.347337	0.422635	0.597516	0.74713	0.932004	0.965409	0.968936	0.973009
Middle-poor – megacities	1.89E-05	0.055396	0.159896	0.217478	0.266305	0.365362	0.688392	0.888538	0.976843	0.983594	0.987847
Small-rich – small-rich	0.000421	0.066765	0.110826	0.148234	0.203264	0.280247	0.57498	0.799176	0.930662	0.957423	0.966022
Large-poor – small-rich	0.000638	0.019434	0.08321	0.138027	0.170486	0.213151	0.418148	0.764387	0.968007	0.972419	0.975652
Middle-rich – small-rich	0.000121	0.079426	0.138946	0.197526	0.266793	0.373491	0.638055	0.856135	0.952566	0.964443	0.976779
Middle-poor – small-rich	0.000566	0.04738	0.203362	0.290017	0.312152	0.374912	0.488309	0.704209	0.948761	0.952711	0.956672
Large-poor – large-poor	1.48E-05	0.032032	0.08456	0.159471	0.263221	0.417552	0.685888	0.895872	0.982618	0.984857	0.987588
Middle-rich – large-poor	0.000486	0.072749	0.169588	0.286273	0.363276	0.48508	0.670913	0.814575	0.956803	0.966222	0.973761
Middle-poor – large-poor	4.01E-05	0.04588	0.078025	0.155572	0.223586	0.418818	0.607302	0.800031	0.965686	0.973204	0.97898
Middle-rich – middle-rich	5.04E-05	0.087421	0.198774	0.312219	0.387762	0.575833	0.808958	0.894182	0.963205	0.97786	0.983875
Middle-poor – middle-rich	0.00063	0.057224	0.116462	0.183356	0.269638	0.345034	0.684664	0.836237	0.947318	0.966504	0.970949
Middle-poor – middle-poor	3.39E-05	0.063627	0.112279	0.148062	0.187477	0.225428	0.476265	0.701192	0.954279	0.960927	0.961243

A.12. Passenger numbers covered by connections from Tab. A.11 for 2012 from 2007 for every cluster pair at specified percentage intervals

	0%	5%	10%	15%	20%	30%	50%	75%	100%	125%	150%
Very small-rich – very small-rich	0.067774	0.092765	0.127306	0.166165	0.210781	0.278556	0.484757	0.653801	0.803349	0.823629	0.846265
Small-poor – very small-rich	0.075581	0.100367	0.131579	0.169217	0.205936	0.282742	0.498776	0.685741	0.851591	0.863219	0.881885
Large-rich – very small-rich	0.039388	0.087436	0.144982	0.2019	0.263147	0.361827	0.56645	0.712899	0.81116	0.837209	0.860954
Middle-middle – very small-rich	0.082819	0.102434	0.131856	0.159099	0.199419	0.277152	0.494007	0.682528	0.848529	0.8587	0.876135
Megacities – very small-rich	0.049429	0.093598	0.151067	0.20536	0.263921	0.364268	0.572407	0.727146	0.832357	0.853598	0.875533
Small-rich – very small-rich	0.055388	0.085672	0.124465	0.166904	0.212621	0.296602	0.497225	0.660959	0.794091	0.813276	0.836689
Large-poor – very small-rich	0.082577	0.107804	0.142449	0.180962	0.230575	0.304911	0.514968	0.697275	0.830979	0.844097	0.863942
Middle-rich – very small-rich	0.044049	0.084845	0.13257	0.180838	0.236517	0.333896	0.538295	0.689846	0.801326	0.823441	0.847062
Middle-poor – very small-rich	0.097364	0.112964	0.144701	0.183432	0.229155	0.303389	0.536848	0.709521	0.845078	0.86014	0.886498
Small-poor – small-poor	0.06495	0.091265	0.12598	0.164054	0.204367	0.283315	0.473684	0.652856	0.820269	0.835386	0.854423
Large-rich – small-poor	0.044935	0.093552	0.152577	0.207794	0.273927	0.373318	0.577812	0.725438	0.834222	0.854278	0.874841
Middle-middle – small-poor	0.056815	0.087231	0.122812	0.153802	0.202869	0.27977	0.474892	0.671736	0.839598	0.856815	0.875753
Megacities – small-poor	0.045958	0.095702	0.153825	0.209516	0.273047	0.371992	0.572317	0.72452	0.831846	0.8505	0.871722
Small-rich – small-poor	0.064599	0.090439	0.128135	0.166743	0.213406	0.293206	0.513756	0.69813	0.832193	0.849825	0.872017
Large-poor – small-poor	0.054374	0.089096	0.132683	0.172281	0.220301	0.308806	0.50266	0.679669	0.827423	0.843528	0.863475
Middle-rich – small-poor	0.05334	0.092063	0.143608	0.198231	0.255417	0.352609	0.569945	0.72676	0.833184	0.851776	0.873317
Middle-poor – small-poor	0.079576	0.099912	0.139257	0.17374	0.214412	0.293988	0.473033	0.639699	0.787356	0.80504	0.830239
Large-rich – large-rich	0.016568	0.157396	0.287278	0.394675	0.481953	0.615089	0.756805	0.842604	0.905917	0.921598	0.935799
Middle-middle – large-rich	0.04248	0.087224	0.142426	0.202803	0.265229	0.363558	0.567116	0.719784	0.836334	0.858544	0.880431
Megacities – large-rich	0.009339	0.151987	0.292254	0.413111	0.508698	0.641824	0.770921	0.854605	0.903864	0.918147	0.93536
Small-rich – large-rich	0.028209	0.098761	0.178034	0.256466	0.328943	0.446169	0.63924	0.770239	0.849814	0.873271	0.895525
Large-poor – large-rich	0.028056	0.108634	0.197896	0.279058	0.356296	0.48163	0.6665	0.792168	0.875167	0.893287	0.90982
Middle-rich – large-rich	0.018782	0.126731	0.233637	0.326978	0.408272	0.538987	0.711535	0.8155	0.885221	0.903434	0.919655
Middle-poor – large-rich	0.03972	0.089286	0.152036	0.218959	0.281876	0.389686	0.585781	0.725467	0.827603	0.853138	0.877837
Middle-middle – middle-middle	0.052413	0.100385	0.147172	0.187741	0.243115	0.328694	0.509328	0.676044	0.849275	0.864377	0.885105
Megacities – middle-middle	0.039492	0.101167	0.16874	0.232081	0.294781	0.402744	0.599051	0.75125	0.859726	0.878061	0.894858
Small-rich – middle-middle	0.084759	0.101711	0.132288	0.165082	0.205165	0.28026	0.489702	0.666033	0.805767	0.822877	0.847592
Large-poor – middle-middle	0.049601	0.086201	0.132088	0.183656	0.236207	0.325358	0.520704	0.683273	0.832295	0.849558	0.870862
Middle-rich – middle-middle	0.057915	0.093139	0.139384	0.189087	0.242464	0.3302	0.525878	0.689033	0.822366	0.841383	0.867855
Middle-poor – middle-middle	0.073425	0.099005	0.134533	0.175746	0.217432	0.296068	0.487447	0.657508	0.806253	0.82378	0.852676
Megacities – megacities	0.00509	0.15826	0.30634	0.433596	0.530773	0.658954	0.795002	0.856085	0.893568	0.908376	0.918556
Small-rich – megacities	0.030958	0.105318	0.184936	0.258184	0.333802	0.456673	0.651829	0.780033	0.861132	0.883721	0.902089
Large-poor – megacities	0.021	0.118089	0.217425	0.306212	0.379371	0.508791	0.689099	0.807775	0.882594	0.901153	0.915511
Middle-rich – megacities	0.018329	0.125533	0.243458	0.347089	0.438847	0.577061	0.741671	0.841499	0.899942	0.916311	0.929452
Middle-poor – megacities	0.032522	0.091637	0.156681	0.223967	0.28869	0.399712	0.596924	0.741269	0.833066	0.850048	0.869913
Small-rich – small-rich	0.041389	0.088047	0.142576	0.196894	0.250931	0.350573	0.553088	0.703183	0.808657	0.833181	0.856932
Large-poor – small-rich	0.069405	0.09597	0.136979	0.182638	0.23626	0.322858	0.525155	0.688064	0.813768	0.832793	0.858935
Middle-rich – small-rich	0.033505	0.093207	0.163894	0.231041	0.299752	0.413319	0.61297	0.752597	0.842908	0.864923	0.885881
Middle-poor – small-rich	0.080769	0.103846	0.138034	0.178205	0.228846	0.313248	0.514957	0.68547	0.813034	0.830128	0.860256
Large-poor – large-poor	0.039605	0.100413	0.164955	0.230697	0.293506	0.390585	0.578877	0.728097	0.838912	0.856914	0.87665
Middle-rich – large-poor	0.04231	0.097622	0.167621	0.231894	0.296167	0.40861	0.608233	0.750724	0.848009	0.868019	0.88823
Middle-poor – large-poor	0.056286	0.097668	0.144135	0.193056	0.243907	0.326846	0.52902	0.681045	0.810801	0.830615	0.852534
Middle-rich – middle-rich	0.026661	0.105562	0.194614	0.277845	0.358506	0.486669	0.673163	0.792259	0.868318	0.88713	0.90337
Middle-poor – middle-rich	0.057945	0.09849	0.151838	0.208634	0.267236	0.367039	0.567137	0.709291	0.815332	0.836671	0.859488
Middle-poor – middle-poor	0.052247	0.08255	0.12069	0.160397	0.215778	0.293626	0.483804	0.64838	0.783699	0.804075	0.828631

