Gravity model for air passenger demand estimation: the addition of big data

By

Wesley Kevin Boelrijk

TU Delft

DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF TECHNOLOGY, POLICY AND MANAGEMENT

MSC ENGINEERING AND POLICY ANALYSIS

MASTER THESIS

Gravity model for air passenger demand estimation: the addition of big data

TU Delft supervisors
Dr. J.A. ANNEMA
Dr.ir. B. ENSERINK

Author
W.K. BOELRIJK
4514505
WESLEYBOELRIJK@GMAIL.COM

Special supervisor (VU Amsterdam)
Dr. F. BLASQUES ALBERGARIA AMARAL

KLM supervisors
B. TESSELAAR
A.J. BEEKS

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on August 23rd of 2019.

August 9, 2019

ABSTRACT

Over the last decades, flying has become increasingly accessible and the global aviation industry is growing rapidly. This causes airports and airspaces to reach their capacity limits. Reliable estimation and forecasting of air passenger demand is very important for airlines in order to allocate resources effectively in a constrained and competitive aviation environment. A model is developed to estimate air passenger demand suitable for routes where currently no direct service exists in order to aid decision-making on new destinations. For this model the traditional gravity model is extended with big data from worldwide flight search engines (meta search data). The model is applied to a case study with real world data from KLM Royal Dutch Airlines. We find that the gravity model enriched with meta search data is able to accurately predict demand and rank destinations that are currently not connected to KLM's network by a direct flight. The data-driven approach allows us to give recommendations about promising destinations that could be added to an airline's network. Ultimately, a more efficient network as a result of advanced demand estimation is expected to lead to higher load factors and lower emissions per passenger.

Keywords: Air Passenger Demand Estimation, Aviation, Networks, Gravity Model, Ordinary Least Squares, Big Data

JEL Classifications: C01, C13, C21, C52, C80, C82

PREFACE

Dear reader,

This master thesis was the final challenge for me in order for obtain the Master of Science degree in Engineering & Policy Analysis from Delft University of Technology. During my exchange semester abroad, as part of the master, I got inspired to pursue a second master degree. After obtaining all credits for the courses, thus excluding the master thesis, I started with a master in Econometrics at the VU Amsterdam. After obtaining all course credits in the second master as well, a double master thesis assignment awaited. Both master theses are being performed sequentially during a extensive internship at KLM Royal Dutch Airlines. Interestingly, this first thesis originally started out to be delivered at the VU Amsterdam, but after a change of plans the work was converted to the TU Delft master thesis.

The idea for the thesis topic was born during a brainstorm session at KLM when they asked the question: "*What is the next best destination we could fly to?*" Since I worked with internal company data that includes sensitive competitive information, specific KLM results on recommended destinations are classified as confidential and are therefore excluded from the published version. The prepared data set and developed code are available upon request for the graduation committee.

As I took a special journey throughout the master and especially during the master thesis, quite a few people got involved in the process. First of all I would like to thank my TU Delft supervisor Dr. Jan Anne Annema and chair of the graduation commitee Dr.ir. Bert Enserink for their feedback and flexible attitude in the process. Secondly I would like to thank my special supervisor from the VU Amsterdam Dr. Francisco Blasques for his enthusiasm, feedback and continued support even after the work converted to become the TU Delft thesis. Furthermore I would like to thank my KLM supervisors Bjørn Tesselaar and Anne Jan Beeks for their guidance and support. Besides that I would like to thank my friend and classmate Charlotte de Bruijn for helping me out with structuring numerous of my ideas and proofreading the entire thesis. Last but not least I would like to thank anyone whom I worked with during my internship at KLM, in special the departments Operations Research and Network Planning.

I hope you enjoy reading my master thesis.

Wesley Kevin Boelrijk
Amstelveen, August 2019

EXECUTIVE SUMMARY

Over the last few decades, flying has become increasingly accessible and the overall aviation industry is growing rapidly. While this benefits the worldwide and local economies, employment and accessibility, it also has its negative impact with emissions and noise pollution. Due to a recent period of strong growth and increasing public resistance against flying, airports start to reach their capacity limits. This implies that airlines need to be more selective in the flights they perform as they are constrained in their growth.

Reliable estimation and forecasting of air passenger demand is very important for airlines to be profitable in a constrained and competitive aviation environment. There is an issues with the current demand estimation method based on historic booking data. The problem is that is that the true demand is unobserved as one is only able to collect historic booking data on routes where a direct service exists or where transfers are registered. Secondly, the source for historic booking data is deteriorating in quality as the booking systems become more and more decentralized over the last years. Therefore, the following research question is presented:

How can air passenger demand be estimated accurately, suitable for city-pairs where currently no direct air service exists in order to assist the flight network decision-making?

In order to answer this question a literature study is performed addressing the current standards of air passenger demand estimation. The gravity model is the most suitable method to estimate demand for destinations where currently no direct service exists (Grosche et al., 2007). The gravity model explains flows of people or cargo between two locations based on geo-economic attraction variables and the distance between these locations. However, this traditional method has shortcomings and its applicability to large-scale networks is questioned (Verleger Jr et al., 1972). In order to improve on these limitations a new source of information is added to the traditional gravity model in the form of big data.

Subsequently an empirical case study is performed on the network of KLM Royal Dutch Airlines, the airline that operates at its capacity constrained hub Amsterdam Airport Schiphol. Through KLM we have access to 15 billion records of searches for flight on flight search engines from the years 2017 and 2018. The so-called meta search data contains information regarding the true origin and destination of a potential passenger, information that is normally unobserved. The meta search enriched gravity model is applied to real world data from KLM in order to measure the benefit of the added big data. Nine different model specifications are made in order to test which combination of variables in the model have the most predictive power for demand estimation.

We find that the most extensive models that contain the traditional gravity variables, the newly added meta search data and numerous control variables performs the best at with R^2 of 0.92 indicating quite a good model fit. The added predictive power of the meta search data seems to outperform the traditional gravity variables completely. The model is validated using k-fold cross validation and a routine to predict back existing KLM destinations. The best models predict 70% of the current KLM destination with a rank in the top 10 based on total demand. Furthermore, the data-driven approach allows us to give recommendations about promising destinations that could be added to an KLM's network based on total predicted demand and under-predicted destinations compared to historic booking data. KLM's five latest additions to the flight network are all ranked within the top four, which seems to indicate that this model has a lot of real world value.

