

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF TRANSPORTATION SCIENCES

Lenka Jonáková

**MODELOVÁNÍ POPTÁVKY PO LETECKÉ PŘEPRAVĚ VE
STOCHASTICKÉ DOPRAVNÍ SÍTI**

**MODELLING OF AIR TRAFFIC DEMAND IN STOCHASTIC
TRANSPORTATION NETWORK**

Diploma Thesis

2015



K621..... Ústav letecké dopravy

ZADÁNÍ DIPLOMOVÉ PRÁCE
(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení studenta (včetně titulů):

Bc. Lenka Jonáková

Kód studijního programu a studijní obor studenta:

N 3710 – PL – Provoz a řízení letecké dopravy

Název tématu (česky): **Modelování poptávky po letecké přepravě
ve stochastické dopravní síti**

Název tématu (anglicky): Modelling of Air Traffic Demand in Stochastic
Transportation Network

Zásady pro vypracování

Při zpracování diplomové práce se řiďte osnovou uvedenou v následujících bodech:

- Úvod
- Popis základního matematického aparátu použitého v úloze klasifikace
- Popis vlastní matematické úlohy a její aplikace na reálný problém
- Popis výsledků aplikace úlohy klasifikace na reálná data
- Rozbor výsledků experimentů
- Analýza výsledků experimentů a zhodnocení metody aplikované na danou problematiku
- Závěr

- Rozsah grafických prací: dle pokynů vedoucího diplomové práce
- Rozsah průvodní zprávy: minimálně 55 stran textu (včetně obrázků, grafů a tabulek, které jsou součástí průvodní zprávy)
- Seznam odborné literatury: SRINIDHI, S. Development of an airline traffic forecasting model on international sectors. IEEE, 2009.
- LESAGE, James P. a R. Kelley PACE. Spatial econometric modeling of origin-destination flows. Journal of Regional Science. 2008.
- NAGY, Ivan. ČVUT V PRAZE, Fakulta dopravní. Stochastické systémy.

Vedoucí diplomové práce: **doc. Ing. Ivan Nagy, CSc.**
Ing. Bc. Jakub Hospodka, Ph.D.

Datum zadání diplomové práce: **30. června 2014**
(datum prvního zadání této práce, které musí být nejpozději 10 měsíců před datem prvního předpokládaného odevzdání této práce vyplývajícího ze standardní doby studia)

Datum odevzdání diplomové práce: **30. listopadu 2015**


a) datum prvního předpokládaného odevzdání práce vyplývající ze standardní doby studia a z doporučeného časového plánu studia

b) v případě odkladu odevzdání práce následující datum odevzdání práce vyplývající z doporučeného časového plánu studia

L. S.


.....
doc. Ing. Stanislav Szabo, PhD. MBA prof.
vedoucí
Ústavu letecké dopravy




.....
Dr. Ing. Miroslav Svítek, dr. h. c.
děkan fakulty

Potvrzuji převzetí zadání diplomové práce.


.....
Bc. Lenka Jonáková
jméno a podpis studenta

V Praze dne..... 30. června 2015



CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Transportation Sciences

Dean's office

Konviktská 20, 110 00 Prague 1, Czech Republic

K621..... Department of Air Transport

MASTER'S THESIS ASSIGNMENT

(PROJECT, WORK OF ART)

Student's name and surname (including degrees):

Bc. Lenka Jonáková

Code of study programme code and study field of the student:

N 3710 – PL – Air Traffic Control and Management

Theme title (in Czech): **Modelování poptávky po letecké přepravě
ve stochastické dopravní síti**

Theme title (in English): Modelling of Air Traffic Demand in Stochastic
Transportation Network

Guides for elaboration

During the elaboration of the master's thesis follow the outline below:

- Introduction
- Description of fundamental mathematical tools used for data classification
- Description of a particular mathematical task and its application on a real problem
- Description of the results given by application of the data classification on real data
- Interpretation of the experiments results
- Analysis of the experiments results and evaluation of the method implemented
- Conclusion

Graphical work range: according to the instructions of the master's thesis supervisor

Accompanying report length: at least 55 pages of text (including figures, graphs and tables, which are part of the accompanying report)

Bibliography: SRINIDHI, S. Development of an airline traffic forecasting model on international sectors. IEEE, 2009.

LESAGE, James P. a R. Kelley PACE. Spatial econometric modeling of origin-destination flows. Journal of Regional Science. 2008.

NAGY, Ivan. ČVUT V PRAZE, Fakulta dopravní. Stochastické systémy.

Master's thesis supervisor: **doc. Ing. Ivan Nagy, CSc.**
Ing. Bc. Jakub Hospodka, Ph.D.


Date of master's thesis assignment: **June 30, 2014**
(date of the first assignment of this work, that has be minimum of 10 months before the deadline of the theses submission based on the standard duration of the study)

Date of master's thesis submission: **November 30, 2015**
a) date of first anticipated submission of the thesis based on the standard study duration and the recommended study time schedule
b) in case of postponing the submission of the thesis, next submission date results from the recommended time schedule

L. S.


.....
doc. Ing. Stanislav Szabo, PhD, MBA
head of the Department
of Air Transport




.....
Dr. Ing. Miroslav Svítek, dr. h. c.
dean of the faculty

I confirm assumption of master's thesis assignment.


.....
Bc. Lenka Jonáková
Student's name and signature

Prague June 30, 2015

Poděkování

Ráda bych touto cestou vyjádřila poděkování panu doc. Ing. Ivanovi Nagymu, CSc. za jeho nezměrnou ochotu, cenné rady a trpělivost při vedení mé diplomové práce. Rovněž bych chtěla poděkovat společnosti Letiště Praha, a. s. za poskytnutí potřebných dat, bez nichž by tato práce nemohla vzniknout.

Zvláštní poděkování patří mým rodičům a blízkým za jejich podporu, které se mi dostávalo po celou dobu studia. Jejich přístup mne motivoval a inspiroval nejen během studií, ale mnohem podstatněji také v osobním životě.

Prohlášení

Předkládám tímto k posouzení a obhajobě diplomovou práci, zpracovanou na závěr studia na ČVUT v Praze Fakultě dopravní.

Prohlašuji, že jsem předloženou práci vypracovala samostatně a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu § 60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V Praze dne 30. listopadu 2015

.....

Bc. Lenka Jonáková

CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Transportation Sciences

MODELOVÁNÍ POPTÁVKY PO LETECKÉ PŘEPRAVĚ VE STOCHASTICKÉ
DOPRAVNÍ SÍTI

MODELLING OF AIR TRAFFIC DEMAND IN STOCHASTIC
TRANSPORTATION NETWORK

Diploma thesis
November 2015
Lenka Jonáková

ABSTRAKT

Tato práce si klade za cíl vytvořit spolehlivý model poptávky po letecké přepravě a následně ohodnotit přiměřenost implementované metody. Vzhledem ke stochastické povaze dané problematiky jsou pro řešení využity metody teorie pravděpodobnosti; konkrétně se jedná o odhad směsového modelu a následnou aplikaci lineární a logistické regresní analýzy. Pro účely objektivního zhodnocení navržené metody jsou získané výsledky porovnány s metodou jednoduché lineární a logistické regrese.

Klíčová slova: směsový model, regresní analýza, letecká doprava, poptávka, stochastická síť

ABSTRACT

The aim of the thesis is to create a reliable mathematical model of air traffic demand and subsequently evaluate the reasonability of the implemented method. Considering the stochastic nature of the issue, the probabilistic approach is utilized; particularly, the estimate of a mixture model and the subsequent linear and logistic regression analysis is exploited. To objectively evaluate the suggested method, results are compared to the simple linear and logistic regression method.

Keywords: mixture model, regression analysis, air traffic, demand, stochastic network

TABLE OF CONTENT

1	INTRODUCTION	7
	Objectives of the Thesis.....	9
2	TASK MOTIVATION	11
2.1	Czech Airlines, a.s.	11
3	FACTORS AFFECTING AIR TRAFFIC DEMAND	13
3.1	Input Data	14
4	MODELLING APPROACHES	15
5	MODEL PRESENTATION	17
5.1	Pre-modelling Data Adjustment.....	18
5.2	Linear Regression Model	19
5.2.1	Statistics Recursion	19
5.2.2	Point Estimate of Parameters	20
5.2.3	Point Estimate of Output.....	21
5.3	Discrete Model	21
5.3.1	Statistics Recursion	21
5.3.2	Point Estimate of Parameters	22
5.3.1	Point Estimate of Output.....	23
5.4	Logistic Model	23
5.4.1	Point Estimate of Parameters	25
5.4.2	Point Estimate of Output.....	26
5.5	Mixture Model.....	26
5.5.1	Derivation of Mixture Estimation	27
5.5.2	Estimation of Parameter θ	28
5.5.3	Estimation of Parameter α	29
5.5.4	Estimation of Weighting Vector w	29
6	SOFTWARE PROCESSING.....	30
6.1	Estimation of Mixture Model	30
6.2	Linear Regression Model	32

6.3	Logistic Regression Model	33
6.4	Predefined Functions	34
6.4.1	Likelihood Function.....	34
6.4.2	Supervised Learning.....	34
6.4.3	Data Testing	35
7	RESULTS.....	36
7.1	Simple Linear Regression	36
7.1.1	Estimation of Relevant Variables	36
7.1.2	Estimation of Linear Regression Model.....	37
7.2	Simple Logistic Regression	37
7.2.1	Estimation of Relevant Variables	38
7.2.2	Estimation of Logistic Regression Model	38
7.3	Mixture Model.....	39
7.3.1	Solution Variability	39
7.3.2	Data Classification	41
7.3.3	Linear Regression Analysis	43
7.3.3.1.	Estimation of Relevant Variables.....	44
7.3.3.2.	Estimation of Linear Regression Models	45
7.3.4	Logistic Regression Analysis	47
7.3.4.1.	Estimation of Relevant Variables.....	47
7.3.4.2.	Estimation of Logistic Regression Models	48
8	DISCUSSION	50
8.1	Model Criticism.....	55
8.1.1	Model Formulation	55
8.1.2	Choice of Variables.....	55
8.1.3	Initial Parameters Setting.....	56
8.2	Data Criticism.....	56
8.2.1	Origin and Destination Data.....	56
8.2.2	Socio-economic Data.....	56

8.3	Comparison with ICAO Methodology.....	57
9	CONCLUSION.....	59
10	RESOURCES	60
11	LIST OF FIGURES.....	62
12	LIST OF TABLES.....	64

1 INTRODUCTION

For more than one hundred years, aviation has increasingly been making its way into our everyday lives. Because of its significant international character, air transport has always played an important role in social, economic and political development, at a global as well as at a local level. Furthermore, as has been indicated in many research works, existing socio-economic conditions can be reverse engineered to estimate air traffic demand. This presumption creates a fundamental keystone for all calculations which proceed in this thesis. In order to estimate such a random variable as air traffic demand, it is necessary to fully understand the historical circumstances and conditions under which this demand was created.

The significant impact of aviation on the global economy, as well as on the world political and social situation, was fully evident from the very beginning of its history. Due to its strong international character, air transport has always played an important role in global trade, tourism, investment, labour supply, consumer welfare, and market efficiency development [1]. However, this is not the picture which the general public has perceived during the last century. Aviation development, as with development of many other fields, was primarily accelerated by the military conflicts of the twentieth century. One of the regrettable examples of military activity carried out via air was the dropping of nuclear bombs on the cities of Hiroshima and Nagasaki. Over time the scenarios have changed, but the stigma with which this field was marked has never disappeared. The threats resulting from the international character of aviation survive to this day, for instance in the form of international terrorism.

In 1944 the international importance of air transport's future development was distinctly communicated in the preamble of the Convention on International Civil Aviation, the essential document of aviation law. The direct impact of international civil aviation on "creation and preservation of friendship and understanding among the nations and people of the world" is strongly emphasized here, together with a threat to general security in the case of its abuse. The Convention also set cornerstones for market regulation restrictions. Based on The Freedoms of the Air, flight frequencies and capacity control were established, as well as price tariffs in civil and cargo transportation [2].

All aspects mentioned above have significantly influenced the overall market conditions and consequently the ability to satisfy air traffic demand. The post war economic situation in airline transportation was comprehensively described by Irston R. Barnes, Professor

of Economics at Yale University, in 1946. In his publication he highlighted three general factors crucial for the development of future demand. From an economic point of view he considered the general level of national income as most important, which is closely related to a high level of production, and the overall utilization of resources. The second factor is the public acceptance of air transport from a non-economic standpoint. Both of these factors had been significantly suppressed by the First and Second World Wars. The competitive position of air transportation in relation to competing surface transportation is the third aspect, and one which dominates in more detailed assessments of air traffic demand. Barnes saw the major political threat to full development of commercial aviation in market regulation, which was absolutely irrespective of any basic economic principles of free market environment and often led to cartel agreements, financing and capital investment issues, reduction of incentives to service, and last but not least to problems with aggressive taxation. As will be comprehensively described in the upcoming chapters, all of these factors remain essential for the estimation of air traffic demand today [3].

Barnes' vision of a free market came to pass more than 30 years after his study had been published. The so-called "Open Skies" policy was firstly implemented in the United States in the eighties; in the nineties it spread into Europe and changed the field forever. Deregulation meant a rapid increase in competition in the market and forced air carriers to radically change their business and network strategies. The majority of 'full-service' airlines have adopted the so-called "hub-and-spoke strategy", which enables them to maximize connection opportunities and discourages other competitors from entering the market because of their strengthened bureaucratic control over the hub airport. Low-cost, no frills airlines have followed a completely different strategy. They found the implementation of point-to-point networks serving only high volume routes and the execution of low price policy to be the most effective strategies [4].

Deregulation practices influenced not just air carriers within Europe and the United States, but also external airlines which had to adapt to the changing business environment. For these purposes the phenomenon of airline alliances was established. This represented the possibility to enhance business cooperation and at the same time to create better conditions for negotiation. Thus, it is reasonable to say that overall, deregulation promoted globalization initiatives [4].

Deregulation of air transport has not only changed the structure of airline networks but also the overall concept of air traffic demand estimation. Due to increased competition and greater effectiveness of provided services, traffic volumes became much more uncertain and

volatile. The system became stochastic, i.e. a system involving random variables, which can be described by probability distribution.

All of the historical events mentioned above have helped to create, step by step, air transport as we experience it nowadays. Military conflicts and threats resulting from aviation abuse forced the industry to reach technological perfection, which is not obviously the case for most other transportation systems. Furthermore, due to deregulation, air transport became a liberalized field respecting the basic laws of economics. These can be utilized in order to mathematically model and estimate variables, such as air traffic demand, which are present in transportation processes.

Objectives of the Thesis

The aim of the thesis is to create a reliable mathematical model of air traffic demand and subsequently evaluate the reasonability of the implemented method. Considering the stochastic nature of air transportation system, the probabilistic approach is utilized.

This thesis thematically follows the content of the previously elaborated bachelor thesis. A consistent element of both works is the fundamental presumption that air traffic demand can be expressed as a function of socio-economic parameters. The parameters selection process is a subject of detailed discussion in the upcoming chapter.

In the bachelor thesis, linear regression was utilized for the purposes of air traffic demand estimation. However, as has been demonstrated, presumption of linearity and stationarity of the modelled system was not sufficient for this purpose. As was suggested in the conclusion of the thesis, the mixture model, which is the main subject of this work, promises the possibility of minimizing errors resulting from the inappropriate presumptions.

The method was chosen because the modelled system is believed to exist in several behaviour modes. Furthermore, these modes are assumed to differ so much that each of them needs to be characterized by a special model.

In order to minimize the uncertainty associated with continuous system modelling, and thus fundamentally improve the model reliability, the dependent variable is considered not only continuous, but also discrete. For these purposes the linear regression analysis as well as the logistic regression analysis is exploited after the mixture estimation.

To objectively evaluate the suggested method, results will be compared to more basic mathematical tools such as simple linear and simple logistic regression.

The data support provided by Prague Airport enabled elaboration on the thesis. Because of the close cooperation between Prague Airport and Czech Airlines, both belonging to the Czech Aeroholding Group, task concretization is tailored to the business strategy of that company. Therefore, air traffic demand between Middle Europe and the Russian Federation market is the main subject of mathematical modelling and estimation.

2 TASK MOTIVATION

Air traffic modelling and estimation are an inseparable part of the overall aviation planning processes. Long-term planning is often associated with strategic planning, which includes a time span more than five-years into the future, and determines critical milestones together with capital expenditures of a company. The estimation supports airlines in defining their network and business strategy, assists in developing airspace and airport infrastructure, and last but not least contributes to manufacturer's strategies of new aircraft production. The medium-term horizon involves looking forward over a one-year to five-year time span and takes into account long-term trends as well as cyclical components of demand. It comprises activities such as budgeting and resource allocation, for instance assignment of aircraft to particular routes. In terms of short-term planning, seasonal factors are considered the most important; for example flight schedule creation, maintenance and catering planning are all incorporated [5].

The majority of participants in air transportation processes have to face extremely high financial risks caused by enormous initial investments, with paybacks only coming in the long term. Exploitation of mathematical modelling techniques can rapidly decrease financial risks and enhance the efficiency of processes, especially in the case of strategic planning. The European market environment is characterized by an extreme level of competition and thus, effective planning methods are crucial for economically sustainable development.