A.13. Correlation approach accuracy for remaining connections with correctly predicted passenger numbers for 2012 from 2011 for every cluster pair at specified percentage intervals

	0%	5%	10%	15%	20%	30%	50%	75%	100%	125%	150%
Very small-rich – very small-rich	0.000762	0.116038	0.200608	0.357908	0.44275	0.520677	0.796505	0.926992	0.986681	0.989529	0.992884
Small-poor – very small-rich	0.00051	0.08465	0.224105	0.381409	0.425653	0.548993	0.795255	0.882806	0.987859	0.989982	0.990644
Large-rich – very small-rich	0.000184	0.222991	0.329351	0.49626	0.585079	0.683824	0.88509	0.972699	0.989414	0.993117	0.994516
Middle-middle – very small-rich	0.000477	0.0636	0.106463	0.257903	0.269938	0.563587	0.699911	0.864507	0.977697	0.978288	0.982471
Megacities – very small-rich	0.000166	0.341332	0.449755	0.548557	0.628452	0.728186	0.937242	0.989041	0.994664	0.996473	0.99686
Small-rich – very small-rich	0.000635	0.122817	0.236572	0.308762	0.382251	0.595015	0.884361	0.962978	0.989845	0.992827	0.994695
Large-poor – very small-rich	0.000827	0.317316	0.518092	0.620133	0.662068	0.708912	0.908167	0.962708	0.991288	0.992063	0.992638
Middle-rich – very small-rich	0.000276	0.220967	0.344709	0.418365	0.504368	0.639169	0.822427	0.955101	0.990242	0.992404	0.99314
Middle-poor – very small-rich	0.001969	0.188717	0.377923	0.432969	0.521243	0.656377	0.860814	0.926763	0.990083	0.990864	0.991545
Small-poor – small-poor	0.000102	0.068845	0.165716	0.218821	0.252088	0.35911	0.62773	0.8927	0.973158	0.975807	0.978805
Large-rich – small-poor	0.00019	0.123613	0.21761	0.293465	0.350545	0.472332	0.835729	0.918171	0.982248	0.99442	0.995129
Middle-middle – small-poor	0.000131	0.057499	0.144488	0.281989	0.350964	0.42281	0.650027	0.884655	0.984159	0.9917	0.992272
Megacities – small-poor	3.78E-05	0.175839	0.378272	0.481005	0.56719	0.733638	0.936655	0.982839	0.995084	0.995357	0.996628
Small-rich – small-poor	0.000346	0.041576	0.077813	0.123791	0.284138	0.393618	0.62252	0.887112	0.986625	0.988718	0.989692
Large-poor – small-poor	5.52E-05	0.140315	0.287096	0.397354	0.464913	0.630101	0.815484	0.906592	0.960743	0.962662	0.967733
Middle-rich – small-poor	0.00022	0.114801	0.192055	0.26994	0.329913	0.423566	0.753418	0.900651	0.992682	0.995185	0.996027
Middle-poor – small-poor	0.000157	0.161125	0.257263	0.408547	0.451227	0.695244	0.866209	0.96503	0.995132	0.996942	0.997384
Large-rich – large-rich	1.02E-05	0.294525	0.536991	0.735616	0.799577	0.896157	0.950265	0.963185	0.978767	0.990954	0.995687
Middle-middle – large-rich	0.000293	0.179108	0.354476	0.504758	0.557994	0.697472	0.898508	0.948656	0.985229	0.990716	0.992028
Megacities – large-rich	4.94E-06	0.320593	0.566767	0.708272	0.797609	0.874507	0.921934	0.945723	0.972399	0.981845	0.990677
Small-rich – large-rich	0.00021	0.225399	0.421954	0.5669	0.650546	0.750194	0.911362	0.960519	0.973191	0.991757	0.993744
Large-poor – large-rich	0.000253	0.145476	0.30723	0.39084	0.501013	0.631459	0.756481	0.823991	0.879373	0.944972	0.964378
Middle-rich – large-rich	6.33E-05	0.243592	0.4978	0.641731	0.758704	0.87259	0.950664	0.972009	0.981769	0.987002	0.988523
Middle-poor – large-rich	0.000245	0.062441	0.126679	0.270284	0.35649	0.478642	0.598129	0.682239	0.787093	0.931908	0.985044
Middle-middle – middle-middle	0.000107	0.216629	0.28252	0.363272	0.44783	0.727482	0.801286	0.934334	0.983694	0.989146	0.990653
Megacities – middle-middle	7.29E-05	0.283151	0.490938	0.603097	0.720289	0.836826	0.945362	0.975833	0.992495	0.994214	0.996066
Small-rich – middle-middle	0.000343	0.044049	0.075158	0.153751	0.174786	0.252926	0.406467	0.953891	0.993252	0.995818	0.996726
Large-poor – middle-middle	6.78E-05	0.185286	0.293562	0.424804	0.543052	0.614622	0.833043	0.936399	0.980207	0.983232	0.985744
Middle-rich – middle-middle	0.000491	0.20052	0.314191	0.41768	0.500066	0.627154	0.818589	0.894694	0.97318	0.980564	0.984399
Middle-poor – middle-middle	0.000931	0.216277	0.293469	0.367001	0.544439	0.623166	0.7951	0.873197	0.972952	0.975594	0.980093
Megacities – megacities	8.23E-07	0.206481	0.467101	0.673386	0.806163	0.890191	0.968336	0.979562	0.984576	0.986945	0.987645
Small-rich – megacities	8.28E-05	0.15382	0.304207	0.419583	0.589607	0.706163	0.950269	0.978958	0.990974	0.993989	0.995812
Large-poor – megacities	2.71E-05	0.327063	0.511667	0.676437	0.790325	0.89082	0.963701	0.988616	0.993855	0.99652	0.997256
Middle-rich – megacities	4.42E-05	0.358604	0.563534	0.718324	0.818301	0.932924	0.982079	0.988539	0.993855	0.99532	0.996016
Middle-poor – megacities	4.68E-05	0.169287	0.374063	0.572933	0.682586	0.770365	0.915509	0.97876	0.990299	0.993452	0.993796
Small-rich – small-rich	0.00075	0.182045	0.250507	0.365262	0.441989	0.622569	0.889576	0.941093	0.978541	0.986767	0.990546
Large-poor – small-rich	0.000721	0.092444	0.140298	0.170621	0.216984	0.34434	0.644458	0.900215	0.979169	0.982963	0.983957
Middle-rich – small-rich	0.000197	0.151211	0.313357	0.423413	0.535644	0.700016	0.915797	0.96758	0.991916	0.99444	0.995302
Middle-poor – small-rich	0.001368	0.061002	0.109386	0.155085	0.191574	0.371948	0.774175	0.963978	0.991259	0.993129	0.994309
Large-poor – large-poor	2.07E-05	0.206711	0.430975	0.54682	0.669703	0.843257	0.934126	0.974668	0.992579	0.995337	0.99666
Middle-rich – large-poor	0.000692	0.145334	0.342369	0.459433	0.563428	0.71448	0.848681	0.941283	0.972939	0.987503	0.989231
Middle-poor – large-poor	0.000175	0.083417	0.281476	0.471094	0.565914	0.748192	0.88425	0.947369	0.985978	0.991234	0.995152
Middle-rich – middle-rich	7.33E-05	0.183689	0.381184	0.594965	0.699351	0.820504	0.94518	0.984903	0.993565	0.995184	0.995912
Middle-poor – middle-rich	0.00187	0.11114	0.286503	0.429675	0.514259	0.744353	0.899582	0.951469	0.982674	0.984162	0.98502
Middle-poor – middle-poor	9.45E-05	0.100388	0.296365	0.397945	0.540241	0.736375	0.882948	0.928584	0.984967	0.985452	0.995036