An academic conclusion drawn is that meta search data seems to be highly valuable for air passenger demand estimation. The results give the impression that adding the meta search data improves the estimation of true air passenger demand and that it partly overcomes the measurement error that is present in the historic booking data. In the gravity model enriched with meta search data, the traditional gravity variables are not significant anymore as they are outperformed by the meta search data. This might indicate that the traditional gravity model is not as relevant today anymore as when it was created.

The gravity model enriched with meta search data seems to shows real world value as it accurately ranks the most recent KLM destinations highly. The model contributes to the network decision-making of the airline by providing a prediction for the unobserved true demand. Therefore it might uncover destinations where previously no attention or interest had been based on the deteriorating historic booking data. Despite the promising results, the meta search data as well as the model has some limitations which one should be aware of. It is the price-sensitive leisure passengers that typically compares flights the most and thus logs the most searches. Therefore the model is limited the type of destinations it can recommend. Destinations consisting mainly of business traffic, which are the most profitable for an airline, are likely to be overlooked using this model. Nonetheless, a well developed and more efficient KLM network helps Amsterdam Airport Schiphol to offer better connectivity within its capacity constrains. A more efficient network in terms of resource allocation might lead to higher load factors resulting in lower emissions per passenger. During this research the case study is applied to the KLM network on the constrained Amsterdam Airport Schiphol. However, the developed method can be generalized to other airlines and airports as well.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AF	IATA designator for Air France
AF-KLM	Air France - KLM
AMS	Amsterdam
BA	IATA designator for British Airways
CSV	Comma-Seperated Values
DBD	Days Before Departure
DL	IATA designator for Delta Air Lines
JSON	JavaScript Object Notation
GDP	Gross Domestic Product
GDS	Global Distribution Service
HV	IATA designator for Transavia
KL	IATA designator for KLM
KLM	KLM Royal Dutch Airlines (Dutch: Koninklijke Luchtvaart Maatschappij)
LCC	Low Cost Carrier
LH	IATA designator for Lufthansa
IATA	International Air Transport Association
MS	Meta Search
MIDT	Marketing Information Data Tape
O&D	Origin and Destination

CONTENTS

1 INTRODUCTION

This master thesis research is performed during an internship at Air France - KLM Royal Dutch Airlines (AF-KLM). Data from KLM TripPlanner and Network tools of the Operations Research department are used from the period January 2017 to December 2018.

Reliable estimation and forecasting of air passenger demand is very important for airlines to be profitable and competitive (Grosche et al., 2007). A successful airline is characterized by optimally dividing its assets to fly to the right destinations. This is particularly important since airlines have large upfront investments in their fleet and need to make early choices about the aircraft types they purchase. Secondly, airlines must commit themselves one year in advance to the destinations and frequencies of their flights, regardless of the actual demand for those destinations at the time of the flights. These investments and commitments are based on what the airline expects from the market. Thus, in order to make a profit in the competitive airline industry it is important to have a sound market estimation, since margins are particularly low.[1]

Before demand can be estimated, it is crucial to note that *true demand* is (partially) unobserved. For example, when a passenger flies from Amsterdam to New York, it is assumed that the true origin was Amsterdam and the true destination is New York. However, this passenger could have a connecting flight from Vienna to Amsterdam and later a connecting flight from New York to Washington DC. This means that the true origin-destination (O&D) combination is Vienna - Washington DC. If the passenger flies all three connecting flights of the route with the same airline or partner airline, the true O&D is known (at that specific airline). However, if the passenger books separate flights, three different passenger flows are registered.

Traditionally, bookings were done by travel agencies connected to the largest Global Distribution Systems (GDSs). These systems contained all information regarding the true O&Ds of all passengers. However in the last ten to fifteen years it has become increasingly normal for customers to book flights directly at the airline's own booking channels, without the use of a travel agency. This causes that the information for airlines, which they use to estimate demand on, to deteriorate. It might be the case that new methods or data has to be used to still estimate demand accurately. This poses the following research question:

How can air passenger demand be estimated accurately, suitable for city-pairs where currently no direct air service exists in order to assist the flight network decision-making?

[1] IATA publication: https://www.iata.org/pressroom/pr/Pages/2018-06-04-01.aspx

Academic knowledge gap
The literature on air passenger demand estimation and forecasting started in the 1950's and is still a topic that has a considerable amount of attention from researchers today (Wang and Song, 2010). A very well known and popular model in the air travel demand estimation literature is the gravity model. Grosche et al. (2007) state that "*Since no single technique guarantees accuracy, airlines in fact compare forecasts from several different models. Within this set of forecasting methods, the most widely used is the gravity model.*" The gravity model explains passenger flows between two entities (airports, cities, countries etc.) and is able to estimate demand for city-pairs where no direct air service exists currently. Grosche et al. (2007) argue that the gravity model is especially appropriate, because it is able to estimate demand for destinations that are not currently in the network and it has minimal dependence on airline related data.

However, the gravity model is quite a traditional method. The main issue is that the model has been proven to work well on a small scale, but on a larger scale (multiple markets) the performance of the gravity model is questionable (Verleger Jr et al., 1972; Grosche et al., 2007; Kopsch, 2012; Hazledine, 2017). Since the gravity model is the best way to estimate demand for destinations that are not currently in the network we are still eager to use it. We suggest that big data would be a good addition to the gravity model. It can arguably solve the problem of the increasingly missing information and the multiple market issue because this information is implicitly incorporated in the data. Therefore we propose to enrich the traditional gravity model with new data sources from the 21^{st} century called meta search. In this way connectivity between city pairs can be captured by other influences that were not available before. It will also show if the gravity model is still useful to estimate demand or if we need completely new methods. We propose to enrich the gravity model by adding meta search data to proxy air passenger demand.