The main purpose of the mathematical model created in the thesis is to enhance processes of long-term planning, particularly of the network strategy of Czech Airlines in the context of the Russian market.

2.1 Czech Airlines, a.s.

Czech Airlines is the flagship carrier of the Czech Republic and was founded in 1923, making it one of the five oldest airlines in the world. With the hub airport located in Prague, the company provides scheduled air carriage to 89 destinations in 45 countries within Europe and Asia [6] while transporting approximately 3 million passengers per year [7].

Since deregulation, Czech Airlines has been adapting its business model in order to deal with strong competition, primarily induced by the expansion of low cost carriers. In the interest of cost reduction, a hub-and-spoke network strategy was implemented. Furthermore, the company launched a hybrid business strategy which utilizes cost reduction mechanisms,

which is comparable to the no-frills policy of low cost airlines while maintaining the codeshare cooperation and corporate clientele. Despite all efforts, Czech Airlines was not able to resist the financial crash and in 2013 was urged to strengthen its position on the market through consolidation with Korean Air, the current owner of 44% of the company's share [8].

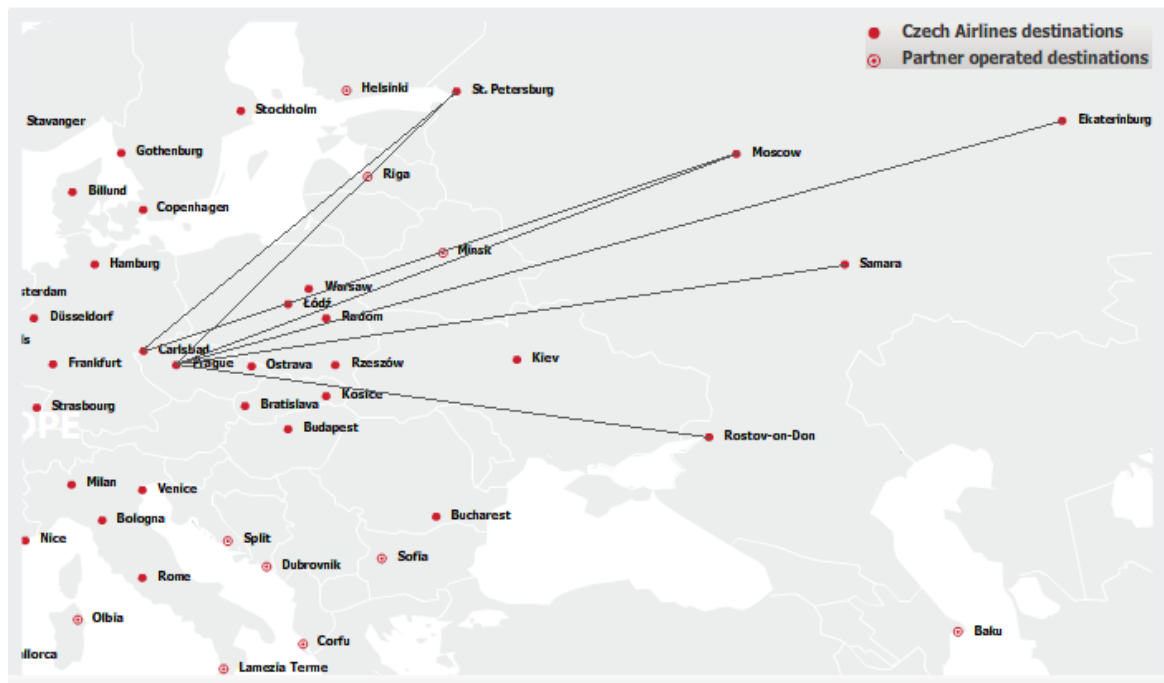


Figure 1: Connections from the Czech Republic to Russia provided by Czech Airlines nowadays [6]

The Russian Federation market has always played an important role in Czech Airlines' network development, specifically in terms of air traffic demand and revenue generation. The company's good reputation on the Russian market, as well as strong economic relations, can be documented in the long term cooperation. The first connection from Prague to Moscow was established in 1936, making it the oldest connection of the carrier. Five connections from the Czech Republic, to Moscow, Saint Petersburg, Ekaterinburg, Samara and Rostov on Don, are provided nowadays, carrying almost a quarter of a million passengers (Figure 1). Despite financial risks resulting from the current unstable political situation, unlike many other carriers Czech Airlines does not plan to back away from the Russian market. On the contrary, the market is predicted to consolidate and grow, thus new activities are the subject of continual planning [9].

Czech Airlines activities on the Russian market are further enhanced by Aeroflot's membership in Skyteam Alliance, which came into effect in 2004. Czech Airlines cooperates with Aeroflot on many air routes via code-sharing agreements [10].

3 FACTORS AFFECTING AIR TRAFFIC DEMAND

As was indicated above, air traffic demand can be defined as a variable dependent on two main types of drivers. These are the geo-economic and service dependent factors [11].

Economic factors refer to commercial, industrial and cultural activities in the respective transportation area. Considering the complex nature and broad span of these activities, selecting variables sufficiently describing the overall economic situation is quite problematic. In the majority of mathematical models, population and income are usually utilized as explanatory variables. However, these variables themselves can be often substituted by many other economic indicators, because of some strong correlations existing between them. Therefore, some studies have also considered other variables, such as the percentage of degree holders and employment composition structure. Nevertheless, their utilization has never been put into a coherent framework [11].

Another important geo-economic aspect in the process of air traffic estimation is transportation distance. As the distance between destinations increases, the relative competitiveness in terms of travel time improves. As a result, smaller players in the market can also reach better economic conditions for operating long-haul connections. On the other hand, with the increase in distance the social and commercial interactions between destinations are narrowing. The negative effects of closely placed competing airports cannot be neglected either. In this case demand is critically determined by the frequency of departures, even at the cost of extra travel times [11].

In micro-economic theory, demand is viewed as a variable directly dependent on price. In the specific context of air traffic, price is perceived as the charge for provided services, and thus modelling often requires utilization of both factors, i.e. the price charged as well as the quality of the product. In terms of quality, frequency of departures and load factor were observed to have the most significant effect. The decrease in load factor is assumed to lower the probability of delay. Together with the increase in frequency, these two factors perceptibly improve the level of the service provided. Furthermore, aircraft size and technology can also be taken into account as dummy variables for the purposes of defining the boundary conditions of the model [11].

3.1 Input Data

Processing of the input data selection was executed with respect to the empirical findings described in Chapter 3. Moreover, the requirement of variables independency was also taken into account.

The geo-economic factors influencing air traffic demand development are captured in variables: population density, gross domestic product per capita, unemployment rate, average monthly nominal wage and distance between the respective destinations. Conversely, variables such as revenue and average fare represent the service dependent aspects.

It is important to stress the fact that the possibilities of choice of variables were significantly limited by the structure of databases used. For that reason, only indicators capturing the situation on a regional level were collected. Data concerning the area of the European Union were, in most cases, obtained from the Eurostat database. In this case, indicators following the NUTS 2 and NUTS 3 methodology were selected. The Russian Federal State Statistics Service was used as a source of information for data regarding Russian Federation regions. The origin-destination data were provided by Prague Airport, as was emphasized earlier.

Generally speaking, the origin-destination data describe demand regarding more than 4,000 different air routes, including 60 destinations in Middle Europe and 125 destinations in Russian Federation. The selected socio-economic data comprise seven different indicators about the particular air connections. In total, the matrix is made up of almost fifty thousand data samples.

4 MODELLING APPROACHES

Modelling approaches vary extensively in terms of intended use and the particular application concerned. The techniques available for air traffic estimation purposes can be divided into three main categories: quantitative, qualitative and decision analysis [5].

One of the methods represented by the quantitative approach is time-series analysis. It is based on the presumption that historical patterns and conditions of an operating environment will continue into the future; thus, this method heavily relies on stability in past development as well as on the availability and reliability of historical data. Widely used mathematical tools in this category include trend projection, i.e. a graphical extrapolation of dataset using various types of trend curves for establishing the best fit possible, and decomposition methods sensitive not only to the trend factor but also to seasonal and cyclical changes. For short and medium term prognoses, time-series methods seem to be an adequate mathematical tool. However, these often cannot reasonably describe significant disturbances appearing in the long-term period. Causal methods offer a suitable alternative for this case. They investigate a cause-and-effect relationship between variables, and when utilized wisely have a huge potential to estimate the ups and downs of a market. Regression analysis is one of the most popular causal methods of forecasting civil aviation demand, often used in the form of multiple regression or econometric analysis. Causal methods also comprise simultaneous equations models and spatial equilibrium models. Simultaneous equations models, as the name suggests, involve more than one equation and all parameters of the model are solved concurrently. Spatial equilibrium models assume air traffic directly proportional to the size characteristic of a particular region, and inversely proportional to the distance [5].

Qualitative methods are best suited to situations where data are not applicable, highly limited or unavailable, and where expert judgement is the most valuable consideration. Delphi technique, for example, as well as technological forecasting, belongs to this category. Generally speaking, their function lies mainly in bringing together information from many experts and using this as the basis for making a final decision [5].

A combination of both quantitative and qualitative approaches is considered in decision analysis. This discipline involves a wide range of techniques such as market research, industry surveys, probabilistic analysis, Bayesian analysis, as well as system dynamics. Decision analysis has proved its undisputed qualities in the assessment of uncertainty and in risk analysis [5]. From the above description it is apparent that the utilization of a combined

approach can provide a significant advantage in terms of model complexity and reliability. Therefore, it is intended to apply this combined approach. For the purposes of mathematical modelling, causal analysis, particularly estimation of a mixture model, linear regression analysis and logistic regression analysis, is exploited as part of a quantitative approach. Within qualitative technique, initial modelling parameters as well as initial variables are chosen with respect to expert knowledge and further evaluated by applying probabilistic methods.

5 MODEL PRESENTATION

The solution provided in the bachelor thesis presumed the linearity and stationarity of the modelled system, and therefore linear regression was used for the purposes of air traffic demand estimation. As was demonstrated, this approach did not provide a sufficiently reliable level of results. Based on the elaborated calculations, the use of a mixture model was suggested as a more suitable approach in the conclusion of the thesis. This is the method which promises the possibility to minimize errors resulting from the inappropriate presumptions of the modelled system.

Mixture models provide a convenient solution in cases where the observed system exists in different behaviour modes [12]. The air traffic demand model is believed to work in such modes. Components of the mixture model, represented by these particular modes, are estimated in the classification task.

Classification is a statistical method used in order to build a predicative model to separate and classify new data points. The classification procedure consists of two phases. Firstly, a classification model is created on the basis of the observed features of a training data set; this step is called *supervised learning*. After that, the classification model is utilized in order to assign new data points into predefined classes [13].

Before initiation of the classification procedure, it is often essential to adjust the input data appropriately. This includes for instance noise reduction, generalization and standardization of the dataset [14].

It is fully evident that attitudes toward air traffic demand modelling vary in terms of model structures as well as the input variables. To ensure the best choice of explanatory variables, four different scenarios are examined and subsequently compared and evaluated during the process of data classification. The scenario which is found to be the most appropriate is further utilized for the modelling purposes.

The first scenario includes the exploitation of all the available data, i.e. of all variables listed in Subchapter 3.1. This setting allows us to generally evaluate the mathematical approach.

The second scenario is based on the practices presented in the Manual for Air Traffic Forecasting, which was published by the International Civil Aviation Organization in 2006;

see [5]. The suggested model structure includes only four variables: revenue, distance and GDP per capita of the origin and destination.

A purely economic approach is presented in the third scenario, where demand is expressed as a function of the average fare only.

The last scenario comprises the variables, which were found to be the most important based on our research and experience with previously computed calculations. The scenario comprises eight indicators; namely fare, distance and the population density, GDP per capita and wage of the respective origin and destination.

After the estimation of mixture model and classification of the data, the components' models are defined. As was emphasized earlier, the behaviour modes of the modelled system are assumed to differ so much that it is necessary to create an independent model for each of the determined components [12]. For these purposes the linear regression analysis is exploited. Moreover, discretization of the dependent variable is assumed to minimize the uncertainty associated with continuous system modelling and significantly improve the model reliability. Thus the logistic regression analysis is also utilized.

Furthermore, the suggested method is compared to more basic mathematical tools such as the simple linear and simple logistic regression. In this way the objective evaluation of the method is enabled.

The mixture model is an extremely complex method requiring knowledge of several mathematical techniques for successful implementation. Firstly, fundamental modelling methods such as simple linear regression, discrete model and simple logistic regression are described in detail. The mixture model is presented afterwards in relation to these methods.

5.1 Pre-modelling Data Adjustment

Data standardization is an easy pre-modelling data adjustment method, which enables the comparability of regression coefficients.

Let X be a random variable with the probability distribution $N(\mu, \sigma^2)$. Then, the standardized variable \bar{X} from the variable X is

$$\bar{X} = \frac{X - \mu}{\sigma} \quad (1)$$

The mean value of \bar{X} equals zero and the variance is equal to one as follows from the definition.

5.2 Linear Regression Model

The linear regression model can be generally described by the equation

$$y_t = \psi_t' \Theta + e_t, \quad e_t \sim N(0, r) \quad (2)$$

- y_t is the dependent variable
- $\Theta = \{\theta, r\}$ is the model parameter carrying the information about the vector of regression coefficients $\theta = \{\theta_0, \theta_1, \dots, \theta_n\}$ and the noise variance r
- $\psi_t = \{1, \psi_{1;t}, \psi_{2;t}, \dots, \psi_{n;t}\}$ is the vector of regression
- e_t is noise with normal probability distribution, mean value equal to one and constant variance r

According to the noise definition, the system model can be expressed in a form of normal pfd as follows

$$f(y_t | \psi_t, \Theta) = \frac{1}{\sqrt{2\pi}} r^{-0.5} \exp \left\{ -\frac{1}{2r} (y_t - \psi_t' \theta)^2 \right\} \quad (3)$$

5.2.1 Statistics Recursion

For the purposes of further calculations, it is convenient to formally rewrite the model into the form

$$f(y_t | \psi_t, \Theta) = \frac{1}{\sqrt{2\pi}} r^{-0.5} \exp \left\{ -\frac{1}{2r} [-1 \ \theta'] D_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right\} \quad (4)$$

- $D_t = \begin{bmatrix} y_t \\ \psi_t \end{bmatrix} [y_t \ \psi_t']$ is the information matrix

To enable the use of the Bayes rule for the purposes of statistics recursion, it is necessary to choose the prior pdf in a specific analytical form, so that by multiplying it with the model structure, the analytical form will reproduce itself. Such a prior pdf is called conjugated distribution [15].

Conjugated distribution to the normal is the inverse Gauss-Wishart distribution. It is obtained as the distribution of likelihood function of the model.

$$f(\Theta|d(t)) \propto r^{-0,5 \kappa_t} \exp \left\{ -\frac{1}{2r} [-1 \ \theta'] V_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right\} \quad (5)$$

- κ_t is the distribution statistics defining the number of data vectors, which have been utilized. It holds $\kappa_t = \kappa_{t-1} + 1$.
- V_t is a symmetrical and positive definite matrix, so called the information matrix. It holds $V_t = V_{t-1} + \Psi_t \Psi_t'$ and can be further break down into submatrices

$$V_t = \begin{bmatrix} V_y & V_{y\psi}' \\ V_{y\psi} & V_\psi \end{bmatrix}$$

After adding the above expressions into the Bayes formula, we obtain the statistics recursion

$$\underbrace{r^{-0,5 \kappa_t} \exp \left\{ -\frac{1}{2r} [-1 \ \theta'] V_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right\}}_{\text{posterior pdf}} \propto \underbrace{r^{-0,5} \exp \left\{ -\frac{1}{2r} [-1 \ \theta'] D_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right\}}_{\text{model}} \underbrace{r^{-0,5 \kappa_{t-1}} \exp \left\{ -\frac{1}{2r} [-1 \ \theta'] V_{t-1} \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right\}}_{\text{prior pdf}} \quad (6)$$

$$V_t = V_{t-1} + D_t \quad (7)$$

$$\kappa_t = \kappa_{t-1} + 1 \quad (8)$$

5.2.2 Point Estimate of Parameters

In order to estimate the maximum of the likelihood function and consequently the parameter θ , the Equation 5 is differentiated with respect to θ and equated to zero

$$\frac{\partial f(\{\theta, r\}|d(t), r)}{\partial \theta} = r^{-0,5 \kappa_t} \exp \left\{ -\frac{1}{2r} [-1 \ \theta'] V_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right\} \left(-\frac{1}{2r} \right) (-2V_{y\psi} + 2V_\psi \theta) = 0 \quad (9)$$

It follows

$$\hat{\theta} = V_\psi^{-1} V_{y\psi} \quad (10)$$

After adding the point estimate $\hat{\theta}$ into the Equation 5 we get

$$f(r|d(t)) \propto r^{-0,5 \kappa_t} \exp \left\{ \frac{V_y - V_{y\psi}' V_\psi^{-1} V_{y\psi}}{2r} \right\} \quad (11)$$

In the next step the Equation 11 is differentiated with respect to r and equated to zero

$$\frac{\partial f(r|d(t))}{\partial r} = -\kappa_t \frac{1}{2r} + \frac{V_y - V_{y\psi} V_{\psi}^{-1} V_{y\psi}}{2r} = 0 \quad (12)$$