A.14. Passenger numbers covered by connections from Tab. A.13 for 2012 from 2011 for every cluster pair at specified percentage intervals

Appendix B

This appendix contains detailed validation results for SAM from 2002 to 2012 annually at 100 km intervals.

Interval		Years													
km	**km**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	mean	mean per 1000km
0 - 100	50	33%	27%	22%	20%	10%	14%	12%	12%	4%	-3%	-2%	0%	12%	
100 - 200	150	12%	4%	7%	3%	-1%	4%	-4%	9%	7%	-5%	-2%	1%	3%	
200 - 300	250	-7%	-11%	-8%	-9%	-14%	-12%	-15%	0%	-1%	-10%	-8%	-3%	-8%	
300 - 400	350	-13%	-17%	-16%	-14%	-18%	-17%	-19%	-3%	-4%	-14%	-11%	-7%	-13%	
400 - 500	450	-18%	-20%	-20%	-18%	-20%	-21%	-22%	-4%	-4%	-12%	-11%	-5%	-15%	**-9%**
500 - 600	550	-20%	-23%	-22%	-18%	-20%	-21%	-21%	-5%	-5%	-11%	-12%	-7%	-15%	
600 - 700	650	-21%	-23%	-22%	-19%	-20%	-20%	-21%	-4%	-4%	-12%	-12%	-5%	-15%	
700 - 800	750	-22%	-23%	-22%	-19%	-19%	-19%	-20%	-4%	-5%	-10%	-11%	-6%	-15%	
800 - 900	850	-20%	-21%	-20%	-19%	-18%	-18%	-19%	-3%	-2%	-8%	-10%	-5%	-14%	
900 – 1,000	950	-20%	-20%	-18%	-15%	-17%	-17%	-18%	-2%	-1%	-6%	-7%	-3%	-12%	
1,000 – 1,100	1,050	-19%	-20%	-16%	-15%	-15%	-15%	-16%	1%	1%	-6%	-7%	-2%	-11%	
1,100 – 1,200	1,150	-18%	-18%	-16%	-13%	-14%	-12%	-14%	2%	3%	-4%	-5%	-1%	-9%	
1,200 – 1,300	1,250	-17%	-16%	-13%	-11%	-11%	-12%	-13%	2%	3%	-3%	-4%	-2%	-8%	
1,300 – 1,400	1,350	-14%	-15%	-12%	-9%	-10%	-10%	-11%	5%	6%	-2%	-4%	1%	-6%	
1,400 – 1,500	1,450	-14%	-14%	-11%	-9%	-8%	-8%	-10%	4%	5%	-1%	-3%	1%	-6%	**-5%**
1,500 – 1,600	1,550	-12%	-13%	-10%	-7%	-7%	-7%	-8%	6%	8%	0%	-3%	1%	-4%	
1,600 – 1,700	1,650	-11%	-12%	-9%	-6%	-5%	-5%	-8%	8%	11%	2%	-1%	3%	-3%	
1,700 – 1,800	1,750	-11%	-11%	-9%	-5%	-5%	-5%	-7%	9%	9%	2%	-1%	3%	-3%	
1,800 – 1,900	1,850	-8%	-10%	-6%	-3%	-3%	-3%	-6%	9%	10%	3%	-1%	4%	-1%	
1,900 – 2,000	1,950	-8%	-8%	-5%	-2%	-3%	-3%	-4%	11%	12%	4%	0%	2%	0%	
2,000 – 2,100	2,050	-9%	-9%	-7%	-3%	-3%	-4%	-4%	9%	10%	3%	-1%	3%	-1%	
2,100 – 2,200	2,150	-8%	-8%	-6%	-2%	-2%	-1%	-4%	9%	13%	4%	0%	3%	0%	
2,200 – 2,300	2,250	-7%	-7%	-6%	-2%	-3%	-4%	-4%	10%	11%	4%	0%	3%	0%	
2,300 – 2,400	2,350	-6%	-7%	-5%	0%	0%	0%	-2%	11%	14%	6%	2%	4%	1%	
2,400 – 2,500	2,450	-4%	-6%	-4%	-1%	1%	2%	-1%	10%	12%	4%	2%	3%	2%	1%
2,500 – 2,600	2,550	-4%	-5%	-3%	2%	1%	0%	-1%	13%	13%	7%	2%	4%	3%	
2,600 – 2,700	2,650	-4%	-3%	-1%	1%	2%	1%	1%	14%	15%	8%	2%	4%	3%	
2,700 – 2,800	2,750	-5%	-5%	-3%	2%	1%	2%	1%	12%	16%	6%	1%	3%	3%	
2,800 – 2,900	2,850	-6%	-3%	-4%	0%	2%	0%	1%	14%	14%	5%	1%	4%	2%	
2,900 – 3,000	2,950	-7%	-5%	-5%	0%	2%	0%	0%	13%	15%	4%	1%	1%	1%	
3,000 – 3,100	3,050	-7%	-6%	-5%	-1%	0%	-1%	0%	8%	15%	4%	-3%	1%	1%	
3,100 – 3,200	3,150	-7%	-3%	-4%	0%	1%	0%	-1%	8%	13%	3%	-3%	-1%	1%	
3,200 – 3,300	3,250	-14%	-7%	-6%	-2%	-1%	-4%	-2%	8%	10%	2%	-3%	-4%	-2%	
3,300 – 3,400	3,350	-12%	-7%	-6%	-2%	-1%	-3%	-2%	7%	10%	3%	-2%	-2%	-1%	
3,400 – 3,500	3,450	-12%	-8%	-6%	-1%	-1%	-3%	-2%	7%	11%	1%	-5%	-2%	-2%	**-2%**
3,500 – 3,600	3,550	-11%	-10%	-7%	-4%	-2%	-4%	-3%	9%	8%	0%	-5%	-4%	-3%	
3,600 – 3,700	3,650	-10%	-8%	-7%	-2%	1%	-3%	-1%	8%	9%	1%	-4%	-5%	-2%	
3,700 – 3,800	3,750	-10%	-8%	-6%	-3%	-2%	-1%	-2%	10%	11%	0%	-6%	-4%	-2%	
3,800 – 3,900	3,850	-10%	-8%	-7%	-6%	-4%	-4%	-5%	7%	8%	-1%	-6%	-5%	-3%	
3,900 – 4,000	3,950	-14%	-8%	-10%	-3%	-4%	-5%	-1%	6%	9%	-2%	-7%	-5%	-4%	
4,000 – 4,100	4,050	-13%	-13%	-12%	-8%	-7%	-7%	-6%	3%	6%	-2%	-9%	-8%	-6%	
4,100 – 4,200	4,150	-12%	-14%	-13%	-9%	-9%	-8%	-6%	0%	3%	-6%	-10%	-10%	-8%	
4,200 – 4,300	4,250	-16%	-14%	-17%	-11%	-10%	-8%	-9%	-1%	2%	-7%	-11%	-10%	-9%	
4,300 – 4,400	4,350	-15%	-15%	-17%	-12%	-10%	-10%	-9%	-3%	3%	-4%	-11%	-11%	-10%	
4,400 – 4,500	4,450	-18%	-17%	-18%	-13%	-11%	-14%	-13%	-6%	-2%	-8%	-16%	-13%	-12%	**-11%**
4,500 – 4,600	4,550	-15%	-16%	-17%	-15%	-11%	-11%	-11%	-5%	-2%	-8%	-11%	-11%	-11%	
4,600 – 4,700	4,650	-17%	-20%	-17%	-16%	-13%	-13%	-12%	-9%	-3%	-9%	-12%	-12%	-13%	
4,700 – 4,800	4,750	-17%	-20%	-19%	-18%	-16%	-16%	-15%	-8%	-5%	-13%	-14%	-10%	-14%	
4,800 – 4,900	4,850	-18%	-22%	-21%	-18%	-14%	-15%	-12%	-8%	-5%	-11%	-16%	-12%	-14%	
4,900 – 5,000	4,950	-17%	-20%	-22%	-18%	-16%	-16%	-13%	-8%	-6%	-10%	-15%	-12%	-14%	
5,000 – 5,100	5,050	-17%	-18%	-20%	-18%	-14%	-16%	-13%	-8%	-6%	-13%	-16%	-12%	-14%	
5,100 – 5,200	5,150	-19%	-21%	-21%	-18%	-15%	-18%	-14%	-11%	-8%	-16%	-17%	-14%	-16%	**-14%**
5,200 – 5,300	5,250	-19%	-20%	-19%	-17%	-15%	-18%	-12%	-5%	-7%	-11%	-16%	-12%	-14%	