Practical relevance and multi-actor perspective
Over the last few decades, flying has become increasingly accessible and the overall aviation industry is growing rapidly.[2] While this benefits the worldwide and local economies, employment and accessibility, it also has its downsides. Many airports to have reached their capacity limits. A constrained airport has a reducing competitive position that threatens her airlines and the slows down the development of international connectivity. National economies are often dependent on travellers and thus being connected and accessible is important. Many airports deal with capacity constraints in the form of physical space, but other limitations such as emission and noise pollution exists as well. We observe an increasing public resistance on flying. Especially concerns regarding noise pollution and CO_2 emissions are raised by residents of the area. While some parties benefit from a growing avi-

[2]https://www.iata.org/publications/store/Pages/20-year-passenger-forecast.aspx

ation industry, others are harmed by it.

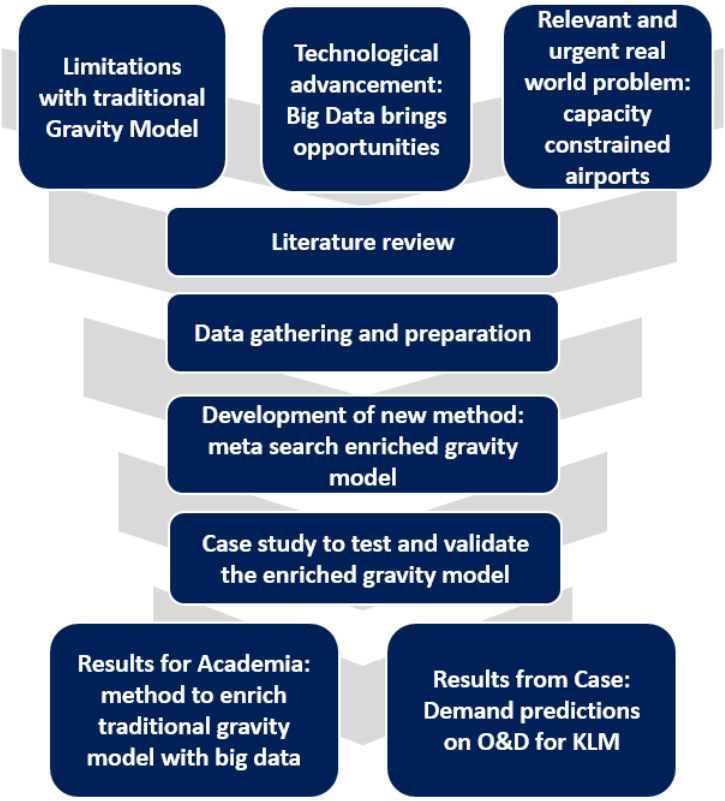
This makes the problem we are addressing is a typical multi-actor problem where the perspectives and interests of the involved parties are unaligned. A couple of actors are explained here. First we have the airports who are interested in expanding, because their livelihood depends on it. The aviation landscape is very competitive and not being able to grow can quickly become problematic. The airlines using the airport are companies with a profit motive and try to capture as many of the available demand as possible. Passengers using the airlines and airports are generally also in favor of a growing aviation industry as this increases accessibility and could reduce prices. Then there are local residents who are often not happy about the airport expanding, because it causes nuisance in the form of emissions and noise pollution. There is a large public debate going on worldwide about the effects of aviation on the environment. Especially extra emission and noise pollution of additional air traffic. The last key actor are the governments that underline the importance of the aviation sector for reasons such as that their economic infrastructure depends on tourism or is an important business location, while at the same time they try to reduce CO_2 emissions as part of their long term objectives to battle climate change.

These contradictory interests make future growth of constrained airports highly uncertain. Airlines and airports have to make decisions on investments in fleet and infrastructure under this large uncertainty. This research aims to develop a model to help the aviation industry make better demand estimations to optimize allocation of current resources such as the fleet and airports flight movements.

Research approach
Several steps will be taken in order to find the answer to the main research question. These steps are represented schematically in flow chart in Figure 1. The motivation for this research is three-fold. The first two motives for the research are academic in nature and are described in the knowledge gap above. The traditional gravity model was developed decades ago and copes with certain limitations. We have a lot of extra data now which can prove to be very helpful. One could imagine that big data as a product of human behaviour captures underlying patterns that traditional approaches such as a gravity model could not capture. Also it contains more information about true demand.

Besides an academic research gap the idea for the research originates from a relevant and urgent real world problem. The aviation industry has been growing for decades and many airports are coping with space limitations. Airports are getting congested, by either capacity constraints, emission legislation or noise pollution. A concrete case is Amsterdam Airport Schiphol. This is the third research motive.

Figure 1: Flow chart on the research approach



To answer our research question we start with a literature review about the current standards for air demand estimation. The gravity model is reviewed and the idea to improve on the traditional gravity model by adding meta search data is born. Of course, adding meta search also comes with some difficulties, which will be discussed.

In order to make any statements on the usefulness, applicability and performance of the new method, a case study will be done. The method will be tested on a case study for KLM. KLM is the largest airline at Schiphol responsible for the majority of the network connectivity. The gravity model will be applied to the flight network of KLM. Due to the choice for a case study we are able to assess the practical feasibility and usefulness of our new method/model and it allows us to use KLM's data. Several model specifications can be compared to find which adjustments and additions to the traditional model yield the best improvement. The new model will be validated according to real world data.

The results from the case study will have real world implications and academic implication. On one side, for academia we will draw conclusions about the 'new' method, which will show if the gravity model is still relevant. On the other hand, we will have predictions for KLM to assist in the network decision-making process.

2 THEORETICAL BACKGROUND

In this section we will examine various ways for air passenger demand estimation from current literature, make the choice of which model to use and argue why in Section 2.1. We will then present the chosen model and explain it in detail in Section 2.2. Then we critically discuss the pros and cons of using the gravity model in Section 2.3. Finally, Section 2.4 deals with literature about adding big data into our model.

2.1 Air passenger demand estimation

This subsection provides an overview on the methods for air passenger demand estimation described in the literature. Each method has its advantages and disadvantages that determine the applicability for the research: determining air passenger demand for O&Ds where currently no air service exists. Subsequently the most suitable method is selected.

Air passenger demand estimation review
The literature on air passenger demand estimation and forecasting started in the 1950's and is still a topic that has a considerable amount of attention from researchers today (Wang and Song, 2010). Verleger Jr et al. (1972) review a series of air transportation demand models and classify three classes of air passenger demand models: 1) Aggregate travel models, 2) City pair or point-to-point models and 3) Gravity and cross-sectional models.