It follows

$$\hat{r}_t = \frac{V_y - V_{y\psi} V_{\psi}^{-1} V_{y\psi}}{\kappa_t} \quad (13)$$

5.2.3 Point Estimate of Output

The point estimate of output, i.e. the mean value of y_t , is derived directly by substituting the point estimate of parameters into model.

$$\hat{y}_t = E[y_t | \psi_t, d(t-1)] = E[\psi_t' \hat{\theta}_t + e_t] = \psi_t' \hat{\theta}_t \quad (14)$$

5.3 Discrete Model

In the case of all the input variables acquiring a finite number of values, we are talking about a discrete model. A discrete system can be characterized by conditional probabilities, which are assigned to every configuration of values taken by explanatory variables. Such a model of discrete system is described by categorical probability distribution [15]

$$f(y_t | \psi_t, \Theta) = \Theta_{y_t | \psi_t} \quad (15)$$

- $y_t | \psi_t = \{y_t, \psi_{1t}, \psi_{2t}, \dots, \psi_{nt}\}$ is the vector of regression
- $\Theta_{y_t | \psi_t}$ is the vector of probabilities

5.3.1 Statistics Recursion

For the purposes of further calculations, it is convenient to formally rewrite the model into the multiplication form

$$f(y | \psi, \Theta) = \prod_{i \in y^*} \prod_{\varphi \in \Psi^*} \Theta_{y | \psi}^{\delta(i|\varphi, y_t | \psi_t)} \quad (16)$$

- $\delta(i|\varphi, y_t | \psi_t)$ is the Kronecker function, which is equal to one when $i|\varphi = y_t | \psi_t$ and equal to zero otherwise
- y^*, Ψ^* are sets of dependent and explanatory variables

Point estimate of model parameters can be defined as a conditional mean value of parameter $\theta_{y|\psi}$. To be able to mathematically express this value, distribution function of parameter $\theta_{y|\psi}$ has to be conjugated to the categorical distribution. An example of such a conjugated distribution is the Dirichlet distribution (Equation 17, 18, 19 and 20) [15].

$$f(\Theta|d(t)) = \frac{1}{B(v_t)} \prod_{i \in y^*} \prod_{\varphi \in \Psi^*} \Theta_{i|\varphi}^{v_{i|\varphi,t}} \quad (17)$$

- $B(v)$ is the generalized beta function

$$B(v) = \prod_{\varphi \in \Psi^*} \frac{\prod_{i \in y^*} \Gamma(v_{i|\varphi})}{\Gamma(\sum_{i \in y^*} v_{i|\varphi})} \quad (18)$$

- $\Gamma(x)$ is the gama function defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt \quad (19)$$

$$\Gamma(x + 1) = x \Gamma(x), x \in R^+ \quad (20)$$

After adding the above expressions into the Bayes formula, the statistics recursion is obtained in a form

$$\begin{aligned} f(\Theta|d(t)) &\propto \underbrace{\prod_{i|\varphi \in \Psi^*} \Theta_{i|\varphi}^{\delta(i|\varphi, y_t|\psi_t)}}_{model} \underbrace{\prod_{i|\varphi \in \Psi^*} \Theta_{i|\varphi}^{v_{i|\varphi;t-1}}}_{prior\ pdf} = \\ &= \underbrace{\prod_{i|\varphi \in \Psi^*} \Theta_{i|\varphi}^{\delta(i|\varphi, y_t|\psi_t) + v_{i|\varphi;t-1}}}_{posterior\ pdf} = \prod_{i|\varphi \in \Psi^*} \Theta_{i|\varphi}^{\overbrace{v_{i|\varphi;t}^{new\ statistics}}} \end{aligned} \quad (21)$$

$$v_{i|\varphi;t} = \delta(i|\varphi, y_t|\psi_t) + v_{i|\varphi;t-1} \quad (22)$$

5.3.2 Point Estimate of Parameters

Point estimate of model parameters is defined as a conditional mean value of parameter $\theta_{y|\psi}$ [15]. Detailed calculation of the conditional mean value of parameter $\theta_{y|\psi}$ is presented below.

$$\begin{aligned}
\hat{\Theta}_{y|\psi} &= E[\Theta_{y|\psi} | d(t)] = \int_0^\infty \Theta_{y|\psi} f(\theta | d(t)) d\theta = \frac{1}{B(\nu)} \int_0^\infty \Theta_{y|\psi} \prod_{i \in y^*} \prod_{\varphi \in \Psi^*} \Theta_{i|\varphi}^{\nu_{i|\varphi}} d\theta = \\
&= \frac{1}{B(\nu)} \int_0^\infty \prod_{i \in y^*} \prod_{\varphi \in \Psi^*} \Theta_{i|\varphi}^{\nu_{i|\varphi} + \delta(i|\varphi, y|\psi)} d\theta = \frac{1}{\prod_{\varphi \in \Psi^*} B(\nu)} \prod_{\varphi \in \Psi^*} \int_0^\infty \prod_{i \in y^*} \Theta_{i|\varphi}^{\nu_{i|\varphi} + \delta(i|\varphi, y|\psi)} d\Theta_{y|\psi} = \\
&= \frac{B(\nu_\psi + \delta(i, y))}{B(\nu_\psi)} = \frac{\prod_{\varphi \in \Psi^*} \frac{\prod_{i \in y^*} \Gamma(\nu_{i|\varphi} + \delta(i, y))}{\Gamma(\sum_{i \in y^*} \nu_{i|\varphi} + 1)}}{\prod_{\varphi \in \Psi^*} \frac{\prod_{i \in y^*} \Gamma(\nu_{i|\varphi})}{\Gamma(\sum_{i \in y^*} \nu_{i|\varphi})}} = \frac{\nu_{i|\varphi}}{\sum_{i \in y^*} \nu_{i|\varphi}} = \frac{\nu_{i|\varphi}}{\sum_{i \in y^*} \nu_{i|\varphi}} \\
\hat{\Theta}_{y|\psi} &= \frac{\nu_{y|\psi, t}}{\sum_{i \in y^*} \nu_{i|\psi, t}}, \forall y \in y^*, \forall \psi \in \psi^* \tag{23}
\end{aligned}$$

Expression of the parameters point estimate (Equation 23) can be viewed as a mere normalization of ν statistics, so that sum of probabilities is equal to one.

5.3.1 Point Estimate of Output

As a point estimate of output y is considered \hat{y} corresponding to the maximal $\hat{\theta}_{y|\psi}$ of the particular value of ψ .

5.4 Logistic Model

Logistic regression is a method with an extremely wide range of uses, especially popular in traffic demand modelling. Contrary to the discrete model, logistic regression can be utilized in cases where the modelled variable is dependent on discrete, as well as continuous explanatory variables. Furthermore, the use of this method is favourable when all variables are discrete; however, they take a large number of values, and thus a purely discrete model would obtain an excessively high dimension [15].

Logistic model can be defined as follows

$$\psi_t \Theta + e_t = \text{logit}(p) \tag{24}$$

- p is the probability $P(y = 1 | \psi)$
- $\psi = \{1, \psi_{1t}, \psi_{2t}, \dots, \psi_{nt}\}$ is the vector of regression
- $\Theta = \{\theta, r\}$ is the model parameter carrying the information about the vector of regression coefficients $\theta = \{\theta_0, \theta_1, \dots, \theta_n\}$ and the noise variance r
- e_t is noise the white noise

As can be observed from Equation 24, probabilities acquiring values on interval (0, 1) are transformed by *logit* function on a real axis. The real values are further expressed by linear regression which utilizes vector of regression ψ . The meaning of *logit* function is graphically presented in Figure 2. On the contrary, inverse *logit* function is utilized in order to transform real numbers into probabilities (Figure 3).

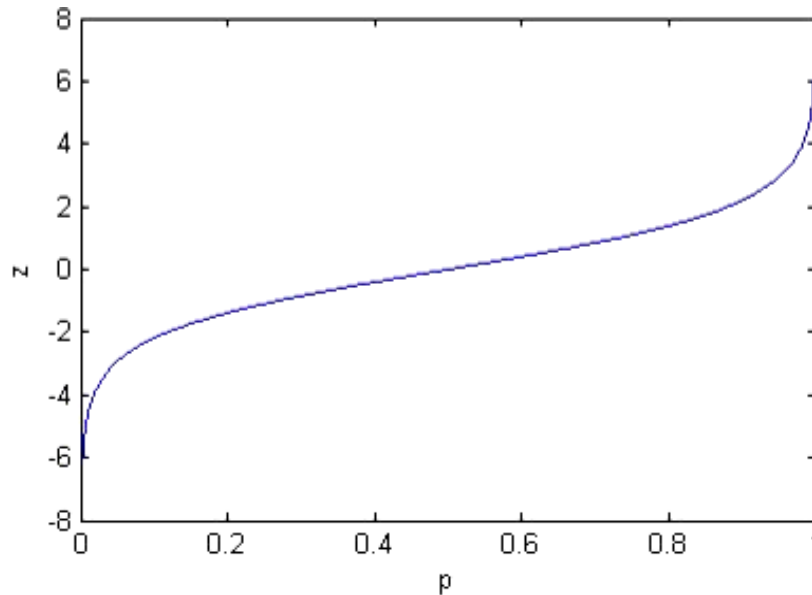


Figure 2: Graphical presentation of the *logit* function: $z = \ln p/(1-p)$

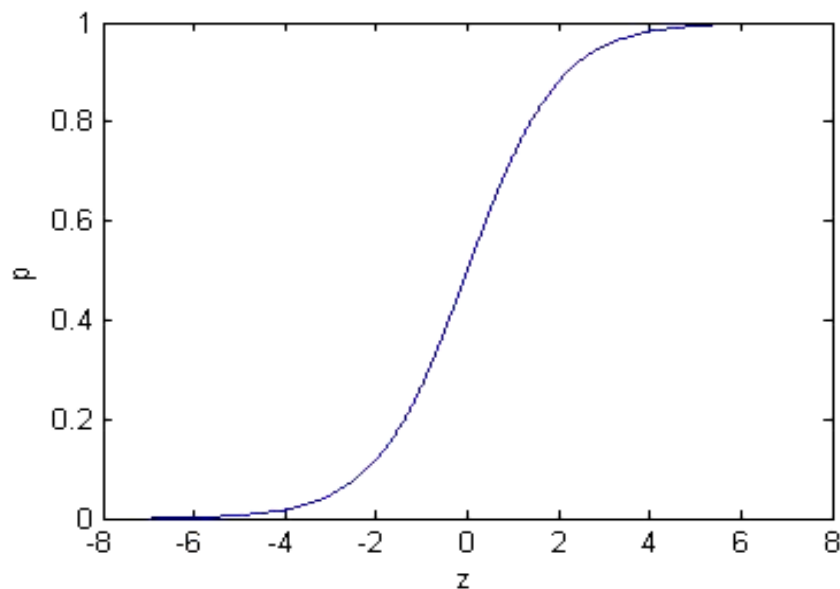


Figure 3: Graphical presentation of the inverse *logit* function: $p = \exp(z)/(1+\exp(z))$

The inverse *logit* function provides the possibility of other model interpretation. Considering Y can take on any of discrete values $\{y_1, \dots, y_K\}$, then the form of $P(Y = y_k | \psi_t, \theta)$ for $Y = y_1, Y = y_2, \dots, Y = y_{K-1}$ is

$$P(Y = y_k | \psi_t, \Theta) = \frac{\exp(\theta_{k0} + \sum_{i=1}^n \theta_{ki} \psi_{it})}{1 + \sum_{j=1}^{K-1} \exp(\theta_{j0} + \sum_{i=1}^n \theta_{ji} \psi_{it})} \quad (25)$$

When $Y = y_K$, it is

$$P(Y = y_K | \psi_t, \Theta) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\theta_{j0} + \sum_{i=1}^n \theta_{ji} \psi_{it})} \quad (26)$$

5.4.1 Point Estimate of Parameters

Unfortunately, sufficient conjugated distribution to the logistic model does not exist. Therefore, point estimate of model parameters is executed by the maximum likelihood method; specifically the conditional log likelihood $\ln L(\Theta)$ is utilized (Equation 27). Because it is a concave function, the gradient converges to a global maximum [16].

$$\ln L(\Theta) = \ln \prod_{\tau}^t f(y_{\tau} | \psi_{\tau}, \Theta) \quad (27)$$

After substitution and minor adjustments we obtain

$$\begin{aligned} \ln L(\Theta) &= \ln \prod_{\tau}^t \frac{\exp(y_{\tau}(\theta_0 + \sum_{i=1}^n \theta_i \psi_{i\tau}))}{1 + \exp(\theta_0 + \sum_{i=1}^n \theta_i \psi_{i\tau})} = \\ &= \sum_{\tau}^t \left[y_{\tau}(\theta_0 + \sum_{i=1}^n \theta_i \psi_{i\tau}) - \ln(1 + \exp(\theta_0 + \sum_{i=1}^n \theta_i \psi_{i\tau})) \right] \end{aligned} \quad (28)$$

In order to find the global maximum of Equation 27, the Newton method is used [16]. For more than one dimension, solution of the Newton method is defined as

$$\theta^{(n+1)} = \theta^{(n)} - H^{-1}(\theta^{(n)}) \nabla f(\theta^{(n)}) \quad (29)$$

- H is the Hessian matrix
- n represents each iteration, which converges the function to the global maximum

5.4.2 Point Estimate of Output

Point estimate of output $f(y|\psi, \hat{\theta}_t)$ is obtained by adding the point estimate of parameters $\hat{\theta}_t$ in the model (Equation 25 and 26). For each regression vector ψ we obtain

$$P(Y = y_k|\psi, \hat{\theta}_t) = \frac{\exp(\hat{\theta}_{k0} + \sum_{i=1}^n \hat{\theta}_{ki}\psi'_i)}{1 + \sum_{j=1}^{K-1} \exp(\hat{\theta}_{j0} + \sum_{i=1}^n \hat{\theta}_{ji}\psi'_i)} \quad (30)$$

$$P(Y = y_k|\psi, \hat{\theta}_t) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\hat{\theta}_{j0} + \sum_{i=1}^n \hat{\theta}_{ji}\psi'_i)} \quad (31)$$

5.5 Mixture Model

As was mentioned earlier, mixture model consists of a set of ordinary models and a pointer model, which at each time instant differentiates the active component.

In Bayesian statistics component model can be expressed as

$$f(d_t|c_t, \theta_c, \psi_t), c \in \{1, 2, \dots, n_c\} = c^* \quad (32)$$

- d_t are the modelled data
- c defines the particular component
- θ_c are parameters of the c^{th} component
- ψ_t is the regression vector

Pointer model is defined as a discrete model

$$f(c_t|\alpha, d(t-1)) = f(c_t|\alpha) = \alpha_{c_t} \quad (33)$$

$$\alpha \in \alpha^* = \left\{ \alpha_{i|j} \geq 0, \forall i, j \in c^*; \sum_{i \in c^*} \alpha_{i|j} = 1, \forall j \in c^* \right\} \quad (34)$$

- α is the parameter of pointer

5.5.1 Derivation of Mixture Estimation

The estimation of mixture model is derived from the posterior probability density function of all unknown parameters conditioned on data, i.e. from

$$\begin{aligned}
 & f(d_t, c_t, \alpha, \theta | d(t-1)) = \\
 & = f(d_t | c_t, \alpha, \theta, d(t-1)) f(c_t | \alpha, \theta, d(t-1)) f(\alpha | d(t-1)) f(\theta, d(t-1)) = \\
 & = \underbrace{f(d_t | c_t, \theta, \psi_t)}_I \underbrace{f(c_t | \alpha)}_{II} \underbrace{f(\alpha | d(t-1))}_{III} \underbrace{f(\theta | d(t-1))}_{IV}
 \end{aligned} \tag{35}$$

- *I* is the model of the c^{th} component
- *II* is the model of pointer
- *III* is the prior for the component model estimation
- *IV* is the prior for the pointer model estimation

Notice that the model of the c^{th} component is not dependent on the parameter α and the historical data, and the pointer model does not depend on data at all (Equation 35). Furthermore, the parameters α and θ are mutually independent [17].

In order to update the Equation 35, i.e. to calculate $f(\alpha, \theta | d(t))$ from $f(\alpha | d(t-1))$ and $f(\theta | d(t-1))$, marginalization of the pdf $f(d_t, c_t, \alpha, \theta | d(t))$ is computed as

$$f(\alpha, \theta | d(t)) = \sum_{c_t} f(d_t, c_t, \alpha, \theta | d(t)) \tag{36}$$

As shown in Equation 36, the posterior pdf has a form of a sum. When repetitively multiplying it with the model according to the Bayes rule, the computation became unfeasible. Hence the appropriate approximation has to be applied [17].