5,300 – 5,400	5,350	-17%	-19%	-20%	-16%	-15%	-17%	-13%	-4%	-3%	-10%	-19%	-11%	-14%	
5,400 – 5,500	5,450	-20%	-20%	-21%	-19%	-15%	-18%	-13%	-9%	-5%	-12%	-16%	-13%	-15%	
5,500 – 5,600	5,550	-19%	-18%	-20%	-18%	-15%	-16%	-14%	-6%	-5%	-12%	-14%	-11%	-14%	
5,600 – 5,700	5,650	-21%	-21%	-20%	-19%	-18%	-16%	-13%	-6%	-2%	-10%	-14%	-10%	-14%	
5,700 – 5,800	5,750	-20%	-20%	-20%	-19%	-17%	-17%	-14%	-4%	-2%	-7%	-13%	-11%	-14%	
5,800 – 5,900	5,850	-19%	-20%	-21%	-21%	-19%	-17%	-11%	-5%	-4%	-10%	-16%	-13%	-15%	
5,900 – 6,000	5,950	-21%	-17%	-19%	-19%	-17%	-19%	-12%	-3%	-3%	-11%	-17%	-10%	-14%	
6,000 – 6,100	6,050	-19%	-16%	-19%	-19%	-18%	-15%	-12%	-2%	-2%	-9%	-12%	-12%	-13%	
6,100 – 6,200	6,150	-19%	-17%	-19%	-18%	-17%	-19%	-13%	-3%	-2%	-8%	-14%	-11%	-13%	
6,200 – 6,300	6,250	-18%	-19%	-20%	-17%	-17%	-17%	-10%	1%	-2%	-9%	-14%	-10%	-13%	
6,300 – 6,400	6,350	-17%	-18%	-17%	-14%	-14%	-13%	-11%	0%	1%	-7%	-14%	-9%	-11%	
6,400 – 6,500	6,450	-14%	-16%	-15%	-13%	-14%	-14%	-12%	4%	0%	-7%	-15%	-8%	-10%	**-11%**
6,500 – 6,600	6,550	-16%	-16%	-16%	-15%	-14%	-15%	-13%	1%	2%	-9%	-11%	-10%	-11%	
6,600 – 6,700	6,650	-14%	-15%	-16%	-15%	-14%	-14%	-9%	3%	0%	-7%	-12%	-10%	-10%	
6,700 – 6,800	6,750	-15%	-15%	-16%	-15%	-15%	-14%	-11%	3%	1%	-9%	-13%	-8%	-10%	
6,800 – 6,900	6,850	-14%	-13%	-14%	-13%	-13%	-12%	-8%	6%	2%	-8%	-12%	-6%	-9%	
6,900 – 7,000	6,950	-11%	-13%	-12%	-12%	-11%	-10%	-7%	8%	6%	-4%	-9%	-5%	-7%	
7,000 – 7,100	7,050	-10%	-12%	-14%	-14%	-9%	-13%	-4%	7%	6%	-7%	-8%	-6%	-7%	
7,100 – 7,200	7,150	-10%	-10%	-13%	-12%	-11%	-12%	-7%	8%	3%	-5%	-11%	-5%	-7%	
7,200 – 7,300	7,250	-10%	-8%	-11%	-10%	-9%	-8%	-6%	8%	5%	-3%	-8%	-4%	-5%	
7,300 – 7,400	7,350	-9%	-10%	-12%	-11%	-8%	-10%	-6%	9%	5%	-4%	-9%	-5%	-6%	
7,400 – 7,500	7450	-8%	-9%	-11%	-9%	-8%	-9%	-6%	11%	6%	-4%	-8%	-4%	-5%	
7,500 – 7,600	7,550	-7%	-10%	-9%	-8%	-7%	-10%	-5%	11%	8%	-3%	-9%	-4%	-4%	**-5%**
7,600 – 7,700	7,650	-7%	-8%	-10%	-7%	-8%	-8%	-5%	9%	5%	-10%	-9%	-4%	-5%	
7,700 – 7,800	7,750	-4%	-4%	-4%	-5%	-3%	-7%	-4%	9%	8%	-4%	-11%	-4%	-3%	
7,800 – 7,900	7,850	-3%	-4%	-7%	-7%	-6%	-6%	-2%	10%	7%	-4%	-8%	-3%	-3%	
7,900 – 8,000	7,950	-3%	-4%	-6%	-5%	-4%	-5%	-2%	11%	7%	-1%	-8%	-3%	-2%	
8,000 – 8,100	8,050	-4%	-3%	-6%	-5%	-3%	-4%	-1%	12%	9%	-2%	-7%	-2%	-1%	
8,100 – 8,200	8,150	0%	-5%	-7%	-6%	-3%	-5%	0%	10%	9%	-1%	-9%	-3%	-2%	
8,200 – 8,300	8,250	-2%	-3%	-6%	-4%	-3%	-4%	-1%	13%	10%	-1%	-6%	0%	-1%	
8,300 – 8,400	8,350	-4%	-5%	-8%	-7%	-4%	-4%	-2%	14%	11%	-1%	-7%	0%	-1%	
8,400 – 8,500	8,450	0%	-3%	-6%	-5%	-6%	-5%	2%	14%	11%	1%	-9%	1%	0%	
8,500 – 8,600	8,550	-2%	-2%	-6%	-5%	-2%	-5%	0%	14%	12%	3%	-4%	3%	1%	**0%**
8,600 – 8,700	8,650	-2%	-3%	-6%	-4%	-3%	-3%	0%	14%	12%	1%	-5%	-1%	0%	
8,700 – 8,800	8,750	-3%	-3%	-5%	-7%	-4%	-3%	-2%	15%	12%	1%	-4%	4%	0%	
8,800 – 8,900	8,850	-4%	-3%	-6%	-4%	-1%	-4%	-1%	16%	13%	2%	-2%	3%	1%	
8,900 – 9,000	8,950	-2%	-2%	-5%	-4%	-2%	-3%	0%	13%	14%	1%	-4%	1%	0%	
9,000 – 9,100	9,050	-2%	-4%	-6%	-6%	-3%	-3%	1%	15%	13%	0%	-5%	3%	0%	
9,100 – 9,200	9,150	-1%	-3%	-5%	-6%	-6%	-4%	-1%	13%	12%	3%	-5%	1%	0%	
9,200 – 9,300	9,250	-1%	-2%	-5%	-5%	-2%	-3%	1%	16%	12%	2%	-2%	1%	1%	
9,300 – 9,400	9,350	-1%	-5%	-5%	-5%	-3%	-3%	-1%	15%	13%	5%	-4%	4%	1%	
9,400 – 9,500	9,450	1%	-3%	-6%	-6%	-4%	-3%	0%	14%	13%	-1%	-6%	4%	0%	
9,500 – 9,600	9,550	-1%	0%	-7%	-6%	-4%	-2%	0%	16%	15%	4%	-4%	2%	1%	**0%**
9,600 – 9,700	9,650	-2%	0%	-4%	-6%	-3%	-4%	4%	13%	11%	3%	-4%	2%	1%	
9,700 – 9,800	9,750	-3%	-3%	-4%	-5%	-3%	-3%	4%	15%	12%	4%	-7%	2%	1%	
9,800 – 9,900	9,850	-3%	-2%	-7%	-6%	-6%	-4%	0%	13%	11%	3%	-7%	1%	-1%	
9,900 - 10,000	9,950	-3%	-2%	-5%	-6%	-2%	-3%	1%	14%	14%	2%	-5%	1%	0%	
10,000 – 10,100	10,050	-2%	0%	-6%	-6%	-5%	-5%	1%	15%	12%	3%	-5%	0%	0%	
10,100 – 10,200	10,150	-2%	-3%	-5%	-6%	-3%	-3%	1%	15%	12%	1%	-5%	1%	0%	
10,200 – 10,300	10,250	0%	0%	-5%	-7%	-2%	-3%	2%	12%	13%	2%	-8%	1%	1%	
10,300 – 10,400	10,350	2%	-1%	-8%	-5%	-3%	0%	2%	18%	9%	4%	-7%	-1%	1%	
10,400 – 10,500	10,450	0%	0%	-9%	-6%	0%	-1%	1%	12%	11%	4%	-8%	0%	0%	**-1%**
10,500 – 10,600	10,550	-2%	-1%	-7%	-8%	-4%	-4%	2%	8%	9%	1%	-10%	-5%	-2%	
10,600 – 10,700	10,650	-2%	-2%	-11%	-7%	-5%	-4%	0%	12%	10%	0%	-11%	-3%	-2%	
10,700 – 10,800	10,750	-1%	-3%	-8%	-7%	-5%	-4%	-2%	7%	5%	-1%	-9%	-2%	-3%	
10,800 – 10,900	10,850	-5%	-7%	-9%	-8%	-7%	-5%	-6%	5%	5%	-2%	-10%	-4%	-4%	
10,900 – 11,000	10,950	-8%	-6%	-11%	-9%	-7%	-7%	-3%	3%	4%	-4%	-8%	-5%	-5%	
11,000 – 11,100	11,050	-3%	-7%	-9%	-8%	-7%	-4%	-2%	13%	9%	1%	-10%	-2%	-2%	
11,100 – 11,200	11,150	-6%	-8%	-11%	-10%	-5%	-9%	0%	8%	5%	-3%	-9%	-5%	-4%	**-5%**
11,200 – 11,300	11,250	-5%	-5%	-12%	-5%	-7%	-8%	-4%	3%	4%	0%	-11%	-5%	-5%	