Verleger Jr et al. (1972) state that aggregate models appear most frequently and that the model usually measures air traffic in revenue passenger kilometres (as a homogeneous commodity). In the model, a measure for aggregated demand is explained by variables such as price, national incomes and a measure for alternative methods of travel. Verleger Jr et al. (1972) explain that the models have been successful for accounting purposes and revenue forecasting under the influence of changing prices and income of the consumers. However, our research goal is to estimate demand for destinations that have currently no air service in place, which implies that no data on revenue and pricing is available for new O&Ds. This makes the class of aggregate models not applicable for our purpose.

The second method described by Verleger Jr et al. (1972) is the city pair or point-to-point method for air transport demand. This class of models creates separate models for each O&D in the data set, such that the changes in demand over time can be analyzed. In this model the most important explanatory variables of demand are income and price. The main benefit of this technique is that an accurate model can be made per city pair to forecast demand into the future.

This method is often used to measure and scientifically explain price and income elasticities (Verleger Jr et al., 1972). The prerequisite of this model is that you need data per city pair over various points in time. Only with enough observations, claims about the significance of parameter estimates can be made. However, since our purpose is to estimate demand for non-served city pairs, we are not specifically interested in how changes over time occur. Secondly, we only have demand data available describing the last two years, which is not enough to feasibly create such a model.

Thirdly, we find the gravity and cross-sectional models. The gravity model assumes that travel between two cities will increase with population and wealth levels and decrease with the distance (Verleger Jr et al., 1972). It is widely used in many disciplines, such as transportation and marketing (Shen, 2004). In contrast with the point-to-point model described above, the gravity model is applied typically to cross-sectional data sets (Verleger Jr et al., 1972), (Gómez-Herrera, 2013). The gravity model requires the assumption that travel across a diverse set of city pairs can be characterized by the same set of variables (Verleger Jr et al., 1972). This often leads to the result that gravity models produce satisfactory results for homogeneous sets of city pairs. This means that the scope of a gravity model is often limited to a country or a group of neighbouring countries such as in studies by Grosche et al. (2007), Kopsch (2012) and Hazledine (2017). One must be cautious with using the gravity model when applying it to explain heterogeneous markets.

Grosche et al. (2007) apply the gravity model to estimate air passenger volumes between city pairs. They specifically state that the gravity model can be applied to city-pairs where currently no air service is established and historical data is not available. Also if factors describing the current service level of air transportation are not accessible or accurately predictable they argue this is the way to go (Grosche et al., 2007). This is exactly what we need for our research purpose, and therefore we proceed with this modelling choice. How the gravity model works and a critical review on gravity model literature is described in the next section.

Variables used in air transport demand models
Jorge-Calderón (1997) distinguish two types of variables in air passenger demand models: geo-economic variables and service-related variables. Geo-economic variables contain information about economic activities and geographical characteristics of the areas where transport takes place, while service-related variables contain information about aspects that are under control of airlines, such as flight frequency, plane size and prices. For new destinations, information regarding flight frequency, plane size and prices (service related variables) are unavailable. This is an extra reason for choosing the gravity model, as it has minimal dependence on airline related data Grosche et al. (2007).

2.2 Gravity model

In this section we explain the intuition for the gravity model and its application in detail. Then we elaborate on which variables the model usually includes and why.

The origin of the gravity model and the intuition behind it
The idea behind the gravity model is based on Newton's gravity law from physics. Newton's law of universal gravitation states: *'Every particle attracts every other particle in the universe with a force which is directly proportional to the product of their masses and inversely proportional to the square of the distance between their centers'* (Newton, 1687). Equation (1) presents the mathematical formulation of the gravitation law where F_{ij} is the attractive force between i and j, explained by a gravitational constant G, the masses M of objects i and j and the distance between the objects D_{ij} squared.

$$F_{ij} = G \cdot \frac{M_i \cdot M_j}{D_{ij}^2} \tag{1}$$

The fundamental idea of Newton's law was the inspiration for creating the gravity model which aims to explain flows of information, people or goods between places by their economic mass and distance. The formula to explain these flows bears a lot of similarities to Equation (1) and is given by:

$$V_{i,j} = k \cdot \frac{(A_i A_j)^\alpha}{d_{i,j}^\gamma}. \tag{2}$$

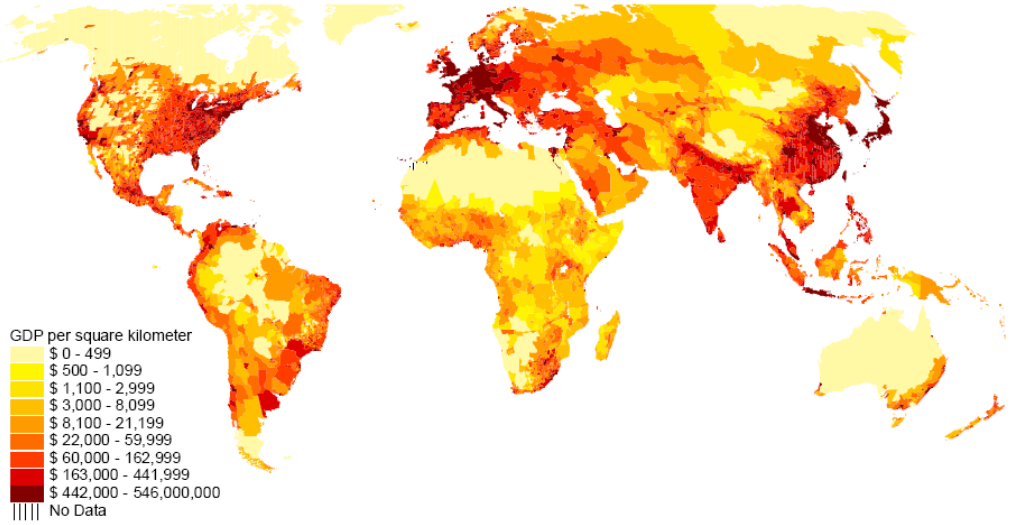
We follow the notation of Grosche et al. (2007). So in essence, the gravity model describes movement or flows between locations. It is expected that the economic size of a country or city has a positive contribution to the flow, while distance has a negative effect (Verleger Jr et al., 1972). Earlier versions of the gravity model are those of Jan Tinbergen in his book *'Shaping the World Economy'*, where he applies is to international trade, the version introduced by Harvey (1951) to explain air passenger traffic patterns and the one given by Isard (1954) to explain trade flows. The gravity model became a widely used work-horse to explain and predict flows of goods, services, finance and information ever since, as reported by Gómez-Herrera (2013) and Hazledine (2017).