For this purpose the active component is considered as known and thus the pdf of pointer model can be approximated by Kronecker function

$$\delta(c, c_t), c \in c^* \tag{37}$$

- c_t is the active component at the time instant t

In fact, the assumption of the knowledge of active component is not fulfilled, which is why the expectation of $\delta(c, c_t)$ is used instead using its value [17].

$$\delta(c, c_t) \rightarrow E[\delta(c, c_t) | d(t)] = \sum_{c=1}^{n_c} \delta(c, c_t) f(c_t | d(t)) = P(c_t = c | d(t)) = f(c_t | d(t)) \tag{38}$$

Denote

$$f(c_t|d(t)) = w_{c,t} \quad (39)$$

- $w_{c,t} = [w_{1,t}, w_{2,t}, \dots, w_{n_c,t}]$ is the weighting vector defining the probability of particular component being active at time t

It holds

$$w_{c,t} \geq 0, \forall c_t \in c^* \quad (40)$$

$$\sum_{c_t \in c^*} w_{c,t} = 1, \forall t \in t^* \quad (41)$$

It is important to emphasize that the task of pdf $f(c_t|d(t))$ computation is closely related to the estimation of pointer model and component model parameters.

The estimation of weights $w_{c,t}$ is possible when all the aprior pdfs are known as follows

$$\begin{aligned} w_{c,t} = f(c_t|d(t)) &= \int_{\theta^*} \int_{\alpha^*} f(c_t, \alpha, \theta|d(t)) d\alpha d\theta = \\ &= \int_{\theta^*} \underbrace{f(d_t | c_t, \theta_c, \psi_t)}_{\mathcal{A}_{c_t}^\theta} \underbrace{f(\theta|d(t-1))}_{\mathcal{B}_{c_t}^\theta} d\theta \int_{\alpha^*} \underbrace{\alpha_{c_t} f(\alpha|d(t-1))}_{\mathcal{C}_{c_t}^\alpha} d\alpha \underbrace{d\alpha}_{\mathcal{D}_{c_t}^\alpha} \end{aligned} \quad (42)$$

- $\mathcal{A}_{c_t}^\theta$ is the posterior pdf for the estimation of θ
- $\mathcal{B}_{c_t}^\theta$ represents the prediction from the component c_t
- $\mathcal{C}_{c_t}^\alpha$ is the posterior pdf for the estimation of α
- $\mathcal{D}_{c_t}^\alpha$ represents the prediction from the pointer model

5.5.2 Estimation of Parameter θ

As was suggested before, parameter θ can be estimated from the component model as

$$f(\theta|d(t)) \propto f(d_t | c_t, \theta_c, \psi_t) f(\theta|d(t-1)) = \mathcal{A}_{c_t}^\theta \quad (43)$$

The model is characterized by normal probability distribution. As was comprehensively explained in Subchapter 5.2.1, to enable the numerable solution of the task, pdf of the model is considered in a form of the inverse Gauss-Wishart distribution [17]. Then the model is expressed in a multiplication form

$$\prod_{c=1}^{n_c} GiV(V_{c;t}, \kappa_{c;t}) \propto \prod_{c=1}^{n_c} f(d_t | c, \psi_t, \theta_c)^{\delta(c;c_t)} \prod_{c=1}^{n_c} GiV(V_{c;t-1}, \kappa_{c;t-1}) \quad (44)$$

After approximation $w_{c,t} \rightarrow \delta(c, c_t)$, the statistics V and κ are defined as follows

$$V_{c;t} = V_{c;t-1} + w_{c;t} \Psi_t \Psi_t' \quad (45)$$

$$\kappa_{c;t} = \kappa_{c;t-1} + w_{c;t} \quad (46)$$

The point estimate of parameter θ as well as the aposterior pdf can be derived directly from these statistics.

5.5.3 Estimation of Parameter α

The estimation of parameter α is analogical to the previously described procedure. Parameter α is estimated from the pointer model as

$$f(\alpha|d(t)) \propto \alpha_{c_t} f(\alpha|d(t-1)) = \mathcal{C}_{c_t}^\alpha \quad (47)$$

The pdf of parameter $w_{c,t}$ is considered in a form of conjugated Dirichler function (see Subchapter 5.3.1). The model is expressed in a multiplication form as

$$\sum_{c=1}^{n_c} \alpha_c^{v_{c,t}} \propto \sum_{c=1}^{n_c} \alpha_c^{\delta(c,c_t)} \sum_{c=1}^{n_c} \alpha_c^{v_{c,t-1}} \quad (48)$$

After approximation $w_{c,t} \rightarrow \delta(c, c_t)$, the statistics v is defined as

$$v_{c;t} = v_{c;t-1} + w_{c;t} \quad (49)$$

5.5.4 Estimation of Weighting Vector w

The estimation of weighting vector $w_{c,t}$ has already been discussed in Subchapter 5.5.1 (see Equations 42). The vector is expressed as

$$w_{c,t} = \mathcal{B}_{c_t}^\theta \mathcal{D}_{c_t}^\alpha \quad (50)$$

Because computation of the component likelihood $\mathcal{B}_{c_t}^\theta$ is quite problematic, the task is being simplified by utilizing the point estimate of parameter θ .

$$\mathcal{B}_{c_t}^\theta = f(d_t|c, \psi_t, \hat{\theta}_{c;t-1}) \equiv \hat{m}_{c;t} \quad (51)$$

The integral $\mathcal{D}_{c_t}^\alpha$ is being expressed as the point estimate of parameter α .

$$\mathcal{D}_{c_t}^\alpha = \int_{\alpha^*} \alpha_c f(\alpha|d(t-1)) d\alpha = \frac{v_{c;t-1}}{\sum_{i \in c^*} v_{i;t-1}} = \hat{\alpha}_{c;t-1} \quad (52)$$

From the above mentioned we obtain

$$w_{c,t} = \mathcal{B}_{c_t}^\theta \mathcal{D}_{c_t}^\alpha = \hat{m}_{c;t} \hat{\alpha}_{c;t-1} \quad (53)$$

6 SOFTWARE PROCESSING

All the programmes presented in this chapter were created in cooperation with the thesis' supervisor doc. Ing. Ivan Nagy, CSc. They stem from the theoretical background described in Chapter 5 and were computed in the Scilab programming environment version 5.4.1.

6.1 Estimation of Mixture Model

The following program is used for the purposes of mixture model estimation, i.e. the estimation of static components as well as the pointer model, and consequently for the purposes of data classification.

```
exec("ScIntro.sce",-1),mode(0),funcprot(0);
getd('func');

nd=4075; // NUMBER OF DATA
nc=3; // NUMBER OF COMPONENTS
nv=14; // NUMBER OF VARIABLES
t_est=5000; // ESTIMATION OF NOISE COVARIANCES

// REGRESSION COEFFICIENTS =====
Sim.nc=nc;
for i=1:nc
    Sim.Cy(i).th=rand(nv,1); // INITIAL COMPONENTS' CENTERS
    Sim.Cy(i).sd=0.1*eye(nv,nv); // COMPONENTS' COVARIANCES
end

// SIMULATED NOISE COVARIANCES =====
Sim.Cp.th=fnorm(rand(1,nc,'u')+1,2);
Sim.ct(1)=1; // INITIAL POINTER

// INITIAL PARAMETERS =====
a=.8; // STANDARD DEVIATION OF INITIAL
for j=1:nc // PARAMETERS SCATTERING
    [mr,mc]=size(Sim.Cy(j).th);
    Ps=[Sim.Cy(j).th;1]; // INITIAL PARAMETERS
    Est.Cy(j).V=Ps*Ps'; // STATISTICS
    Est.Cy(j).th=Sim.Cy(j).th+a*rand(nv,1,'n'); // ESTIMATE OF REGRESSION COEFFICIENT
    Est.Cy(j).sd=.1*eye(nv,nv); // STANDARD DEVIATION
end
Est.ka=ones(1,nc); // COUNTER
Est.Cp.V=ones(1,nc); // POINTER STATISTICS
Est.Cp.th=fnorm(ones(1,nc)); // POINTER PARAMETER
w=fnorm(ones(1,nc)); // WEIGHTS

// DATA =====
load data.dat // DATA LOADING
//nd=size(data,1);
dt=data(1:nd, 3:14)';
ddt=data(1:nd, 3:14)';
for i=1:size(dt,1)
    dt(i,:)=(dt(i,:)-mean(dt(i,:)))/stdev(dt(i,:)); // DATA STANDARDIZATION
end
Sim.yt=dt;
```

```

// ESTIMATION =====
printf('running .....|\n '),itime=0;
for t=1:nd
    itime=itime+1; if itime>(nd-1)/20, mprintf('.', itime=0); end

    for j=1:nc
        [xxx,G(j)]=GaussN(Sim.yt(:,t),Est.Cy(j).th,Est.Cy(j).sd); // LIKELIHOOD
    end
    Lq=G-max(G);
    q=exp(Lq);

    ww=q'.*Est.Cp.th;
    w=ww/sum(ww); // GENERATION OF WEIGHTS
    wt(:,t)=w';

// UPDATE OF STATISTICS =====
Ps=[Sim.yt(:,t)' 1]; // EXTENDED REGRESSION VECTOR
for i=1:nc
    Est.Cy(i).V=Est.Cy(i).V+w(i)*Ps*Ps; // INFORMATION MATRIX
    Est.ka(i)=Est.ka(i)+w(i); // COUNTER
    Est.Cp.V(i)=Est.Cp.V(i)+w(i); // POINTER STATISTICS

    Vyy=Est.Cy(i).V(1:nc,1:nc);
    Vy=Est.Cy(i).V($,1:nc);
    V1=Est.Cy(i).V($,$);
    Est.Cy(i).th=inv(V1+1e-8*eye(V1))*Vy;
    Est.Cy(i).tht(:,t)=Est.Cy(i).th;
    if t>t_est
        Est.Cy(i).cv=(Vyy-Vy*inv(V1+1e-8*eye(V1))*Vy)/Est.ka(i);
    end
end
Est.Cp.th=fnorm(Est.Cp.V,2); // POINT ESTIMATE OF POINTER PARAMETERS
end
[sss,ct]=max(wt,'r');

// RESULTS =====
disp(Est.Cp.th,'pt.pars_est')

s=2:nd;
set(gcf(2),'position',[550 50 400 800])
for i=1:nc
    subplot(1,nc,i)
    plot(Est.Cy(i).tht')
    title('Component '+string(i))
    ylabel('theta');
    xlabel('d(t)');
end

dd=list();
jj=list();
ddd=list();
for i=1:max(ct)
    j=find(ct==i);
    jj(i)=j;
    dd(i)=dt(:,j);
    ddd(i)=ddt(:,j);
end

```

6.2 Linear Regression Model

The programme presented below is used for the purposes of the linear regression model estimation.

```

exec("ScIntro.sce",-1),mode(0)
getd("func")

// DATA =====
load data.dat                                     // DATA LOADING

for i=1:2
    data(:,i)=(data(:,i)-mean(data(:,i)))/stdev(data(:,i));    // DATA STANDARDIZATION
end

s=1:max(size(data));
y= data (s,1);
u1= data (s,2);
u2= data (s,3);
u3= data (s,4);
u4= data (s,5);
u5= data (s,6);
u6= data (s,7);
u7= data (s,8);
u8= data (s,9);
u9= data (s,10);
u10= data (s,11);
u11= data (s,12);

nd=length(y);
V=1e-8*eye(13);

// ESTIMATION =====
for t=1:nd
    Ps=[y(t)' u1(t)' u2(t)' u3(t)' u4(t)' u5(t)' u6(t)' u7(t)' u8(t)' u9(t)' u10(t)' u11(t)' 1];
                                                // EXTENDED REGRESSION VECTOR
    V=V+Ps'*Ps;                               // UPDATE OF STATISTICS
    Vy=V(1,1);                                 // PARTITIONING OF
    Vyp=V(2:$,1);                             INFORMATION MATRIX
    Vp=V(2:$,2:$);
    Eth=inv(Vp+1e-8*eye(Vp))*Vyp;              // POINT ESTIMATE OF REGRESSION
End                                              COEFFICIENTS

// SIMULATION =====
t=1:nd;
X=[u1(t)' u2(t)' u3(t)' u4(t)' u5(t)' u6(t)' u7(t)' u8(t)' u9(t)' u10(t)' u11(t)' ones(t)'];
                                                // REGRESSION VECTOR
Esty=X*Eth;                                    // OUTPUT
Ep=y'-Esty;                                    // PREDICTION ERROR
SE=sqrt(Ep'*Ep)/length(Ep);                   // SUM OF SQUARES OF PREDICTION ERROR

// RESULTS =====
disp('Estimated parameters')
Et=disp(Eth)

scf(1)
s=1:length(y);
plot(s,y(s),s,Esty(s));
set(gcf(),'position',[50 50 800 500]);
title('Linear Regression');
legend('Demand','Estimated Demand');
xlabel('d(t)');

```

6.3 Logistic Regression Model

The following programme is used for the purposes of the logistic regression model estimation.

```
exec("ScIntro.sce",-1),mode(0)
getd("func")

// DATA =====
load data.dat                                // DATA LOADING

nd=size(data,1);
y=data(1:nd,1)';
n=zeros(1,5);
z=ones(y);                                    // DISCRETIZATION
j=find(y>3);
n(2)=length(j);
z(j)=2;
j=find(y>12);
n(3)=length(j);
z(j)=3;
j=find(y>64);
n(4)=length(j);
z(j)=4;
j=find(y>313);
n(5)=length(j);
z(j)=5;
n(1)=nd-sum(n(2:5));
x=data(1:nd,2:12)';

// ESTIMATION =====
[Est,al]=lrLearn(z,x,2);                     // SUPERVISED LEARNING

al=n/sum(n);
ct=lrTest(x,Est,al,2);                       // CLASSIFICATION

printf(' Wrong %d from %d\n',sum(z~=ct),nd)

scf(1)
s=1:length(nd);
plot(s,z(s),'o:',s,ct(s),'o:');
set(gcf(),'position',[50 50 800 500]);
set(gca(),'data_bounds',[0 18 1 6])
title('Logistic Regression');
legend('Demand','Estimated Demand');
xlabel('d(t)');
```

6.4 Predefined Functions

To enable the execution of algorithms above, the knowledge of the inserted functions is necessary. The most important functions are presented in the following subchapters.

6.4.1 Likelihood Function

To define the value of multivariate Gaussian probability density function, i.e. the likelihood function, the following algorithm was created.

```
function [p, Lp]=GaussN(x, m, R)                // p - PROBABILITY
                                                // Lp - LOGARITHM OF PROBABILITY
x=x(:);                                       // REALIZATION
m=m(:);                                       // EXPECTATION
n=max(size(R));                               // R - COVARIANCE MATRIX
Lp=-.5*(n*log(2*%pi)+log(det(R)));
ex=(x-m)*inv(R+1e-8*eye(n,n))*(x-m);
Lp=Lp-.5*ex;
p=exp(Lp);

endfunction
```

6.4.2 Supervised Learning

Learning process of the logistic regression model was executed as follows

```
function [Est, al]=lrLearn(y, x, typ)

if argn(2)<3,
    typ='c';
end
if typ==1, typ='r'; end
if typ==2, typ='c'; end

nc=max(y); n1=min(y);
if n1~=1,
    disp('Error: y must start with 1');
    return
end
for i=1:nc
    c=find(y==i);
    Y=y(c);                               // y – CLASS LABEL
    if typ=='r'
        X=x(c,:);                         // x – DATA VECTOR
    else
        X=x(:,c);
    end
    th=mean(X,typ);                       // MEAN VALUE
    cv=cov(X,typ);                        // COVARIANCES
    if det(cv)<1e-5,
        cv=.1*eye(cv)+cv;
    end
    Est(i).th=th;
    Est(i).cv=cv;
end
ga=vals(y);
al=ga(2,:)/sum(ga(2,:));                 // STATIONARY PROBABILITIES

endfunction
```

6.4.3 Data Testing

During the classification task, data are tested with the model created in the process of supervised learning, which was described in Subchapter 6.4.2.

```
function ct=lrTest(x, Est, al, typ)

    if argn(2)<4,
        typ='c';
    end
    if typ==1, typ='r'; end
    if typ==2, typ='c'; end
    if typ=='r', x=x'; end

    nc=max(size(Est));
    nd=size(x,2);
    md=zeros(nc,nd);
    for t=1:nd
        xt=x(:,t);
        dL=zeros(1,nc);
        for i=1:nc
            [xxx dL(1,i)]=GaussN(xt,Est(i).th,Est(i).cv);
        end
        dL=dL-max(dL);
        d=exp(dL);
        mm=d.*al;
        md(:,t)=mm'/sum(mm);
    end
    [xxx ct]=max(md,'r');

endfunction
```

7 RESULTS

In this chapter, results obtained from the application of simple modelling methods, namely the simple linear and simple logistic regression analysis as well as results of the mixture model estimation, are presented.

7.1 Simple Linear Regression

Process of the regression model formulation consists of two phases. Firstly, it is necessary to define the relevant variables of the model. In order to do so, the dataset needs to be standardized as was described in Subchapter 5.1. In the second phase, where only relevant variables and non-standardized data are involved, a particular regression model is estimated.

7.1.1 Estimation of Relevant Variables

Coefficients obtained from the regression analysis, which was performed on the standardized dataset, are presented in Table 1. As can be observed here, except for one variable, all of the regression coefficients take on insignificant values. Thus, only revenue can be determined as the relevant variable.