11,300 – 11,400	11,350	-4%	-6%	-10%	-6%	-7%	-6%	-3%	8%	8%	-2%	-10%	-7%	-4%	
11,400 – 11,500	11,450	-9%	-9%	-12%	-10%	-6%	-5%	-7%	4%	5%	-4%	-14%	-8%	-6%	
11,500 – 11,600	11,550	-9%	-7%	-9%	-8%	-8%	-6%	-4%	4%	3%	-2%	-10%	-6%	-5%	
11,600 – 11,700	11,650	-9%	-9%	-13%	-13%	-9%	-9%	-5%	5%	4%	-4%	-13%	-5%	-7%	
11,700 – 11,800	11,750	-8%	-11%	-12%	-9%	-6%	-6%	-5%	6%	4%	-6%	-10%	-5%	-6%	
11,800 – 11,900	11,850	-2%	-9%	-11%	-8%	-7%	-7%	0%	5%	6%	-4%	-10%	-4%	-4%	
11,900 – 12,000	11,950	-3%	-8%	-10%	-7%	-7%	-4%	-1%	8%	3%	-1%	-12%	-5%	-4%	
12,000 – 12,100	12,050	-1%	-5%	-10%	-7%	-1%	-7%	0%	4%	3%	-3%	-10%	-5%	-3%	
12,100 – 12,200	12,150	-1%	-4%	-9%	-7%	0%	-4%	0%	8%	11%	-1%	-10%	-6%	-2%	
12,200 – 12,300	12,250	-7%	-9%	-10%	-8%	-2%	-5%	-1%	8%	9%	0%	-8%	-3%	-3%	
12,300 – 12,400	12,350	-5%	-6%	-7%	-5%	-6%	-4%	-4%	5%	4%	-1%	-9%	-4%	-3%	
12,400 – 12,500	12,450	-4%	-3%	-9%	-5%	-3%	-4%	1%	10%	13%	4%	-6%	-4%	-1%	**-2%**
12,500 – 12,600	12,550	-4%	-3%	-8%	-4%	-1%	-4%	2%	6%	9%	-1%	-9%	-1%	-1%	
12,600 – 12,700	12,650	1%	-3%	-5%	-6%	-5%	-2%	1%	12%	6%	4%	-6%	-2%	0%	
12,700 – 12,800	12,750	-4%	-3%	-8%	-8%	-4%	-1%	-1%	9%	5%	2%	-7%	-2%	-2%	
12,800 – 12,900	12,850	-5%	-6%	-10%	-8%	-5%	-5%	-2%	7%	5%	-2%	-10%	-6%	-4%	
12,900 – 13,000	12,950	-3%	-8%	-11%	-5%	-6%	-3%	-2%	9%	6%	2%	-11%	-3%	-3%	
13,000 – 13,100	13,050	-4%	-5%	-8%	-2%	0%	2%	3%	12%	9%	2%	-8%	-3%	0%	
13,100 – 13,200	13,150	-2%	-3%	-8%	-6%	1%	0%	0%	14%	9%	1%	-10%	-4%	-1%	
13,200 – 13,300	13,250	-2%	-4%	-9%	-4%	-2%	-3%	1%	15%	10%	2%	-8%	0%	0%	
13,300 – 13,400	13,350	-4%	-9%	-5%	-3%	0%	-2%	-1%	10%	7%	1%	-8%	-3%	-1%	
13,400 – 13,500	13,450	-4%	-6%	-5%	-5%	-2%	-4%	0%	14%	10%	4%	-5%	-2%	0%	
13,500 – 13,600	13,550	2%	-1%	-8%	-4%	-2%	3%	3%	19%	18%	8%	-1%	1%	3%	**2%**
13,600 – 13,700	13,650	4%	2%	0%	-2%	2%	4%	2%	17%	16%	9%	-1%	4%	5%	
13,700 – 13,800	13,750	0%	0%	-6%	0%	2%	5%	7%	17%	16%	7%	-4%	2%	4%	
13,800 – 13,900	13,850	2%	-1%	-5%	-1%	3%	4%	-4%	16%	14%	7%	-4%	0%	3%	
13,900 – 14,000	13,950	3%	1%	-3%	6%	3%	4%	6%	12%	16%	9%	-3%	2%	5%	
14,000 – 14,100	14,050	-2%	1%	-5%	-3%	2%	7%	9%	18%	16%	5%	-4%	5%	4%	
14,100 – 14,200	14,150	-1%	-3%	-2%	-3%	7%	2%	3%	16%	17%	10%	-5%	2%	4%	
14,200 – 14,300	14,250	0%	1%	-1%	-3%	1%	-1%	3%	14%	19%	5%	-3%	1%	3%	
14,300 – 14,400	14,350	4%	-4%	-7%	-3%	5%	1%	9%	14%	14%	3%	-6%	-4%	2%	
14,400 – 14,500	14,450	2%	-3%	-9%	0%	-1%	2%	-2%	9%	9%	-2%	-11%	-6%	-1%	
14,500 – 14,600	14,550	0%	0%	-7%	-2%	0%	-3%	3%	15%	6%	2%	-4%	-1%	1%	**1%**
14,600 – 14,700	14,650	1%	-3%	-8%	1%	6%	0%	5%	17%	14%	7%	-8%	-4%	2%	
14,700 – 14,800	14,750	-2%	-4%	-10%	-4%	-1%	0%	-4%	12%	9%	3%	-7%	2%	0%	
14,800 – 14,900	14,850	-1%	-5%	-10%	-6%	0%	-2%	-3%	12%	11%	1%	-7%	-4%	-1%	
14,900 – 15,000	14,950	-2%	0%	-15%	-8%	-3%	-1%	-1%	16%	8%	-1%	-8%	-2%	-1%	
15,000 – 15,100	15,050	-1%	-5%	-14%	-7%	-1%	1%	3%	13%	13%	-2%	-8%	-2%	-1%	
15,100 – 15,200	15,150	-1%	-1%	-11%	-7%	1%	-1%	2%	21%	12%	6%	-7%	2%	1%	
15,200 – 15,300	15,250	-6%	-5%	-9%	-3%	4%	-2%	-1%	10%	8%	-3%	-10%	-2%	-2%	
15,300 – 15,400	15,350	-7%	-9%	-6%	-4%	-2%	-7%	-2%	17%	11%	0%	-5%	-2%	-1%	
15,400 – 15,500	15,450	-2%	-7%	-9%	-7%	-3%	-9%	-1%	9%	8%	-3%	-10%	-2%	-3%	
15,500 – 15,600	15,550	-1%	0%	-13%	-8%	2%	-5%	-2%	8%	13%	3%	-6%	-4%	-1%	**-1%**
15,600 – 15,700	15,650	4%	-5%	-5%	-10%	-6%	0%	-1%	14%	7%	0%	-10%	3%	-1%	
15,700 – 15,800	15,750	1%	-2%	-6%	-2%	-3%	-7%	2%	16%	5%	-6%	-10%	0%	-1%	
15,800 – 15,900	15,850	-1%	-8%	-9%	-6%	0%	-1%	6%	13%	10%	2%	-11%	2%	0%	
15,900 – 16,000	15,950	-1%	-7%	-11%	-2%	-5%	-3%	-3%	14%	8%	4%	-9%	2%	-1%	
16,000 – 16,100	16,050	-2%	-6%	-9%	-6%	-1%	-5%	-2%	9%	8%	-1%	-6%	-1%	-2%	
16,100 – 16,200	16,150	1%	-3%	-6%	-1%	0%	-4%	6%	15%	9%	3%	-7%	2%	1%	
16,200 – 16,300	16,250	3%	-4%	-8%	-4%	-1%	0%	3%	13%	15%	8%	-7%	1%	1%	
16,300 – 16,400	16,350	-1%	0%	-10%	-5%	2%	1%	3%	9%	11%	5%	-6%	-1%	1%	
16,400 – 16,500	16,450	-6%	-11%	-10%	-1%	-2%	-6%	-6%	9%	5%	-1%	-10%	-8%	-4%	
16,500 – 16,600	16,550	0%	-3%	-12%	-5%	-3%	0%	4%	16%	12%	5%	-9%	1%	1%	**0%**
16,600 – 16,700	16,650	2%	0%	-10%	-2%	-3%	-5%	1%	16%	11%	5%	-5%	1%	1%	
16,700 – 16,800	16,750	1%	0%	-6%	-9%	-3%	-5%	3%	13%	15%	2%	-6%	2%	1%	
16,800 – 16,900	16,850	-1%	-1%	-11%	-4%	-7%	-6%	1%	18%	14%	9%	-1%	3%	1%	
16,900 – 17,000	16,950	6%	-3%	-10%	-2%	-5%	-2%	0%	12%	13%	7%	-9%	0%	1%	
17,000 – 17,100	17,050	5%	-1%	-7%	-4%	-1%	-3%	0%	21%	20%	0%	-10%	1%	2%	
17,100 – 17,200	17,150	4%	-5%	-10%	-8%	10%	1%	-3%	19%	11%	7%	-5%	3%	2%	**1%**
17,200 – 17,300	17,250	2%	6%	-1%	-5%	6%	1%	-1%	16%	11%	8%	-6%	-3%	3%	