To apply the gravity model to the estimation of air passenger as a measure of demand between city-pairs we need to find measures for V, A and d from Equation 2. In our model V represents the passenger flows between cities i and j. In order to get an idea of the magnitude of global air travel flows nowadays, a visualization of these flows is presented in Figure 2.[3]

[3]Figure by David Kossowsky from Bio.Diaspora on 2013 global flight data (http://www.capsca.org/Meetings/Global2012/2012CAPSCAGlobal-2-6.pdf)

Figure 2: Global flight network

Figure 3: GDP density



Next we investigate variables that can represent A in the gravity model. These attraction variables translate to the economic size of a city, which is measurable in various ways. Most common is using variables that measure income through GDP (per capita) and population in the catchment area of an airport or a combination of

these variables (Jorge-Calderón, 1997).

An example of a combination variable is GDP density, which is defined as the multiplication of GDP per capita and the population density, resulting in a variable measured as GDP per square kilometer. Gallup et al. (1999) visualize GDP density in their paper on *Geography and Economic Development*[4], displayed by Figure 3. This figure shows the intuition behind the variables for economic size in the gravity model: darker areas with larger economic size and wealth are associated with more air travel flows as presented in Figure 2. Grosche et al. (2007) report that another, aggregated, measure can be historical passenger volumes at each airport: *airport popularity*, as proposed by Doganis (1966). Other geo-economical measures such as income distribution, education levels and employment are also mentioned in the literature (Russon and Riley, 1993). The addition of one or multiple so-called attraction variables can help to explain the level of connectivity between cities such as former colonial ties, language or other political- and cultural relationships should be taken into consideration according to Russon and Riley (1993).

The aforementioned variables are in the category of geo-economic variables, other attraction variables that contribute to the A in the gravity model are service-related factors, as introduced by (Jorge-Calderón, 1997). Service-related factors are under control of the airline and determined by both the quality and price of the airline product. In the literature the quality of airline service is defined through three main variables: the frequency of departures, the load factor and the aircraft size or technology (Wang and Song, 2010). Grosche et al. (2007) state in their research: '*By excluding service-related or market-specific input variables, and using cross-sectional calibration data, the models are particularly applicable to city-pairs where no air service exists, historical data is unavailable, or factors describing the current service level of air transportation are not available.*' This is exactly what we need for out-of-network estimation and thus the usage of service related variables is minimized.

Distance d is the other core variable in the gravity model, which is expected to have a negative relation. Getting a clear estimate on the impact of distance is problematic (Hazledine, 2017; Chaney, 2018). For shorter distances the distance variable can be positively related to air transport, because it is affected by the main substitute of air travel: road transport and public transport. Besides that, distance has an effect through price, because airfares are somewhat price related due to flight costs reasons wuch as fuel and labour time. However fixed costs such as airport fees, aircraft ownership costs and crew are also reflected in the price, thus price and distance far from perfectly correlated.

[4]Since the newer version of this map is not available, the figure from 1999 is used. The numbers corresponds with other numbers for GDP an population compared to today. However, these numbers change relatively slowly and the goal is to visualize areas of wealth and economic size.

Next to that, the price contains information on the service quality of the flight, which is not driven by distance. Price will not be included in this research since, 1) there is no unambiguous price data available, especially not for out of network destinations and 2) a price variable complicates the model a lot, because of its relation to other variables.

In the gravity model it is also common to control for other variables that affect the explanatory variables and demand, see Hazledine (2017). These are factors such as (1) whether there is a direct service, (2) intensity of competition (low cost carriers), (3) size of the airport, (4) colonial ties.

2.3 Critical reflection on the gravity model

The critical review of the gravity model will mainly revolve around three papers. One of the most recent cases where the gravity model is applied to air transport is the study by Hazledine (2017). He applies a gravity model to two distinctive data sets: cross-border flights between Canada and the USA, and domestic air travel within New Zealand. He tries to disentangle exogenous and endogenous effects after he claims that supply and demand for air travel are not independent. We find this claim rather ambiguous, since supply and demand should be independent and set by price levels. He argues that extra flights generates extra demand, but this should be handled through differences in price. On the other hand an additional flight to the same destination is not a homogeneous product. Especially business oriented passenger need to be somewhere at a certain time. This makes a flight to the same destination at another time a different product. Hazledine (2017) uses a variable for price in his model. The data for price is collected by finding minimum and average prices of tickets at certain point in time before the flight. This approach is problematic since the price for tickets fluctuate a lot over time and over different flights. Ticket prices are determined based on the percentage of tickets sold and the expectation an airline has on the booking curve. Furthermore each airline has different service levels and optional extra services, which may or may not be included in the price. Passengers on the same flight rarely pay the same tariffs. This makes adding a reliable price variable to the model a difficult task. Therefore price variables will not be considered for this research. Also Hazledine (2017) applies the gravity model to two separate national market: Canada and New Zealand. However, it seems that the coefficients are estimated on the combined data set of two distinct market. Contradictory, he also argues that there is an important difference in traffic flows as Canadians can fly cross-border to the United States quite frequently, while for New Zealanders international flying is less common, because of geographic location. This is referred to as the border problem. Subsequently no validation is performed on the parameter estimates, while claiming to have a model that successfully entangles exogenous and endogenous effects.