Table 1: Regression coefficients obtained from the linear regression performed on the standardized data

Variable $x_i \in X$	Revenue	Fare	Distance	Origin Region			
				Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
Regression Coefficient	0.908169	0.000096	- 0.007543	0.054078	- 0.056003	- 0.017053	0.010485

Destination			
Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
0.051433	- 0.048633	- 0.014097	-0.00512

7.1.2 Estimation of Linear Regression Model

Considering that only revenue has the major impact on air traffic demand development, the model was computed as follows

$$y = 0.0043267x_1 + 97.510945 \quad (54)$$

$$PE = 68.95534152172598 \quad (55)$$

- PE is the sum of squares of prediction error

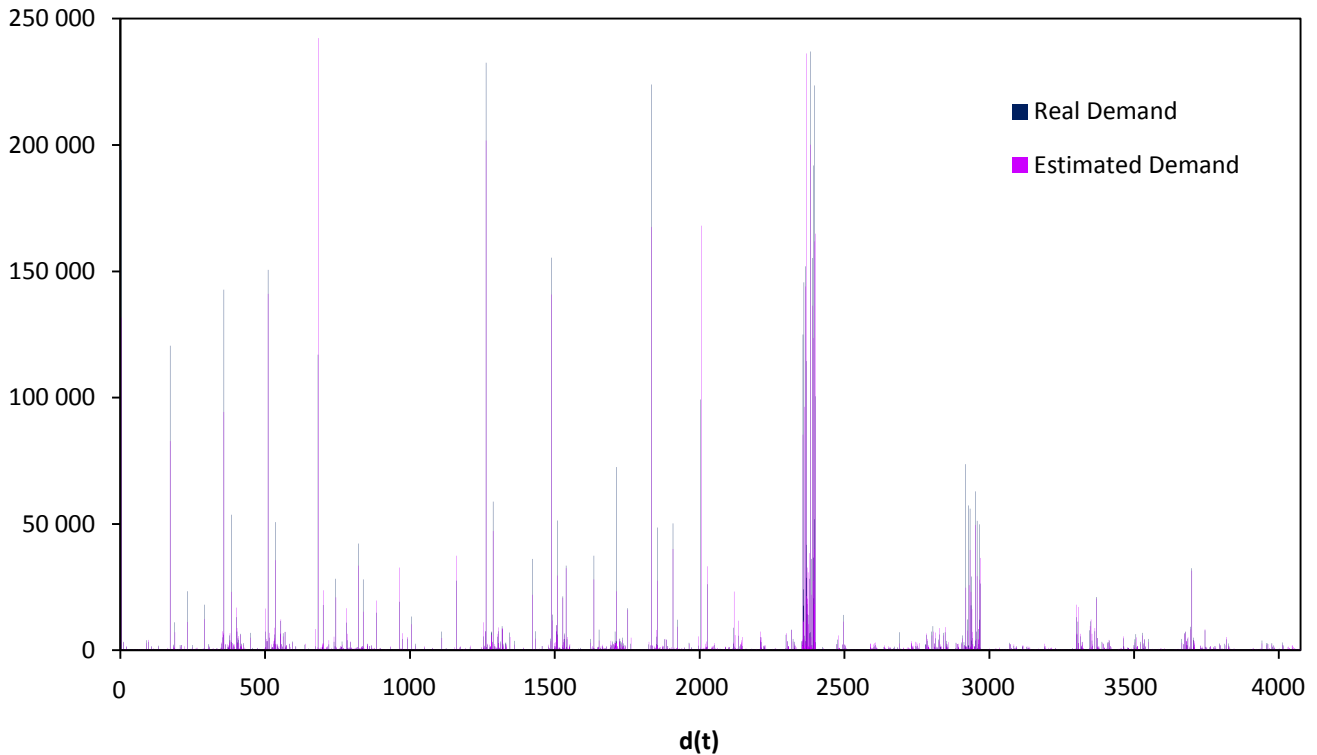


Figure 4: Comparison of the real demand and the demand estimated by the simple linear regression

The real demand as well as the demand estimated by the simple linear regression is presented in Figure 4.

7.2 Simple Logistic Regression

For the purposes of logistic regression, the output is discretized into five categories. The first category, i.e. extremely low demand, includes demand lower or equal to three passengers carried per year. The second one, i.e. low demand, contains values higher than three and lower or equal to twelve passengers per year. Demand higher than twelve and lower or equal to 64 passengers per year is considered in the third category, i.e. medium demand. The fourth category, i.e. high demand, includes demand between 64 and 313

passengers per year, whilst the rest of the data belongs to the fifth category, i.e. extremely high demand.

7.2.1 Estimation of Relevant Variables

In order to estimate the logistic regression model, the relevant variables already defined by regression analysis in Subchapter 7.1.1 are utilized.

7.2.2 Estimation of Logistic Regression Model

Figure 5 presents the real and the predicted demand. 1532 predictions from the dataset consisting of 4075 values do not match with the real demand.

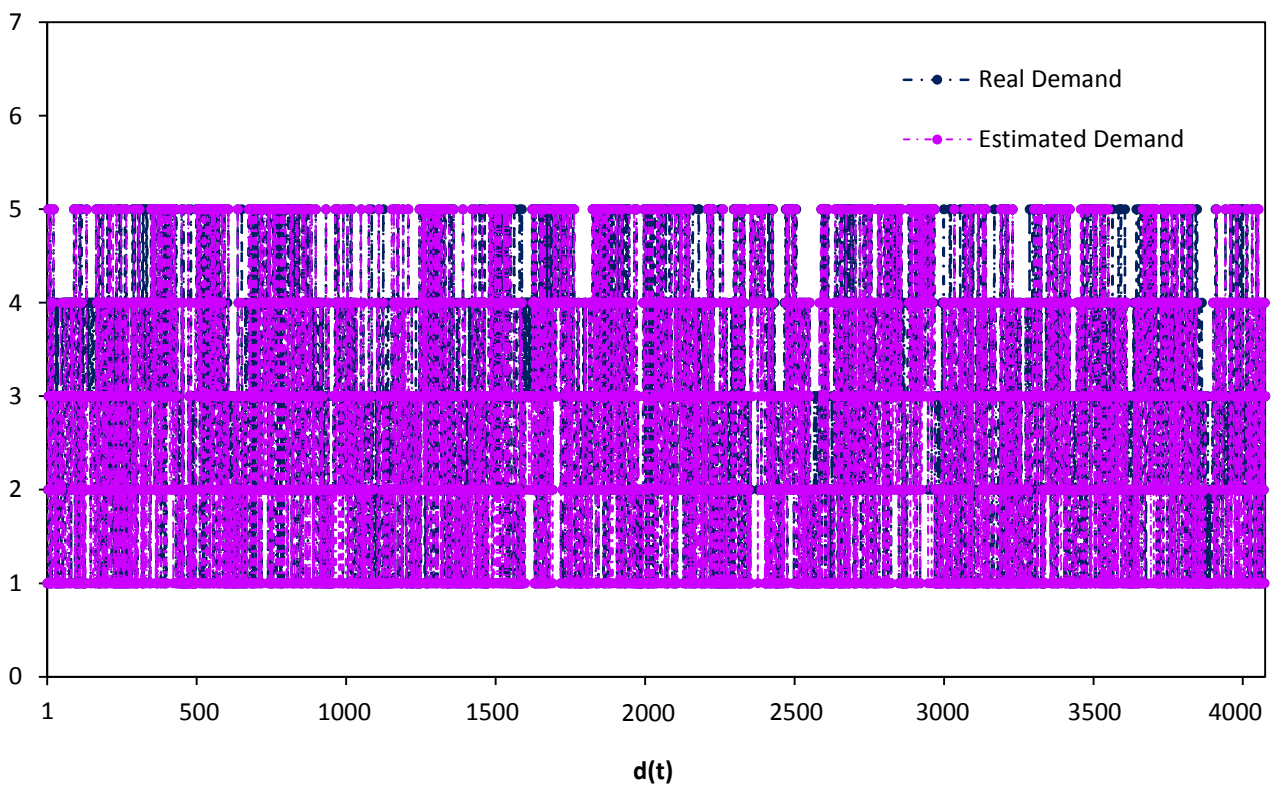


Figure 5: Comparison of the real demand and the demand estimated by the simple logistic regression

7.3 Mixture Model

As was explained in Subchapter 5.5, mixture model distinguishes the modes of the system and classifies data accordingly into particular components. Each component is further characterized by an independent model.

The formulation of initial parameters setting of the mixture model is a very complex task, which is crucial in the process of solution derivation. The mathematical procedure is not a subject of further discussion in the thesis. For the purposes of this work, initial parameters were defined according to expert knowledge. The estimation of noise covariance is defined as bigger than the number of data samples. In this way components are not getting larger during the statistics recursion, they are only allowed to move in the multidimensional space. Component covariance, representing the width of components, equals to 0.1. Standard deviation of scattering initial parameters is defined as 0.8. Furthermore, the initial position of components' centres is determined randomly, thus the repetitive computation provides more objective viewpoint of the solution variability.

7.3.1 Solution Variability

In order to investigate the solution variability, three different computations of pointer parameters were executed. In each of the calculations components' centres were determined randomly and thus the solutions are not identical.

Figures 6, 7 and 8 show the varying development of components' parameters during the three different computations in case all of the available variables are utilized. Table 2 presents the pointer parameters obtained from each of the calculations.

Table 2: Estimation of pointer parameters in three different calculations

	Pointer Parameter 1	Pointer Parameter 2	Pointer Parameter 3
1st Estimation	0.9040425	0.0046591	0.0912984
2nd Estimation	0.0046591	0.3724607	0.6228802
3rd Estimation	0.1186070	0.0046591	0.8767339

From the three different computations can be deduced that initial location of components' centres has a fundamental importance for the solution derivation. The results differ significantly, thus for an objective evaluation of a mixture estimate it is reasonable to conduct repetitive computations and to compare the results.

As presented in figures below, the appropriate estimation of noise covariance positively influenced the process of components' parameters derivation. Development of parameters is no longer dependent on the sequence of data samples, thus it is more effective.

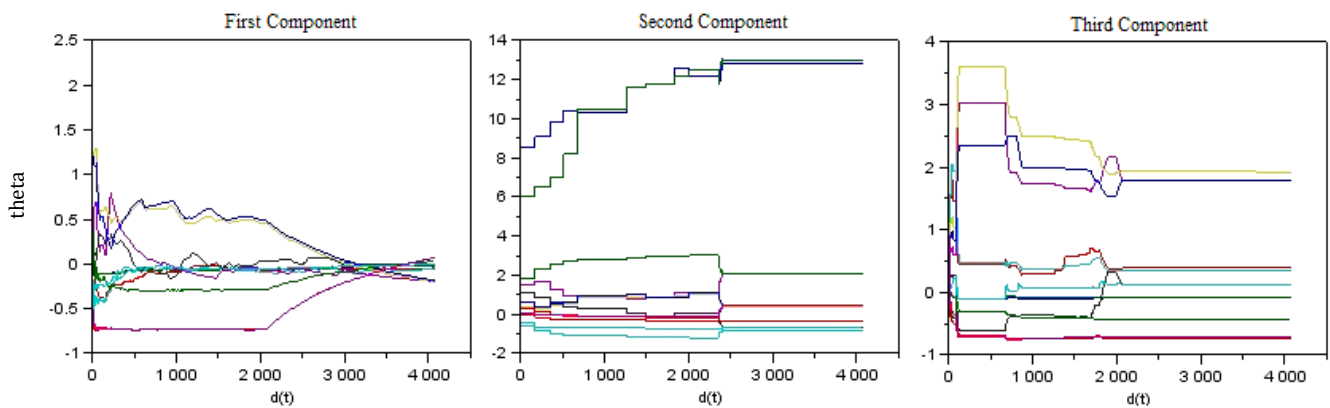


Figure 6: Development of the components' parameters during the statistics recursion in the first computation

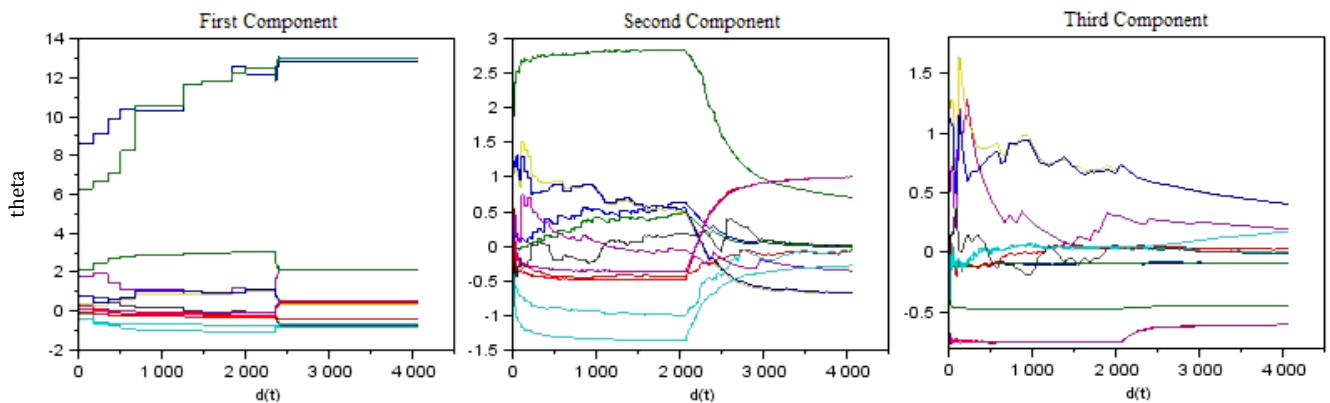


Figure 7: Development of the components' parameters during the statistics recursion in the second computation

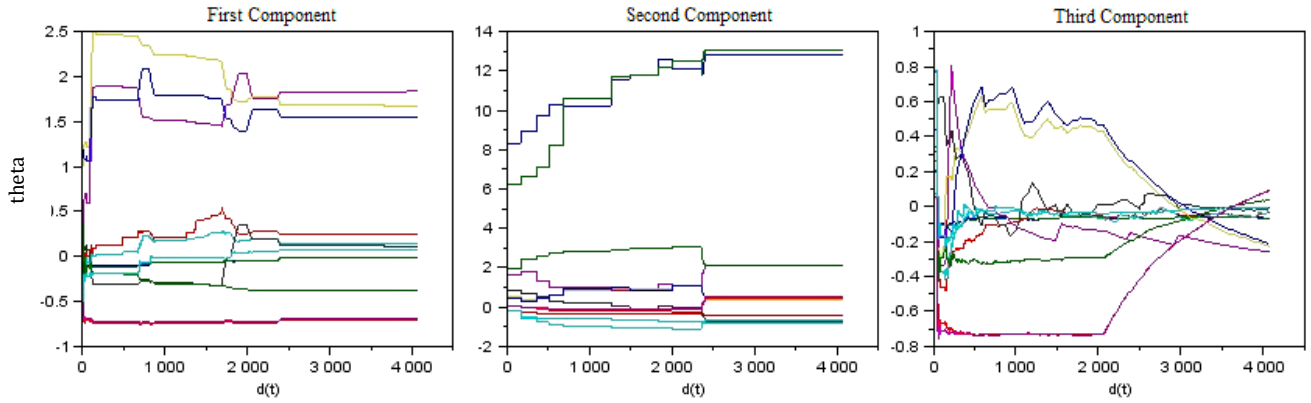


Figure 8: Development of the components' parameters during the statistics recursion in the third computation

7.3.2 Data Classification

As was mentioned in Chapter 5, during the process of data classification four different scenarios, all following different modelling approaches, were considered. Because the classification into three components did not always provide a sufficient level of estimate, data were further classified into five components. It is important to point out that the values of pointer parameter represent the percentages of data contained in the particular component.

Table 3: Data classification obtained from the utilization of a mixture model

	Number of Components	Pointer Parameter 1	Pointer Parameter 2	Pointer Parameter 3	Pointer Parameter 4	Pointer Parameter 5	Number of Relevant Components
1st Scenario	3	0.6208572	0.3744837	0.0046591	-	-	2
	5	0.1216452	0.2272309	0.0328380	0.6136290	0.0046569	3
2nd Scenario	3	0.9281506	0.0046635	0.0671859	-	-	1
	5	0.6987580	0.0046569	0.0216115	0.0730678	0.2019059	2
3rd Scenario	3	0.0620495	0.9324365	0.0055141	-	-	1
	5	0.4181858	0.0046513	0.5081731	0.0092148	0.0597749	2
4th Scenario	3	0.3809840	0.6138733	0.0051427	-	-	2
	5	0.3040073	0.4884203	0.0051451	0.1049064	0.0975209	3

As can be observed from Table 3, the first scenario, where all available variables are utilized, provides a sufficient level of classification in the case of three as well as five components involved. It implies that diffraction of the system into the estimated components can be beneficial for the purposes of mathematical modelling.

The second scenario, representing the modelling approach of the International Civil Aviation Organization, shows a higher level of system consistency, especially in the case of only three components being considered. However, in the case of classification into five components, the estimate did discover deeper system structures, and thus the one equation modelling approach proposed by The Manual for Air Traffic Forecasting is questionable.

A purely economic approach is presented in the third scenario, where only two variables, i.e. demand and price, are incorporated. This scenario like the previous one is characterized by a higher level of system consistency. However, in the case of favourable parameters setting, the system can also be divided into several components as can be observed in the Table 3.

The last scenario is based on our expert knowledge and experience with previously computed calculations. In this case diffraction of the system also seems to be highly beneficial.

Parameters of the first scenario with three components involved were evaluated as the most convenient and thus this scenario is used for the purposes of further calculations. The development of the components' parameters during the statistics recursion is presented in Figures 9, 10 and 11.