17,300 – 17,400	17,350	9%	-1%	-10%	-5%	-5%	-5%	1%	18%	11%	9%	-6%	6%	2%	
17,400 – 17,500	17,450	7%	-10%	-7%	-3%	-4%	-8%	-2%	21%	19%	8%	-10%	9%	2%	
17,500 – 17,600	17,550	-3%	-3%	-7%	-7%	-4%	-4%	-2%	18%	3%	-4%	-12%	-3%	-2%	
17,600 – 17,700	17,650	-1%	4%	-12%	-7%	-6%	0%	-13%	13%	8%	-6%	-12%	-4%	-3%	
17,700 – 17,800	17,750	11%	8%	6%	-3%	-4%	7%	1%	30%	8%	2%	-7%	1%	5%	
17,800 – 17,900	17,850	1%	-12%	-8%	-12%	0%	-6%	-6%	20%	0%	-1%	-16%	2%	-3%	
17,900 – 18,000	17,950	-4%	3%	-9%	-3%	2%	-3%	-6%	15%	4%	1%	-16%	-1%	-1%	
18,000 – 18,100	18,050	5%	5%	-8%	2%	2%	-2%	-3%	28%	7%	3%	-9%	0%	2%	
18,100 – 18,200	18,150	17%	2%	-5%	-3%	8%	8%	-2%	17%	14%	7%	-13%	1%	4%	
18,200 – 18,300	18,250	10%	-2%	-5%	-1%	1%	0%	5%	21%	24%	2%	-15%	-3%	3%	
18,300 – 18,400	18,350	1%	-5%	-4%	-8%	-3%	3%	2%	14%	11%	-3%	-17%	-2%	-1%	
18,400 – 18,500	18,450	-3%	-3%	-6%	-2%	3%	12%	8%	25%	8%	3%	-17%	6%	3%	
18,500 – 18,600	18,550	0%	3%	-2%	-5%	5%	0%	4%	22%	12%	2%	-19%	8%	3%	**3%**
18,600 – 18,700	18,650	-3%	-2%	-3%	-2%	0%	2%	7%	12%	10%	10%	-18%	-1%	1%	
18,700 – 18,800	18,750	-4%	-1%	-13%	0%	5%	10%	10%	27%	12%	-2%	-13%	-2%	2%	
18,800 – 18,900	18,850	-2%	10%	2%	6%	5%	11%	13%	23%	10%	10%	-10%	6%	7%	
18,900 – 19,000	18,950	2%	-4%	-4%	3%	15%	8%	12%	15%	21%	8%	-10%	6%	6%	
19,000 – 19,100	19,050	7%	14%	-6%	16%	19%	23%	8%	13%	25%	17%	0%	7%	12%	
19,100 – 19,200	19,150	6%	-5%	1%	5%	-2%	14%	1%	39%	19%	10%	-8%	13%	8%	
19,200 – 19,300	19,250	-1%	7%	3%	7%	-5%	3%	12%	22%	14%	1%	-10%	13%	5%	
19,300 – 19,400	19,350	-3%	18%	8%	7%	4%	9%	20%	29%	2%	11%	-10%	20%	10%	
19,400 – 19,500	19,450	18%	17%	-5%	1%	12%	14%	-3%	46%	7%	21%	-22%	15%	10%	
19,500 – 19,600	19,550	14%	20%	9%	20%	-2%	7%	4%	15%	31%	15%	10%	6%	12%	**9%**
19,600 – 19,700	19,650	12%	13%	2%	19%	10%	18%	15%	36%	25%	-4%	-23%	-2%	10%	
19,700 – 19,800	19,750	14%	17%	6%	23%	9%	4%	5%	12%	22%	2%	-5%	0%	9%	
19,800 – 19,900	19,850	45%	-14%	6%	4%	10%	-7%	-2%	31%	10%	4%	-4%	-5%	6%	
19,900 – 20,000	19,950	-1%	7%	12%	-3%	-12%	3%	0%	40%	13%	6%	2%	11%	6%	