Grosche et al. (2007) use the gravity model to explain passenger flows between Germany and 28 European countries. They focus predominantly on variables describing economic activity and geographic characteristics of city-pairs, to ensure that the model can be used to find demand for O&Ds where currently no air service exists. In order to produce a valid model they exclude O&Ds that meet one of the following requirements. They exclude typical low-cost routes, where the incentive to fly is argued to be the low price, otherwise these passengers would not have flown at all (Tacke and Schleusener, 2003). Moreover, they exclude routes under 500 km of range, since there would be strong competition by other means of transportation such as road and rail that offset the results of the gravity model. We argue that instead of excluding this data, it might be better to control for it. It might be the case that by removing these instances, you throw away valuable information. Grosche et al. (2007) use historic booking data to estimate the model, while they argue that historic bookings only partly reflect true/unconstrained demand since passengers cannot book flights that are full or do not exist. However, they argue that this is the best data available for their purpose. They perform cross-validation to examine the sensitivity of changes in all explanatory variables. Grosche et al. (2007) conclude that the model is fairly robust with an R^2 of 0.76 and is a meaningful tool in determining demand on routes where currently no direct service is offered by an airline. They create two separate models for cities that have multiple airports. They deal with it by creating extra variables capturing the effects of competing airports. However, they could also have aggregated the data on city-pair level instead of airport level. The only minor point of critique we have is the existence of (multi)-collinearity between the variables distance and travel time, which obviously correlate strongly and thus make the coefficients corresponding to these variables unreliable.

Lastly, we review the gravity model from the Verleger Jr et al. (1972) paper. They present a gravity model about air traffic in the United States for 441 city pairs. They argue that the gravity model should only be used when dealing with homogeneous cities. This means that the effect of one extra dollar of income should have the same increase in air traffic among all city pairs. As this assumption is likely to fail as the scope of the model grows larger, they advice against using one gravity model to analyze an entire air transport market. However, only analyzing a small sub-part of the market is not our goal. Therefore we might subset our data specifically to Amsterdam Airport Schiphol in order to (better) meet the requirement. Apart from this, Verleger Jr et al. (1972) added phone call data to the model as it could proxy the connectivity between cities in order to better analyze the traffic flows. Unfortunately their phone call data only contributed a little. Nonetheless the idea of adding different types of variables that capture connectivity between cities is innovative and opens up opportunities to enrich the gravity model.

2.4 Big data in air passenger demand forecasting

The magnitude of data generated and shared by businesses, public administrations and industrial and non-profit organizations has increased immeasurably (Agarwal and Dhar, 2014). Organizations use analytical techniques to explore data that can aid product and process discovery, productivity and policy-making. Michael and Miller (2013) explain that since the wide-spread adoption of the internet we are transitioning from text-based data to richer data formats that that includes associated meta data such as geo-location and date-time stamps, among other things. With the rise of big data new opportunities appear, but it also brings new challenges.

New opportunities
Michael and Miller (2013) explain the new opportunities that big data brings for companies, such as analyzing consumer purchasing trends to better target marketing. They state that: *"advances in data storage and mining technologies make it possible to preserve increasing amounts of data to yield valuable new insights. ... Big data can expose people's hidden behavioral patterns and even shed light on their intentions."* Sivarajah et al. (2017) speak of opportunities that include value creation, superior business intelligence for more data-driven business decisions and support in performance and flexibility improvements of supply chains and resource allocation. These opportunities line up exactly with what we are aiming for in our research: trying to find demand insights for routes currently not operated by airlines. The answer to our question may be found in the area of big data that uncovers consumer patterns that have previously been unobserved.

New challenges
While big data offers a wide range of new opportunities and applications, it also brings new challenges. Michael and Miller (2013) states that: *"While big data can yield extremely useful information it also presents new challenges with respect to how much data to store, how much this will cost, whether the data will be secure and how long it must be maintained."* Moreover Sivarajah et al. (2017) notice that with the reality of big data comes the challenge of analyzing it in a way that brings value. The worldwide growth of data volumes seems to out-speed the advance of computing infrastructures. Currently, only a fraction of all the data generated is analyzed for insights. Besides the technical challenges, big data also presents new ethical challenges on a topic such as privacy. Lastly (Michael and Miller, 2013) warns that big data applications could have unintended and unpredictable results. This is a consequence of the trends that data scientists seek to discover new patterns and trends that previously have been latent. When working with big data it is important to be aware of these difficulties and work act accordingly.

Big data in the aviation industry
There has already been some research in the field of applications of big data in the aviation industry. Howell (2016) performs a case study to find out if big data can bring value for London Gatwick airport, i.e. if the benefits outweigh the costs. He concludes that directly and indirectly big data analytics support the decision-making at the airport. Kim et al. (2016) developed a forecasting model for short-term air passenger demand fluctuations by making use of big data from search engine queries such as Google. However, non of these models seem to generate insight in demand on O&D level, i.e. where to and where from the passengers are flying. Zhao et al. (2018) mentions that the development of big data presents new challenges to travel-demand forecasting methods specifically in the area of data acquisition, data processing, data analysis, and application of results.

Through our case study for KLM we are granted access to meta search data. This data contains online searches for flight by potential passengers on flight comparison websites such as skyscanner.com. We think this meta search data has the potential to provide novel insights into passenger flows and demand estimation. The data we got from KLM is not personal. What these the meta data records exactly include and more specifically how we can use them as addition into the gravity model will be explained in Section 4.1. Besides that, we expect to encounter some of the challenges mentioned before, especially on the topics of big data storage and processing. How we aim to tackle these challenge will be explained in Chapter 6 as well.

3 KLM CASE STUDY

In this section the empirical case study is introduced. One of the reasons to do a case study is that we can test the suitability of the traditional gravity model versus the gravity model enriched with meta search data by applying it to real world data. We obtain this meta search data from KLM and therefore apply the model to their network. For KLM we hope to discover destinations where the current methods of KLM are unable to accurately display demand, i.e. to discover new destinations that are possibly overlooked now. The results from this case study could be the starting point for a KLM business case to evaluate the full potential of actually adding this destination to the network. Another reason to do a case study (about KLM) is that KLM is the biggest airline located at Amsterdam Airport Schiphol.

In the first chapter the problem of congested airports is introduced. An example of this problem is Amsterdam Airport Schiphol. Schiphol has been growing a lot in the last decades and has reached its current capacity limit. The *Omgevingsraad Schiphol* is a comity consisting of a variety of actors that try to balance the development of aviation and improve the quality of the environment and the area around the airport. In 2008 these parties agreed to have a maximum number of flights at Schiphol of 500,000 a year until the year 2020 Alders (2008). In the year 2017 a number or 497,747 flight movements were registered and in 2018 the maximum of 500,000 flight movements was reached, see Figure 4. The red line corresponds to the 500,000 threshold. However the number of passengers is increasing fast which suggests that demand is growing, see Figure 5. A in-depth analysis on the situation at Amsterdam Airport Schiphol including the perspectives and interests of the involved actors is provided in Appendix I.