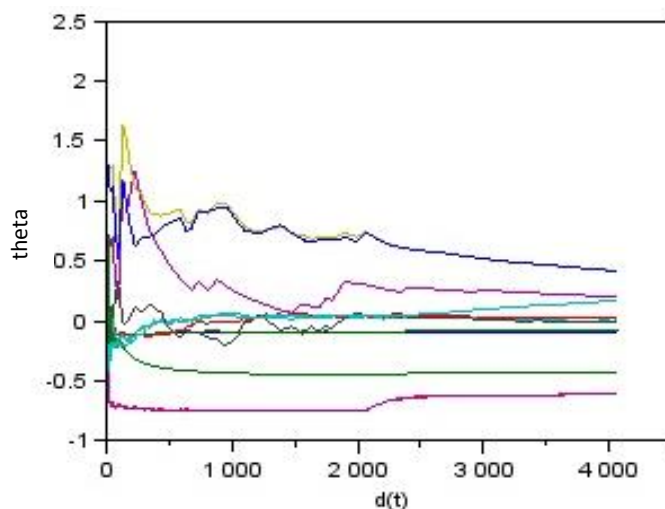


Figure 9: Development of the first component parameters during the statistics recursion of the first scenario

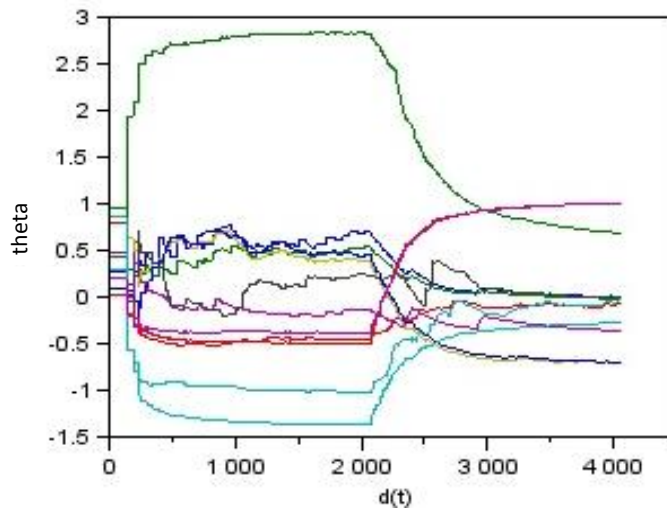


Figure 10: Development of the second component parameters during the statistics recursion of the first scenario

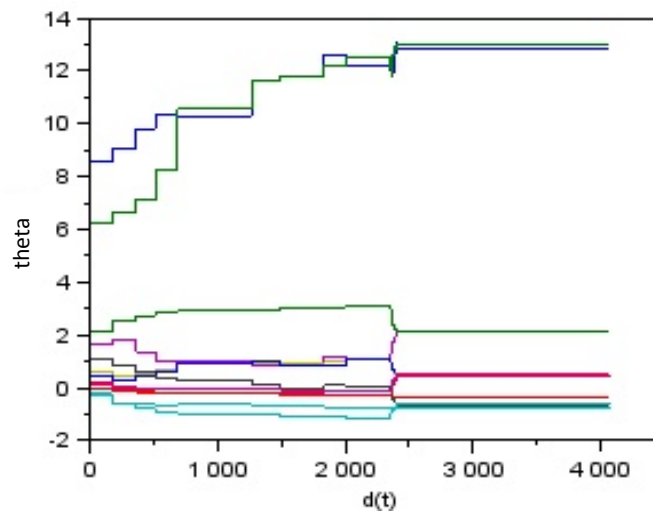


Figure 11: Development of the third component parameters during the statistics recursion of the first scenario

7.3.3 Linear Regression Analysis

Procedure of the regression analysis is similar to the one already presented in Subchapter 7.1. Firstly, the relevant variables are distinguished, thereafter the particular linear regression model is estimated.

7.3.3.1. Estimation of Relevant Variables

Table 4: Regression coefficients obtained from the standardized data of the first component

Variable $x_i \in X$	Revenue	Fare	Distance	Origin			
				Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
Regression Coefficient	0.922236	- 0.005790	0.017992	0.059426	- 0.074614	- 0.035038	0.019238

Destination			
Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
0.009872	- 0.004163	- 0.007290	-0.01820

In the first and second component models air traffic demand was found to be dependent only on one variable, which is revenue (Table 4 and 5). The same result was also obtained in the case of the simple linear regression; see Subchapter 7.1.1.

Table 5: Regression coefficients obtained from the standardized data of the second component

Variable $x_i \in X$	Revenue	Fare	Distance	Origin			
				Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
Regression Coefficient	0.875542	0.000631	- 0.003235	0.102362	0.075775	0.011891	-0.08072

Destination			
Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
0.022569	- 0.016106	0.001492	-0.01694

On the contrary, the third component model shows a well-balanced relevance of all the utilized variables (Table 6).

Table 6: Regression coefficients obtained from the standardized data of the third component

Variable $x_i \in X$	Revenue	Fare	Distance	Origin			
				Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
Regression Coefficient	1.226056	- 0.804571	- 1.234123	- 2.303143	1.126370	1.552037	-0.80478

Destination			
Population Density	GDP per Capita	Unemployment Rate	Monthly Nominal Wage
- 1.948942	1.597787	1.599510	-0.91473

7.3.3.2. Estimation of Linear Regression Models

Regression models of the first (Equation 56) and the second component (Equation 58), where demand is considered to be dependent only on revenue, as well as the model of the third component (Equation 60), where all the variables are included, were estimated as follows

$$y = 0.0040089x_1 - 14.191719 \quad (56)$$

$$PE = 13.98823700968075 \quad (57)$$

$$y = 0.0056442x_1 - 127.80321 \quad (58)$$

$$PE = 62.08225109159252 \quad (59)$$

$$y = 0.0051579x_1 - 747.12458x_2 - 194.94444x_3 - 59.004527x_4 + 3.4827018x_5 + 25053.798x_6 - 53.493786x_7 - 49.930191x_8 + 4.9403101x_9 + 25820.132x_{10} - 60.802573x_{11} + 631528.15 \quad (60)$$

$$PE = 2250.075166316838 \quad (61)$$

In Figures 12, 13 and 14, the real and the estimated demand are presented for each of the defined components.

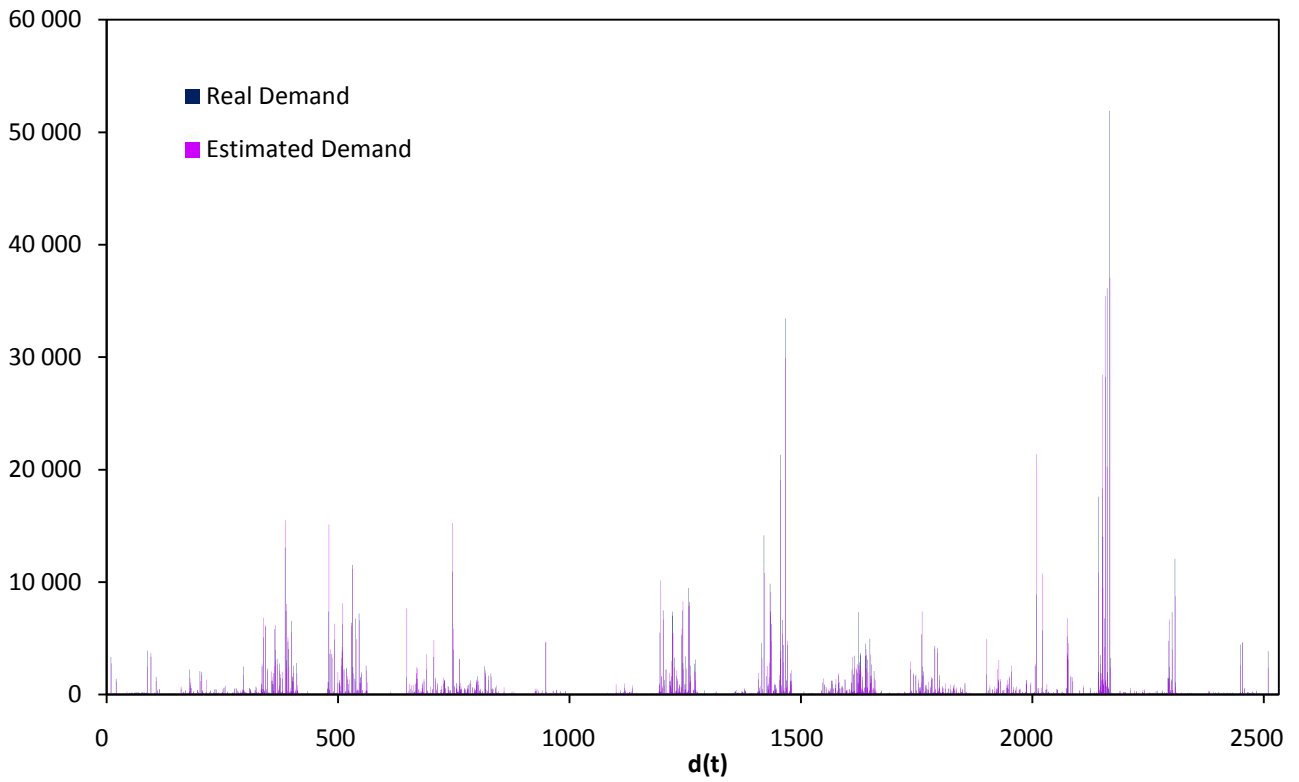


Figure 12: Comparison of the real demand and the demand estimated by the first component linear regression model

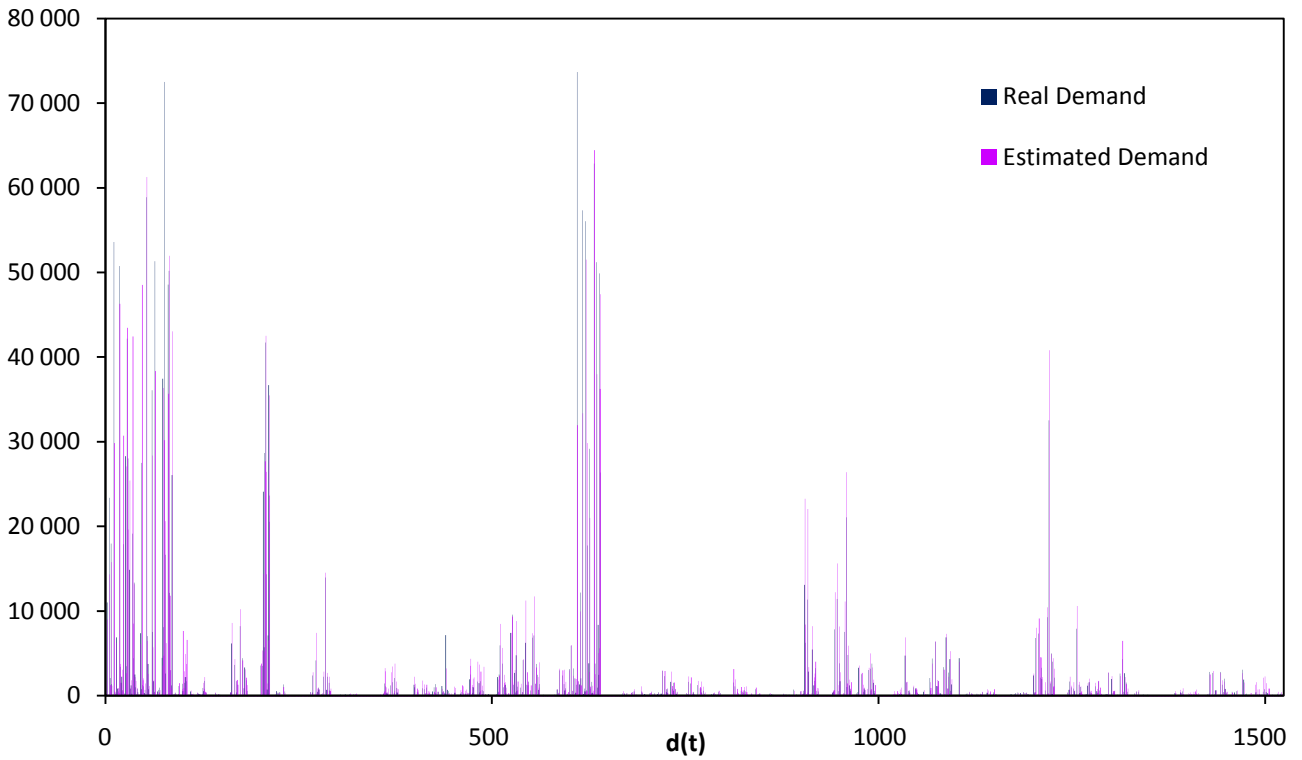


Figure 13: Comparison of the real demand and the demand estimated by the second component linear regression model

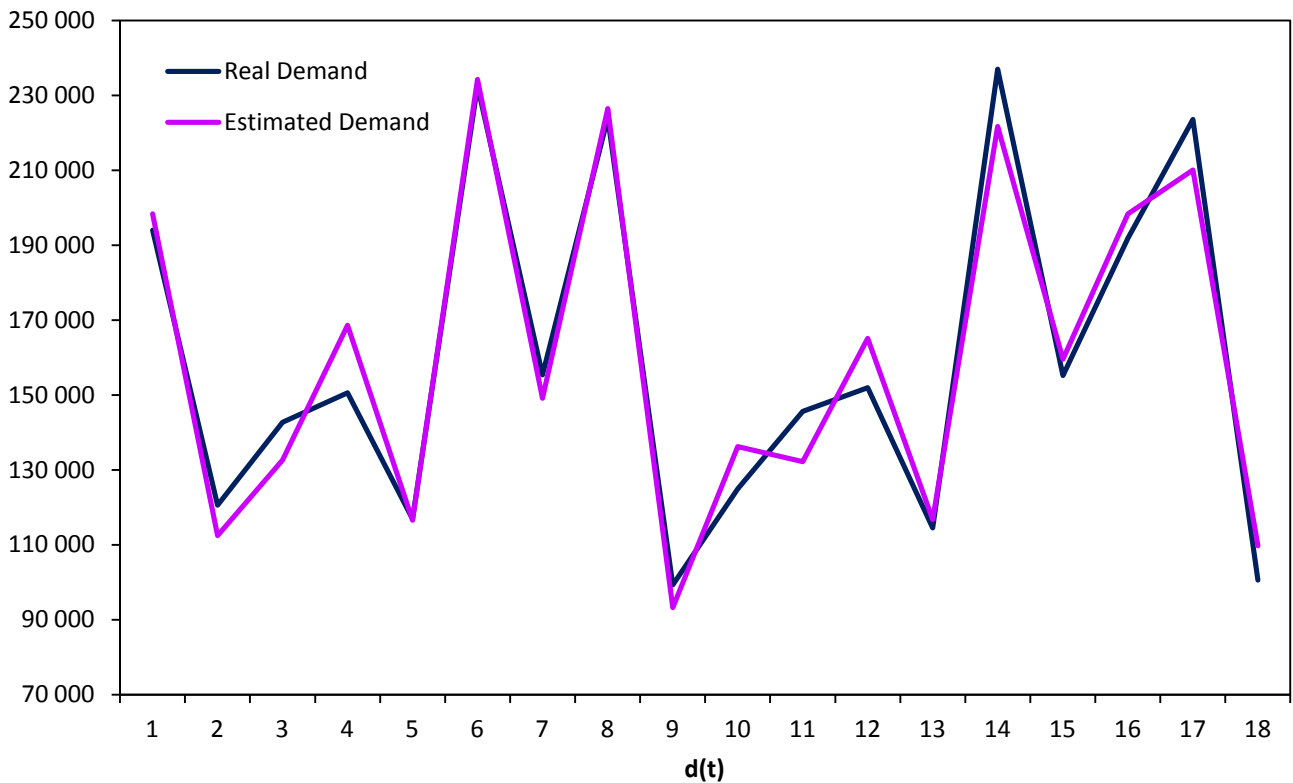


Figure 14: Comparison of the real demand and the demand estimated by the third component linear regression model

7.3.4 Logistic Regression Analysis

Procedure of the logistic regression model estimation is comparable to the one already described in Subchapter 7.2. Also in this case the values taken on by the dependent variable are discretized into five categories, i.e. extremely low, low, medium, high and extremely high.

7.3.4.1. Estimation of Relevant Variables

In order to estimate the logistic regression models of particular components, the relevant variables already determined in Subchapter 7.3.3.1 are exploited. Namely, in the first and the second component model demand is expressed as dependent only on one relevant variable, i.e. revenue. Within the third component, demand is modelled as a function of all the available variables.

7.3.4.2. Estimation of Logistic Regression Models

In Figures 15, 16 and 17 the real and the estimated demand are presented for each of the defined components.