B.1. Detailed validation results for SAM from 2002 to 2012 annually at 100 km intervals

Appendix C

This appendix contains fundamental assumptions for GEO-4 scenarios.

Appendix C

		Fundamental assumption			
Driver category	**Critical uncertainty**	**Markets First**	**Policy First**	**Security First**	**Sustainability First**
Institutional and socio-political frameworks	What is the dominant scale of decision making?	International	International	National	None
	What is the general nature and level of international cooperation?	High, but with focus on economic issues (trade)	High	Low	High
	What is the general nature and level of public participation in governance?	Low	Medium	Lowest	High
	What is the power balance between government, private and civil sector actors?	More private	More government	Government and certain private	Balanced
	What is the overall level and distribution of government investment across areas (e.g., health, education, military and R&D)?	Medium, fairly evenly distributed	Higher, more emphasis on health and education	Low, focus on military	Highest, more emphasis on health and education
	What is the general nature and level of official development assistance?	Low	Higher, increasingly as grants not loans	Lowest	Highest, increasingly as grants not loans
	To what degree is there mainstreaming of social and environmental policies?	Low, for example, little or no specific climate policy, reactive policies with respect to local air pollutants	High, for example, aims at stabilization of CO2-equivalent concentration at 650 ppmv, proactive policies on local air pollutants	Lowest, for example, little or no specific climate policy, reactive policies with respect to local air pollutants	Highest, for example, aims at stabilization of CO2-equivalent concentration at 550 ppmv, proactive policies on local air pollutants
Demographics	What actions are taken related to international migration?	Open borders	Fairly open borders	Closed borders	Open borders
	How many children do women want to have when the choice is theirs to make?	Continued trend towards fewer births as income rises	Accelerated trend	Slowing trend	Accelerated trend
Economic	What actions are taken related to	Move to increased openness,	Increasingly open, with	Moves towards	Increasing open, with

demand, markets and trade	the openness of international markets?	with few controls	some embodiment of fair trade principles	protectionism	strong embodiment of fair trade principles
	To what degree is there an emphasis on sectoral specialization vs. diversification in the economy?	Specialized	Balanced	Diverse, but with emphasis on sectors of interest to governments and powerful private sector actors	Diverse
	How much do people choose to work in the formal economy?	Most work in formal economy	Most work in formal economy	Larger underground economies	Variable by region and societal groups
	What is the general level and emphasis of government intervention in the economy?	Low, efficient markets	High, efficient but also fair markets	Variable by region and sector	Medium, greater emphasis on fairness of markets
Scientific and technological innovation	What are the levels, sources, and emphases of R&D investment?	High, primarily private or by government at behest of private sector, for profit	High, primarily government	Variable, government and certain private sector actors	High, from range of sources
			Benign, but still with eye on profit	Military/security	Benign, appropriate
	What is the emphasis in terms of energy technologies?	Focus on economic efficiency	Focus on general efficiency and environmental impact	Emphasis on security of supply	Focus on general efficiency, environmental impact
	What is done with respect to the access and availability of new technologies?	What you can pay for, primarily through trade	Promotion of technology transfer and diffusion	Closely guarded	Promotion of technology transfer and diffusion, and encouragement of open source technologies
Value systems	What actions are taken related to cultural homogenization vis-à-vis diversity?	Little overt action	Little overt action	Diverse, tending towards xenophobia	Efforts to maintain diversity and tolerance
	What is the emphasis on individualism vis-à-vis the community?	Individual	More community	Individual	Community
	What is the relative rank of conflicting priorities in fisheries?	Profits	Balance between profits, total catch and jobs	Total catch	Focus on ecosystem restoration, but also

				emphasis on jobs and landings
What are the key priorities with regard to protected areas?	"Sustainable use," emphasizing tourism development and some genetic resource protection	Species conservation and ecosystem services Maintenance, then sustainable use, including benefit sharing	Tourism development, and some genetic resource protection	Sustainable use, including benefit sharing, then ecosystem services maintenance and species conservation
How do resource demands shift, independent of changing prices and income?	Follow traditional patterns	Follow traditional patterns for most resources, but some relative reduction in water use	Follow traditional patterns	Slower uptake of meat consumption, energy use, water use and other resource use with rising income