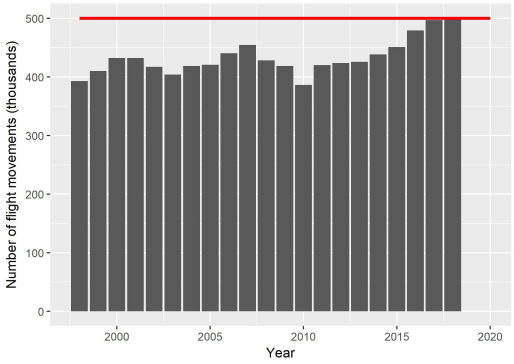


Figure 4: Schiphol flight movements

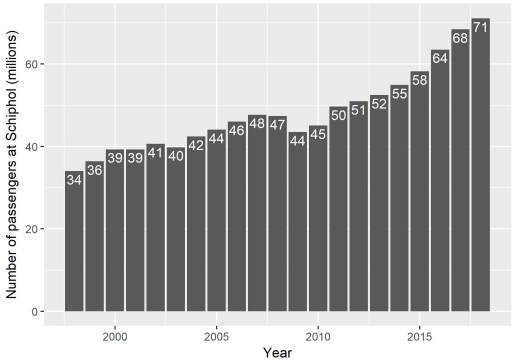


Figure 5: Schiphol number of passenger

The case study is explained in detail in the following sections. Section 3.1 introduces KLM shortly. Collaboration between airlines is increasingly common nowadays and essential for a successful business model. How this creates opportunities

for an airline to offer more destinations is explained in Section 3.2. Subsequently the current methods of air passenger demand estimation at KLM, its associated limitations and the way they decide which destinations to add to the network will be explained in Section 3.3. The benefits and risks associated with introducing meta search data will be explained in Section 3.4.

3.1 KLM Royal Dutch Airlines

KLM is the oldest airline still operating under its own name, established in 1919 by Dutch aviation pioneer Albert Plesman. KLM operates in a hub-and-spoke network and therefore classifies as a network carrier. KLM is located at her home base Amsterdam Airport Schiphol. Currently KLM offers 151 direct destinations.[5] Figure 6 visualizes the network that KLM operates by displaying all routes starting in Amsterdam. KLM, together with Air France as part of the Air France-KLM group, is one of the three main players in the competitive European aviation landscape. The other two players are International Airline Group and Lufthansa Group.

Figure 6: KLM Flight Network

3.2 Network strategy and collaboration

Different network designs for airlines exists. For this research we consider the hub-and-spoke and point-to-point networks. In a hub-and-spoke network the passengers are transported to a central node in the network (the hub) and transported to their final destination after a transfer of aircraft. In a point-to-point network only direct flights between two O&Ds are considered, without the option to transfer. An in-depth analysis of airline networks is done in Appendix II. KLM is operating in

[5] According to KLM's internal network tool Skyline

a hub-and-spoke network and therefore every destination that is added to the network (lets call it X) creates dozens of new routes for her passengers. Besides the new link created from Amsterdam to destination X, every other destination in the KLM network is connected to destination X via a transfer in Amsterdam. Transfer traffic is the main business model for KLM (70%).

Furthermore, it is important to understand a bit about collaboration in network strategy to understand the decision making process for new destinations. The (potential) profitability of a route depends, besides demand, on collaboration and competition on that route. Collaboration is common practice and is required to make routes profitable by exploiting economies of scale. Many network airlines are connected to an alliance. The main benefits of an alliance is offering a larger network through code sharing to increase market share and to lower the costs. When two airlines have a code share agreement, the airlines can sell there own tickets on a flight performed by the partner airline. Code sharing is widely used to connect the networks of partner airlines to provide many more transfer options to their customers. However, KLM also stated that for each destination they consider to add to her network, there should be a significant local market in order to prevent that they are too dependent on the transfer markets. Relying too much on transfer traffic would make them vulnerable for changes in the strategy and network of competitors. Therefore the decision on whether or not a destination is a promising candidate to be added to the network depends on the network of her partners and the network of her competitors. An extensive explanation of the multiple forms of collaboration in the airline industry is provided in Appendix III.

3.3 Current network decision-making and demand estimation at KLM

In this section the current network decision-making process and demand estimation method of KLM will be presented and how they might be advanced.

KLM's network decision-making
Whether or not a destination is a valuable addition to the network does not only depend on demand, but on many other factors. Examples of those factors are economic development, stability of the local currency and political environment, fuel price, competitor analysis and partner collaborations and operational limitations. Destination with minimal (macro-economic) risk are preferred. Based on the estimated demand an indication for the potential revenue is made and balanced against the estimated costs. The result from the newly developed gravity model is to be a starting point for the decision-making process of KLM. The advice on potential destinations is a data-driven approach for KLM to select destinations for which a business case is relevant. Currently this is done by expert judgment. The data-driven approach could discover opportunities that have not been on the radar

before. An in-depth explanation of how new destination are evaluated and what factors impact the decision-making is given in Appendix IV.

Air passenger demand estimation data at KLM
As introduced in chapter 1, traditionally bookings were done by booking agencies connected to the largest Global Distribution Services (GDS). The largest GDS such as Amadeus and Sabre provide market intelligence in the form of Market Information Data Tapes (MIDT). The MIDT data source contains booking information on true O&D level of all airlines affiliated with those GDS. Both the airline industry and academia use MIDT data for insights regarding passenger flows. Also at KLM it is common practice to use this data for market insights for network planning. However this data is not nearly a perfect substitution for market demand because of the following reasons.

1. MIDT data is incomplete
In the last ten to fifteen years it has become increasingly normal for customers to book flights directly at the airline's own booking channels, without the use of a travel agency. These bookings are logged only by the airline at which the customer booked the flight and airlines are reluctant to share this information among each other. It has become very easy and convenient to book (separate) flights at an airlines private website, a comparison site or a mobile application. The MIDT data becomes more and more incomplete and less reliable. Especially passenger flows on typical low cost routes are usually not visible in the data, since they are almost always booked directly at the airline. Therefore a considerable proportion of the market is overlooked when decisions are made purely on the demand insights of MIDT data.