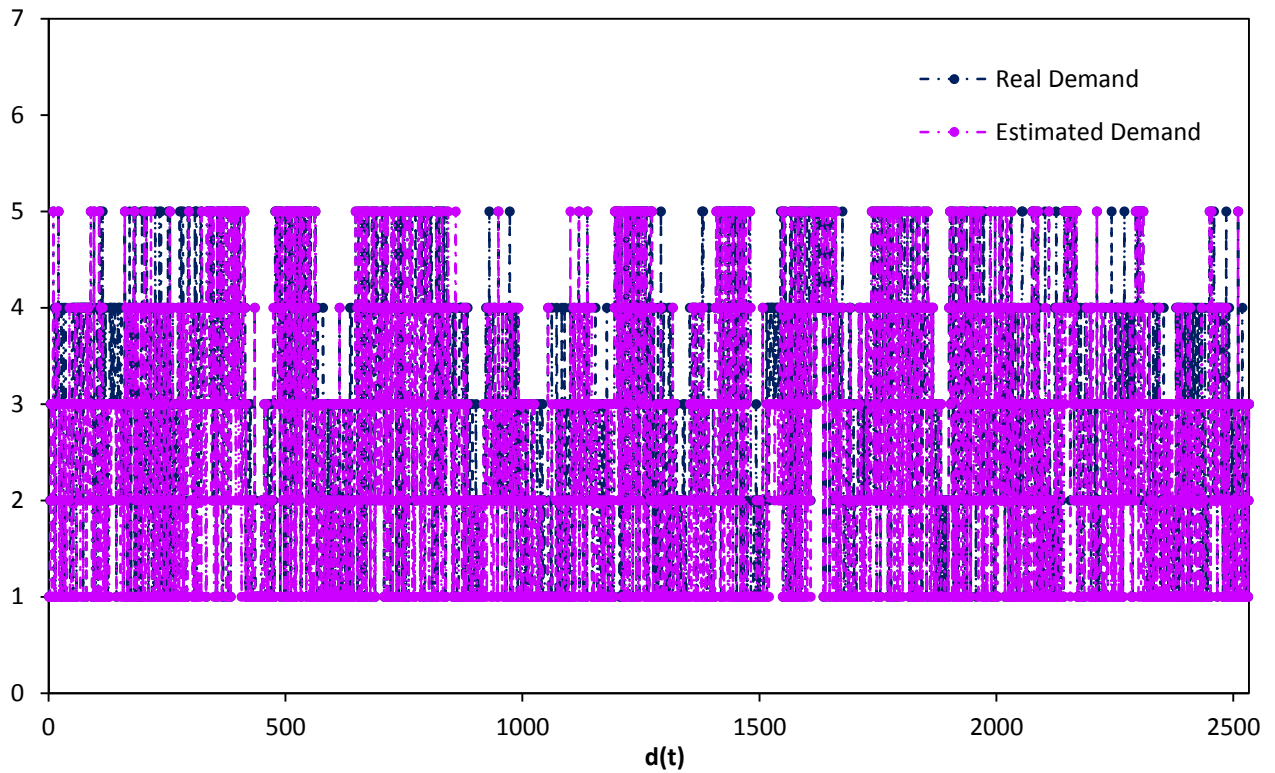


Figure 15: Comparison of the real demand and the demand estimated by the first component logistic regression model

The first component model shows a perfect match of the estimated and the real demand in 1563 from 2533 cases and the second component in 538 from 1524 cases. The third component model consists of extremely high demand value in all cases, thus the match of the predicted and real demand is perfect.

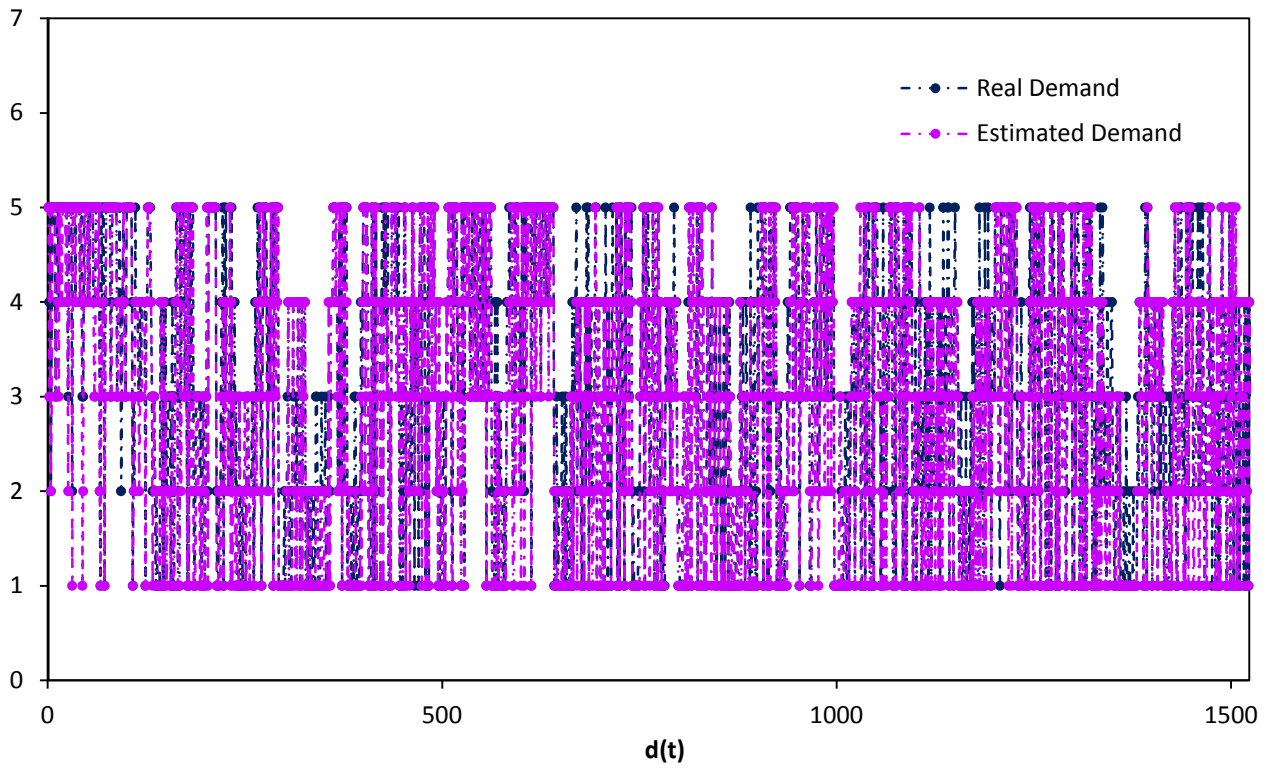


Figure 16: Comparison of the real demand and the demand estimated by the second component logistic regression model

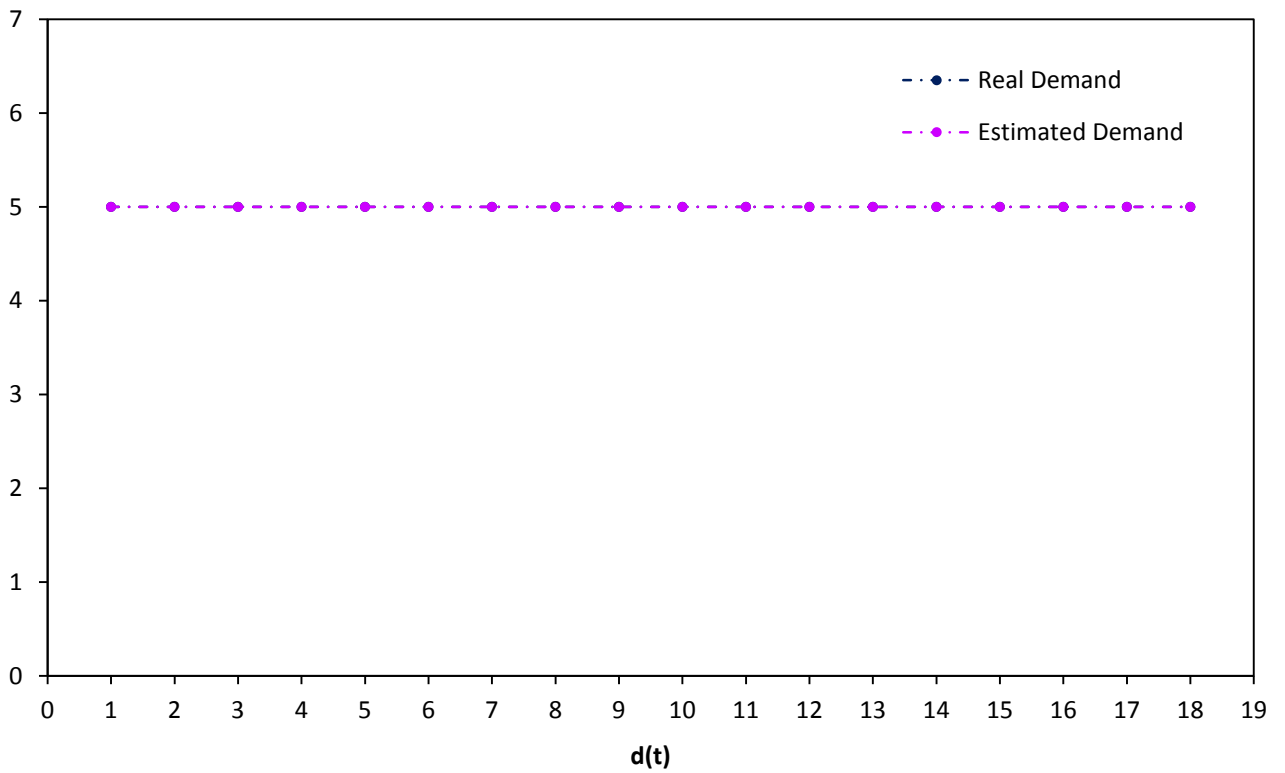


Figure 17: Comparison of the real demand and the demand estimated by the third component logistic regression model

8 DISCUSSION

The results presented in Figure 4 show that at a limited scale, simple linear regression has the potential to satisfactorily estimate air traffic demand. Its limitations were described in detail in the bachelor thesis, as well as in the introduction of this work.

As can be observed from Figures 12, 13 and 14, the estimate of the mixture of linear regression models seems to be very accurate. From the results it was determined that 92 % of estimates obtained from the first component model and 89 % of estimates obtained from the third component model show higher precision than the simple linear regression. However, only 35 % of estimates acquired from the second component model shows the higher precision. Generally speaking, a remarkable 71 % of mixture model estimates is more accurate than those obtained from the simple linear regression.

The extraordinary reliability of the suggested method can also be documented on the sum of squares of prediction error, whose value is relatively small when compared to the values taken on by the modelled variable (see Equation 57, 59 and 61).

Furthermore, as presented in Figure 18, prediction error of the mixture model is one order of magnitude lower than the error of the simple linear regression model.

For the purposes of easier understanding, some concrete samples of the estimated demand were randomly chosen and compared to the real data in Table 7. Even in the cases where the proposed solution does not offer a more accurate result compared to the simple linear regression (see the highlighted rows), the differences between estimates are mostly insignificant.

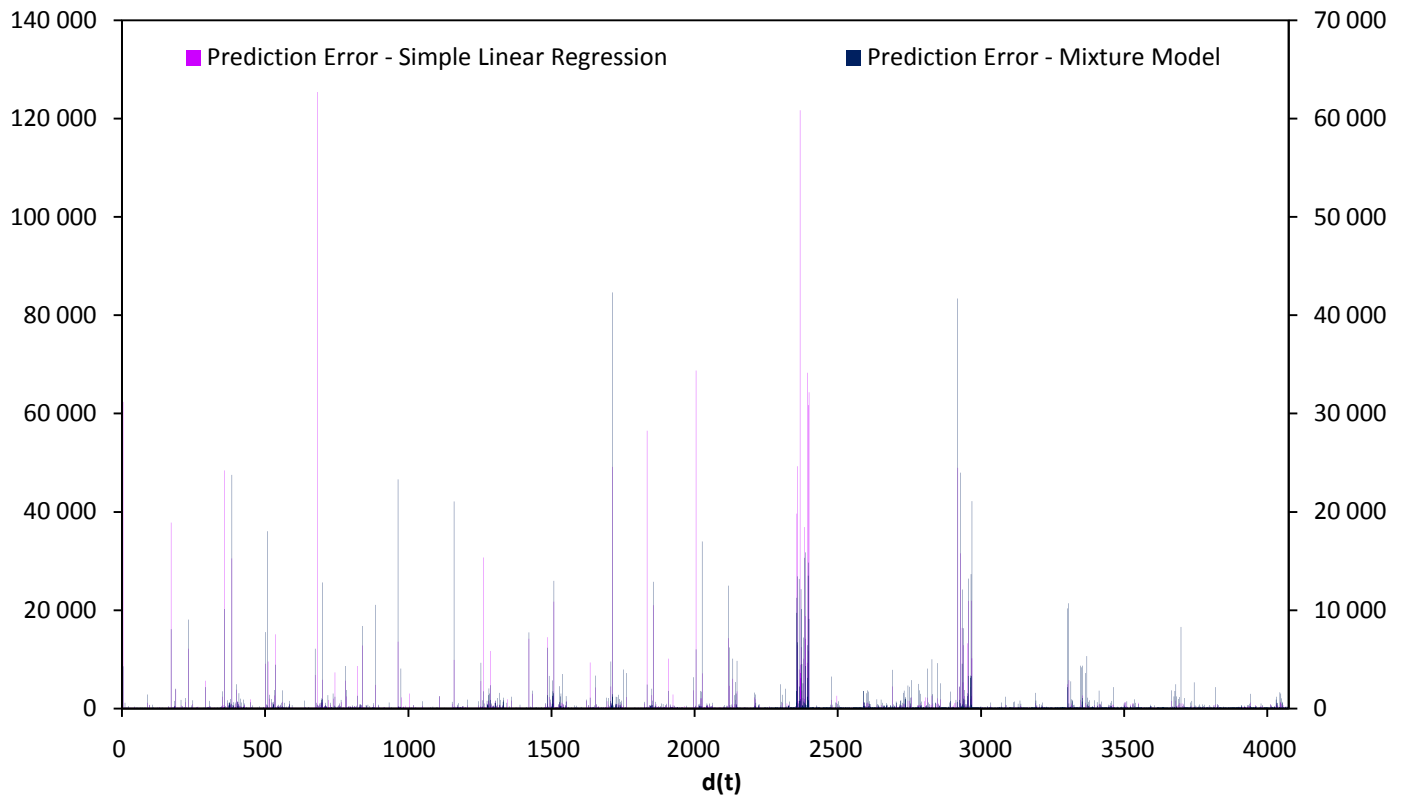


Figure 18: Absolute value of prediction error of the linear regression analyses

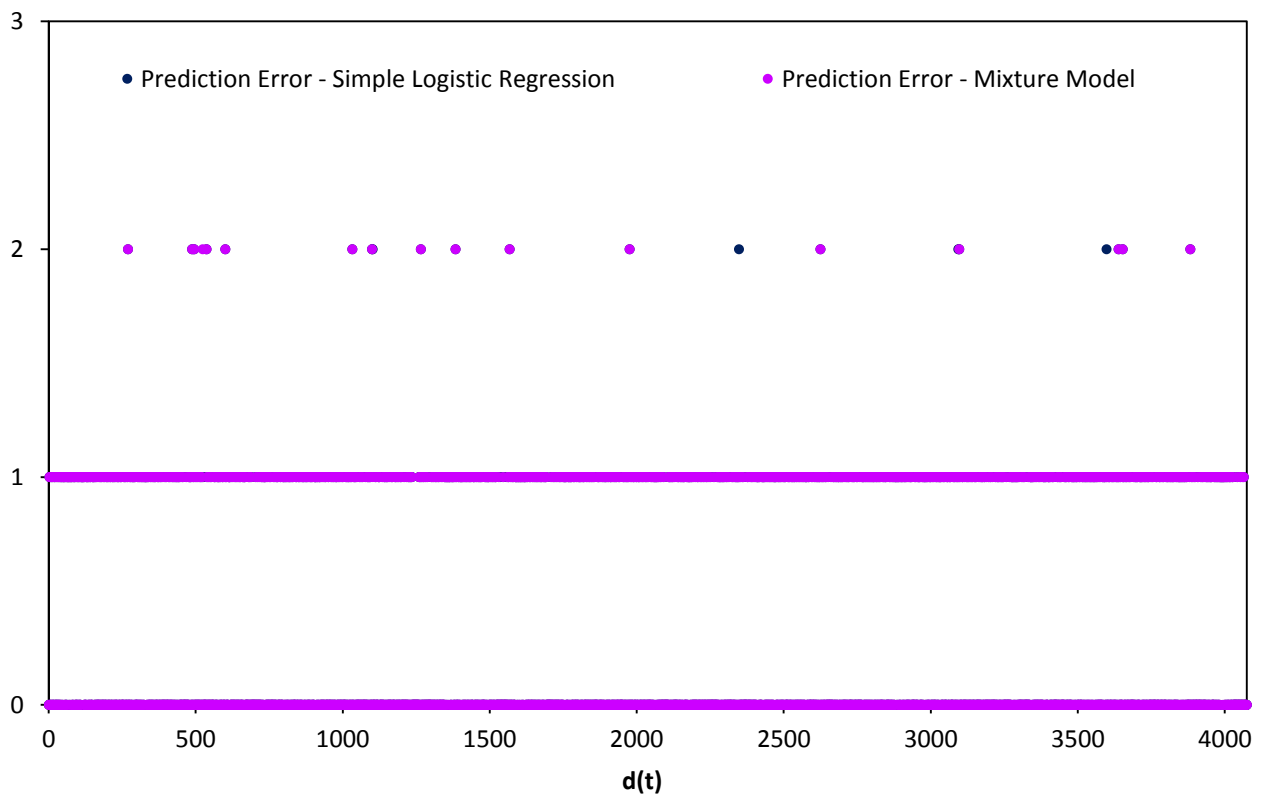


Figure 19: Absolute value of prediction error of the logistic regression analyses