C.1. Key questions related to scenario assumptions (GEO-4, 2007)

Appendix D

The appendix contains the top 15 APD connections according to the forecasted passenger numbers for GEO-4 scenarios:

D.1. Top 15 APD connections in 2042 by APD numbers for the Markets First scenario

D.2. Top 15 APD connections in 2042 by APD numbers for the Policy First scenario

D.3. Top 15 APD connections in 2042 by APD numbers for the Security First scenario

D.4. Top 15 APD connections in 2042 by APD numbers for the Sustainability First scenario

Settlement 1	Country 1	Region 1	Settlement 2	Country 2	Region 2	APD	Average airfare, 2005 US$
Beijing	China	Asia	Shanghai	China	Asia	58,862,120	244
Shanghai	China	Asia	Shenzhen	China	Asia	39,683,246	257
Beijing	China	Asia	Guangzhou	China	Asia	35,578,017	318
Beijing	China	Asia	Chengdu	China	Asia	35,036,931	287
Guangzhou	China	Asia	Shanghai	China	Asia	33,205,948	254
Beijing	China	Asia	Shenzhen	China	Asia	31,643,330	325
Delhi	India	Asia	Mumbai	India	Asia	31,358,054	248
Hanoi	Vietnam	Asia	Ho Chi Minh	Vietnam	Asia	23,652,385	250
Jakarta	Indonesia	Asia	Surabaya	Indonesia	Asia	22,821,502	205
Jakarta	Indonesia	Asia	Medan	Indonesia	Asia	22,588,928	272
Shanghai	China	Asia	Xiamen	China	Asia	22,277,926	217
Beijing	China	Asia	Xian	China	Asia	21,812,261	228
Melbourne	Australia	Oceania	Sydney	Australia	Oceania	21,749,443	207
Denpasar Bali	Indonesia	Asia	Jakarta	Indonesia	Asia	21,698,431	233
Beijing	China	Asia	Hangzhou	China	Asia	21,550,055	249

D.1. Top 15 APD connections in 2042 by APD numbers for the Markets First scenario

Settlement 1	Country 1	Region 1	Settlement 2	Country 2	Region 2	APD	Average airfare, 2005 US$
Beijing	China	Asia	Shanghai	China	Asia	57,137,447	246
Shanghai	China	Asia	Shenzhen	China	Asia	38,520,516	259
Beijing	China	Asia	Guangzhou	China	Asia	34,535,579	321
Beijing	China	Asia	Chengdu	China	Asia	34,010,342	290
Guangzhou	China	Asia	Shanghai	China	Asia	32,233,007	256
Beijing	China	Asia	Shenzhen	China	Asia	30,716,171	328
Delhi	India	Asia	Mumbai	India	Asia	29,355,603	249
Hanoi	Vietnam	Asia	Ho Chi Minh	Vietnam	Asia	25,361,721	251
Jakarta	Indonesia	Asia	Surabaya	Indonesia	Asia	22,393,764	206
Jakarta	Indonesia	Asia	Medan	Indonesia	Asia	22,165,557	274
Shanghai	China	Asia	Xiamen	China	Asia	21,625,174	218
Denpasar Bali	Indonesia	Asia	Jakarta	Indonesia	Asia	21,291,754	234
Beijing	China	Asia	Xian	China	Asia	21,173,158	230
Beijing	China	Asia	Hangzhou	China	Asia	20,918,626	250
Chengdu	China	Asia	Guangzhou	China	Asia	20,874,103	258

D.2. Top 15 APD connections in 2042 by APD numbers for the Policy First scenario

Settlement 1	Country 1	Region 1	Settlement 2	Country 2	Region 2	APD	Average airfare, 2005 US$
Beijing	China	Asia	Shanghai	China	Asia	33,710,931	240
Shanghai	China	Asia	Shenzhen	China	Asia	22,726,994	252
Delhi	India	Asia	Mumbai	India	Asia	22,328,455	244
Hanoi	Vietnam	Asia	Ho Chi Minh	Vietnam	Asia	21,410,633	246
Beijing	China	Asia	Guangzhou	China	Asia	20,375,892	312
Beijing	China	Asia	Chengdu	China	Asia	20,066,003	282
Jakarta	Indonesia	Asia	Surabaya	Indonesia	Asia	19,920,279	203
Jakarta	Indonesia	Asia	Medan	Indonesia	Asia	19,717,279	267
Guangzhou	China	Asia	Shanghai	China	Asia	19,017,380	250
Denpasar Bali	Indonesia	Asia	Jakarta	Indonesia	Asia	18,939,986	229
Rio De Janeiro	Brazil	South America	Sao Paulo	Brazil	South America	18,726,589	170
Beijing	China	Asia	Shenzhen	China	Asia	18,122,455	319
Abuja	Nigeria	Africa	Lagos	Nigeria	Africa	16,421,394	186
Melbourne	Australia	Oceania	Sydney	Australia	Oceania	16,306,412	204
Jeju	South Korea	Asia	Seoul	South Korea	Asia	14,108,187	180

D.3. Top 15 APD connections in 2042 by APD numbers for the Security First scenario

Settlement 1	Country 1	Region 1	Settlement 2	Country 2	Region 2	APD	Average airfare, 2005 US$
Beijing	China	Asia	Shanghai	China	Asia	56,748,222	239
Shanghai	China	Asia	Shenzhen	China	Asia	38,258,115	251
Beijing	China	Asia	Guangzhou	China	Asia	34,300,312	309
Beijing	China	Asia	Chengdu	China	Asia	33,778,660	280
Guangzhou	China	Asia	Shanghai	China	Asia	32,013,429	248
Beijing	China	Asia	Shenzhen	China	Asia	30,506,932	316
Delhi	India	Asia	Mumbai	India	Asia	28,265,859	242
Hanoi	Vietnam	Asia	Ho Chi Minh	Vietnam	Asia	24,294,366	244
Shanghai	China	Asia	Xiamen	China	Asia	21,477,862	213
Beijing	China	Asia	Xian	China	Asia	21,028,923	224
Beijing	China	Asia	Hangzhou	China	Asia	20,776,130	243
Chengdu	China	Asia	Guangzhou	China	Asia	20,731,906	250
Chengdu	China	Asia	Shanghai	China	Asia	20,149,781	293
Jakarta	Indonesia	Asia	Surabaya	Indonesia	Asia	19,977,439	202
Jakarta	Indonesia	Asia	Medan	Indonesia	Asia	19,773,854	265

D.4. Top 15 APD connections in 2042 by APD numbers for the Sustainability First scenario