Table 7: Samples of the estimated demand obtained from the simple linear regression as well as the mixture model

Origin	Destination	Explicitly Measured Demand	Simple Linear Regression Estimate	Mixture Model Estimate / Component
Berlin	Moscow	194 039	131 642,1	198 327 3
Prague	Krasnodar	9 873	7 683,15	7 014,33 1
Altenrhein	Rostov	4	99,45	-12,3917 1
Budapest	Barnaul	78	247,46	124,74 1
Duesseldorf	Kazan	3 178	3 015,88	2 689,84 1
Hamburg	Barnaul	853	1 210,25	1 016,82 1
Ostrava	Rostov	13	117,42	4,25 1
Arkhangelsk	Vienna	121	370,34	228,10 2
Samara	Munich	6 846	5 859,90	7 389,33 2
Bremen	Ekaterinburg	83	191,85	73,21 1
Bratislava	Nizhneartovsk	37	134,65	20,22 1
Dresden	Ulan-Ude	13	116,92	3,79 1
Frankfurt	Magas	12	108,09	-4,39 1
Graz	Novyj Urengoj	8	116,08	3,01 1
Hamburg	Omsk	2 522	2 454,76	2 169,93 1
Innsbruck	Barnaul	99	211,59	91,51 1
Klagenfurt	Nizhniy Novgorod	11	114,95	1,97 1
Karlovy Vary	Ekaterinburg	4 632	5 208,37	4 721,31 1
Stuttgart	Orenburg	683	766,18	605,37 1
Moscow	Geneva	114 489	236 192,40	116 719,90 3
Zurich	Moscow	99 289	168 010,4	93 276,88 3
Dresden	Moscow	17 962	12 295,05	15 784,09 2
Brno	Nizhnekamsk	146	210,07	90,10 1
Sochi	Karlovy Vary	13	120,90	7,48 1
Friedrichshafen	Moscow	487	517,79	420,46 2

Logistic regression was assumed to smooth these little nuances through the output discretization and to further improve the reliability of the estimate. Nevertheless, the opposite was verified. Due to the simplification of the task, i.e. discretization of the output, a lot of information is supposed to have been lost, and the estimate becomes even more imprecise. Moreover, the simplification entirely cancelled out the mentioned advantages of the mixture model utilization. The simple logistic regression provided an accurate estimate in 62.5 % of cases, while when the mixture model was utilized, logistic regression provided a precise prediction in 63 % of all cases. In total, the prediction of the simple logistic model was similar to the prediction of the mixture model in 97.5 % of cases. Only 1.5 % of the mixture model estimates were more accurate than the simple logistic model estimates.

Figure 19 shows that in the overwhelming majority of all cases, the prediction error was not higher than one class difference only. Specifically, in case of the simple logistic model as well as the mixture model, 99 % of prediction errors was equal to one class difference only and 1 % corresponded to two class difference.

Some of the concrete samples of the results are presented in Table 8. Only one of these samples shows the mixture model estimate as worse than the simple logistic model estimate (see the highlighted row). All of the other mixture model estimates are the same or better than the simple logistic model estimates.

Table 8: Samples of the estimated demand obtained from the simple logistic regression as well as the mixture model

Origin	Destination	Explicitly Measured Demand	Simple Logistic Regression Estimate	Mixture Model Estimate / Component
Berlin	Moscow	5	5	5 3
Prague	Krasnodar	5	5	5 1
Altenrhein	Rostov	2	1	1 1
Budapest	Barnaul	4	4	4 1
Duesseldorf	Kazan	5	5	5 1
Hamburg	Barnaul	5	5	5 1
Ostrava	Rostov	3	2	2 1
Arkhangelsk	Vienna	4	4	4 2
Samara	Munich	5	5	5 2
Bremen	Ekaterinburg	4	3	3 1
Bratislava	Nizhneartovsk	3	3	3 1
Dresden	Ulan-Ude	3	2	2 1
Frankfurt	Magas	2	2	1 1
Graz	Novyj Urengoj	2	2	2 1
Hamburg	Omsk	5	5	5 1
Innsbruck	Barnaul	4	3	3 1
Klagenfurt	Nizhniy Novgorod	2	2	2 1
Karlovy Vary	Ekaterinburg	5	5	5 1
Stuttgart	Orenburg	5	4	5 1
Moscow	Geneva	5	5	5 3
Zurich	Moscow	5	5	5 3
Dresden	Moscow	5	5	5 2
Brno	Nizhnekamsk	4	3	3 1
Sochi	Karlovy Vary	3	2	2 1
Friedrichshafen	Moscow	5	4	4 2

8.1 Model Criticism

Generally applicable doubts relate to the issue of the appropriate model formulation, the choice of variables, and furthermore the initial parameters setting.

8.1.1 Model Formulation

The weaknesses of the simple linear regression model, such as the misleading presumption of system linearity and stationarity, have already been comprehensively discussed in the bachelor thesis as well as in the introduction of this work. In the mixture of the linear regression models, the system idealization is still present. However, as was verified, the data classification enabled to discover deeper system structures and prominently improve the estimate reliability.

The output discretization, which is essential for the logistic regression analysis, causes a significant information loss. As a result, the precision of the estimate decreases and the advantages, which mixture model provides in the case of continuous model, are significantly suppressed.

The criticism of the mixture model lies mostly in the approximation $w_{c,t} \rightarrow \delta(c, c_t)$. The active component c_t at the time instant t is considered as known and thus pdf of the pointer model is approximated by the Kronecker function. However, practical experience shows that the simplification does not have a significant impact on the reliability of the results.

8.1.2 Choice of Variables

Another important factor in deciding the model's reliability is the appropriate choice of independent variables. For the purposes of this work, the initial set of variables was selected according to the various types of methodologies as well as expert knowledge; see Chapter 3. Variables relevant to the particular model were further selected from this set by performing the regression analysis on the standardized data.

Even though the selection of variables was undertaken with respect to all the significant aspects referred in the literature, concerns regarding this issue can never be neglected, because of the complexity of the task. Variables, which were not included into the modelling process due to the databases insufficiencies, or due to the specific focus of the task, are the secondary variables such as percentage of degree holders and employment composition

structure. Thus, for the purposes of further studies an examination of these secondary variables in the context of air traffic demand modelling is strongly suggested.

8.1.3 Initial Parameters Setting

Setting of the initial parameters has a crucial impact on the overall process of solution derivation. For the purposes of this work, the parameters were determined according to the expert knowledge. The noise covariance was defined as bigger than the number of data samples, and thus the size of components was assured to remain unchanged. In this way the components were only allowed to move in the multidimensional space. The setting is thought to provide an objective result, independent on the sequence of data samples. Other parameters were determined according to the expert knowledge.

Even though the expert approach provided a sufficient level of the mixture estimate, a missing analysis of the initial parameters setting is still perceived as a weak point of the thesis. For the purposes of further studies, closer examination in this field is strongly encouraged.

8.2 Data Criticism

Another aspect essentially influencing the reliability of modelling results is the accuracy of input data. In the upcoming subchapters the problems regarding the origin and destination data are discussed, along with socio-economic data insufficiencies.

8.2.1 Origin and Destination Data

Because of the deficient interconnections between airline reservation systems, the absolute completeness of origin and destination data cannot be guaranteed. According to unofficial sources, the error of origin-destination data can reach up to 30 %.

8.2.2 Socio-economic Data

The socio-economic data were drawn from two electronic databases, the Eurostat database and the Russian Federal State Statistics Service. It is important to emphasize that the reliability and range of the socio-economic data were significantly limited by the structure of the databases used.

Unfortunately, it was not possible to ensure the time consistency of the input data; data from the years 2011 to 2014 were collected. Taking into account the data standardization as well as the presumption of the linear development of the indicators in time, the mentioned insufficiency can be neglected.

The databases also do not provide complex information about cities, thus only indicators capturing the situation on the regional level were collected. As was presented in Subchapter 7.3, the mixture model has discovered deeper system structures based on this information. That is why it is believed that regional indicators are a sufficient alternative for the purposes of air traffic demand estimation.

Another inconvenience related to the use of information from several databases is that information is collected and processed according to different methodologies. Therefore, the comparability of such information can be strongly limited.

The last but not least important critique is the absence of a great deal of information in the electronic databases, which negatively influenced the coherency and range of the input data.

Despite all the insufficiencies stated, it is reasonable to say that on the overall level the collected data provided a satisfactory framework for the proceeded calculations.

8.3 Comparison with ICAO Methodology

Even though mixture models are used in a variety of applications, their exploitation for the purposes of air traffic demand estimation is only very marginally mentioned in scientific articles or in International Civil Aviation methodology. Most air traffic forecasting methods utilize only one equation model.

According to ICAO methodology, for most long-term air traffic demand forecasts the one equation model is utilized in the following form

$$\log(y) = a \cdot \log(GDP) + b \cdot \log(yield) + \log(c) \quad (62)$$

As can be observed, the model considers the log linear dependency of the variables involved instead of the purely linear one.

The results presented in Subchapter 7.3.2 emphasize that while considering only GDP and yield as relevant variables, the system shows a higher level of coherency. However, as was verified, the coherency can be easily disrupted by the use of a mixture model with an appropriate initial parameters setting. Therefore, the one equation approach is considered questionable and it is believed that the mixture model exploitation can mean a significant improvement even in the case of the modelling technique presented by ICAO.

The process of relevant variables estimation revealed the importance of variables other than those given by the ICAO model. Thus, for the purposes of further studies it is suggested the combined utilization of a mixture model and the log linear regression analysis be examined; furthermore a review of the composition of variables involved is also strongly recommended.

9 CONCLUSION

The computations executed in the thesis verified that the use of the mixture model offers the possibility to significantly improve the reliability of air traffic demand estimation.

The system was proved to exist in several behaviour modes, which is the main precondition for successful implementation of the suggested method. As presented in Subchapter 7.3.2, while utilizing all the available variables, the mixture model provides a sufficient level of data classification. In the cases where variables are selected according to the ICAO methodology as well as according to the economic theory, the results show a higher level of system coherency. However, in both cases the coherency can be easily disrupted by an appropriate initial parameters setting.

The utilization of mixture of linear regression models proved to be especially beneficial. In the elaborated task it provided a significant improvement of estimates in 71 % of all cases when compared to the simple linear regression model. Thus, this method credibly verified its qualities in the area of air traffic demand modelling.

The logistic regression model was assumed to provide even more accurate estimates due to the output discretization. However, the discretization caused a significant loss of information and limited the previously observed advantage of mixture model implementation. Nevertheless, due to the relatively small prediction error observed, for special types of applications, where the discretized output is considered sufficient, this approach can also be recommended.

For the purposes of further studies, it is suggested the utilization of a mixture model and the log linear regression analysis be merged. The solution is believed to combine the advantages of the system decomposition presented by the mixture model as well as the exceptional properties of the log linear regression model, which is abundantly applied by the International Civil Aviation Organization.

10 RESOURCES

- [1] Air Transport Action Group, "The economic and social benefits of air transport," *International Civil Aviation Organization*, 2005.
Available: http://www.icao.int/Meetings/wrdss2011/Documents/JointWorkshop2005/ATAG_SocialBenefitsAirTransport.pdf.
- [2] "Convention on International Civil Aviation," *International Civil Aviation Organization*, 1944.
Available: http://www.icao.int/publications/Documents/7300_orig.pdf.
- [3] I. R. Barnes, "The Economic Role of Air Transportation," *Law and Contemporary Problems*, 1946.
Available: <http://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=2257&context=lcp>.
- [4] G. Burghouwt, „Airline Network Development in Europe and its Implications for Airport Planning,“ Ashgate Publishing Group, 2008.
Available: <http://80.site.ebrary.com/dialog/cvut.cz/lib/cvut/detail.action?docID=10209165&p00=airline+network+development+europe+implications+airport+planning>.
- [5] International Civil Aviation Organization, "Manual for Air Traffic Forecasting," 2006.
Available: http://www.icao.int/MID/Documents/2014/Aviation%20Data%20Analyse%20Seminar/8991_Forecasting_en.pdf.
- [6] "Where we fly," *Czech Airlines*, 2015. Available: <http://www.csa.cz/en/portal/info-and-services/travel-information/flight-map.htm>.
- [7] J. Sůra, „České aerolinie se dál zmenšují. Loni jim ubyly pětina cestujících,“ *iDnes*, 2015. Available: http://ekonomika.idnes.cz/pocty-cestujicich-csa-2014-klesaly-dom-eko-doprava.aspx?c=A150217_111835_eko-doprava_suj.
- [8] Český Aeroholding, „Konsolidovaná výroční zpráva představenstva o podnikatelské činnosti společnosti a o stavu jejího majetku za rok 2013,“ 2014.
Available: http://www.cah.cz/cs/o-nas/vyrocnizprava/Contents/0/cah_consol_annual_report_2013.pdf.
- [9] Czech Airlines, „České aerolinie si připomínají 55. výročí zahájení provozu na mezinárodní letiště Moska-Šeremetěvo,“ 2015.
Available: http://www.csa.cz/cs/portal/quicklinks/news/news_tz/news_tz_data/tz_20082015.htm.

- [10] „Aeroflot chce opustit alianci Skyteam, odchod by mohl poškodit i ČSA,“ *E15*, 2013. Available: <http://zpravy.e15.cz/byznys/doprava-a-logistika/aeroflot-chce-opustit-alianci-skyteam-odchod-by-mohl-poskodit-i-csa-1002196>.
- [11] S. Srinidhi, “Development of an Airline Traffic Forecasting Model on International Sectors,” *IEEE Xplore*, 2009.
Available: <http://80.ieeexplore.ieee.org/dialog.cvut.cz/stamp/stamp.jsp?tp=&arnumber=5234138>.
- [12] I. Nagy, S. Evgenia a M. Kárný, „Bayesian Estimation of Mixtures with Dynamic Transitions and Known Component Parameters,“ *ÚTIA*, 2011.
Available: <http://www.utia.cas.cz/files/soutez11/kyber/nagy.pdf>.
- [13] K. Minsoo, “Statistical Classification,” Pomona College, 2010.
Available: <http://pages.pomona.edu/~jsh04747/Student%20Theses/MinsooKim10.pdf>.
- [14] M. Rychlý, „Klasifikace a predikce,“ Vysoké učení technické v Brně.
Available: <http://www.fit.vutbr.cz/~rychly/public/docs/classification-and-prediction/classification-and-prediction.pdf>.
- [15] I. Nagy, „Stochastické systémy,“ 2013.
Available: <http://staff.utia.cas.cz/suzdaleva/pdfka/StSysTexty.pdf>.
- [16] T. M. Mitchell, “Machine Learning,” Carnegie Mellon University, 2015.
Available: <https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf>.
- [17] I. Nagy, „Pokročilé statistické metody a jejich aplikace,“ České vysoké učení technické v Praze, Fakulta dopravní, 2015.
Available: <http://nagy.rudolfpohl.cz/Doktorandi/LecturesPhD.pdf>.

11 LIST OF FIGURES

- Figure 1 Connections from the Czech Republic to Russia provided by Czech Airlines nowadays
- Figure 2 Graphical presentation of the *logit* function: $z = \ln p/(1-p)$
- Figure 3 Graphical presentation of the inverse *logit* function: $p = \exp(z)/(1+\exp(z))$
- Figure 4 Comparison of the real demand and the demand estimated by the simple linear regression
- Figure 5 Comparison of the real demand and the demand estimated by the simple logistic regression
- Figure 6 Development of the components' parameters during the statistics recursion in the first computation
- Figure 7 Development of the components' parameters during the statistics recursion in the second computation
- Figure 8 Development of the components' parameters during the statistics recursion in the third computation
- Figure 9 Development of the first component parameters during the statistics recursion of the first scenario
- Figure 10 Development of the second component parameters during the statistics recursion of the first scenario
- Figure 11 Development of the third component parameters during the statistics recursion of the first scenario
- Figure 12 Comparison of the real demand and the demand estimated by the first component linear regression model
- Figure 13 Comparison of the real demand and the demand estimated by the second component linear regression model
- Figure 14 Comparison of the real demand and the demand estimated by the third component linear regression model
- Figure 15 Comparison of the real demand and the demand estimated by the first component logistic regression model
- Figure 16 Comparison of the real demand and the demand estimated by the second component logistic regression model

- Figure 17 Comparison of the real demand and the demand estimated by the third component logistic regression model
- Figure 18 Absolute value of prediction error of the linear regression analyses
- Figure 19 Absolute value of prediction error of the logistic regression analyses

12 LIST OF TABLES

Table 1	Regression coefficients obtained from the linear regression performed on the standardized data
Table 2	Estimation of pointer parameters in three different calculations
Table 3	Data classification obtained from the utilization of mixture model
Table 4	Regression coefficients obtained from the standardized data of the first component
Table 5	Regression coefficients obtained from the standardized data of the second component
Table 6	Regression coefficients obtained from the standardized data of the third component
Table 7	Samples of the estimated demand obtained by the simple linear regression as well as the mixture model
Table 8	Samples of the estimated demand obtained by the simple logistic regression as well as the mixture model