2. MIDT booking data is not unconstrained demand
MIDT booking data, in the form of historical passenger flows on O&D level, is not the same as unconstrained demand (Grosche et al., 2007). Unconstrained demand is the demand for a particular route/flight/date irrespective of the capacity. Weatherford and Pölt (2002) state: "*Accurate forecasts of passenger demand are the heart of a successful revenue management system. The forecasts are usually based on historical booking data. These bookings do not reflect historical demand in all cases because booking requests can be rejected due to capacity constraints or booking control limits.*" Obtaining unconstrained demand from booking data is a topic on its own and will not be considered here.

3. Historic flows do not provide insight for new routes/destinations
Currently KLM's air passenger estimation (and forecasts) are largely based on historical passenger flows, which is a limitation itself. The main shortcoming is that historical information on existing routes does not provide insight on potential pas-

senger flows for new destinations, which is essential for decision-making about new destinations. In order words, a passenger can not book a flight to a destination that is not connected to the network.

The three reasons mentioned above make the MIDT data increasingly less reliable. KLM is very aware of these limitations of MIDT data and has various ways to deal with this.

KLM solution 1: Manual up-scaling
KLM has a rather quick fix solution available for the issue of incomplete data due to missing direct bookings in MIDT. The data is mostly incomplete on the typical low cost routes. In order to control for the missing data, passenger flows are manually up-scaled based on expert judgment from the network planning department. The risk with this is that it is not a consistent solution. Every network planner does this manually based on his/her experience with the market portfolio he/she manages. This results in a non data-driven solution with limited accuracy and efficiency.

KLM solution 2: Enriching MIDT
The second solution to counteract the incomplete data is to enrich it with additional data sources, such as T100 from the US government, Eurostat from the European Union and KLM Actuals. Each added data source only contains information on very specific parts of the market. This complex enrichment process takes a lot of time and yields only marginal improvements.

3.4 A new opportunity: meta search data

In order to solve the difficulties with the estimation of demand due to the incomplete data, we propose a new solution: combining meta search data with the gravity model. Meta search data is very promising because it can solve the limitations of MIDT data, however it also creates some new challenges.

Since December 2016, KLM logs the data stream of online search queries for all flights. Every day 25-30 million air travel search requests are send to the KLM servers by flight search engines, such as Skyscanner or Kayak. The data of these requests are logged search by search and accessible for this research. With suitable aggregation and cleaning of this data it can be used as additional source of information to perform air passenger demand estimation on city pair level. Since people search for trips from their origin to their final destination and every combination of origin and destionation can be searched for, meta search data is presumably a better source for the true O&D demand, which (partly) solves limitation 1 and 3. Regardless of whether the preferred booking class is available, potential passengers can search for flights. This solves limitation 2.

Of course, meta search data also has limitations that should be considered. We can not directly use the searches as demand, since there is a very high look-to-book ratio. This means that a lot of people search for flights, but only a small number of people actually book a flight. This is because flight search engines are often used as comparison method for passengers to gain information on the options for their travel and to compare prices. Secondly meta search data can contain biases in the type of traffic: MS data mainly contains information on the leisure market where passengers are generally price-sensitive (Kopsch, 2012). Furthermore the data reveals that business travelers generally do not search for their tickets on such websites. Another bias can arise when the flight search engines are more popular in one country than another. This is also true for the look-to-book ratio that may vary per country. We will come back on how to deal (or not deal) with these biases. Lastly there is a challenge in the size of the meta search data. The individual searches are logged in a database that contains approximately 10 TB of data and grows with a speed of 100 GB per week (approximately). This is a huge amount of data and will bring challenges accordingly. Suitable aggregation, cleaning and editing is required to create a data set of feasible size to work with.

4 THE DATA CHALLENGE

The (newly) proposed gravity model requires the availability of (high) quality data. Therefore, a large part of this research is devoted to creating an extensive data set, obtained from various data sources. This chapter is devoted to the collection, preprocessing and exploration of the data. In total seven different data sources are used, each contributing a different piece of information.

This research is heavily based on empirical data to extract insights for real world decisions. However, when working with data one should always be cautious, since data is never perfect. It will never be complete, nor free of errors. When relevant, the limitations, biases and struggles will be discussed.

This chapter is divided into four sections. First meta search, a big data source from the worlds largest flight search engines, is explained in Section 4.1. Then the other data sources we require to create the traditional gravity model are presented. The variables and features extracted from these data are explained in Section 4.2. In Section 4.3 we explain how all data sources are joined together into one big data set, and subsequently sub-setted such that we only keep the data that is relevant for the case study. Lastly we explore the data, highlighting some preliminary insights. This is carried out in Section 4.4.

4.1 Big data source: meta search

As explained, there is large potential in meta search data when we manage to extract it and prepare it in the right way. We start by looking at how the data is generated, recorded and acquired.

Nowadays, many people search for flights online. In Europe, Skyscanner is one of the leading flight comparison websites, displaying all options for a search between two cities or airports. When a search is entered by a person online, Skyscanner forwards this question to all airlines. The airline's IT systems respond to this request by returning a list of flights and their corresponding prices. For the customer, Skyscanner provides an overview of all possible flights and the corresponding information such as departure times, transfer connections (if applicable) and prices of all airlines in their system.

The reason we are interested in the meta search data is because it contains implicitly a lot of information on consumer preferences. A lot of flights are flown via the hub-and-spoke network, so a passenger flies from A via B to C. Often this passenger books flights from A to B and from B to C, possibly with different airlines, which hides the true demand from A to C. Besides that, a bias is created on the demand to large hub airports.

The added value of meta search is that it shows the consumer's preference of the true demand from A to C, since that is what consumers will search for naturally on the meta search channels. This can provide insights in the demand on routes that are not currently operated.

Data acquisition
KLM does not only collect the worldwide searches from Skyscanner, but also from 34 other Meta Search Providers (MSPs). Which MSPs these are and their associated market share for searches answered by KLM is provided in Appendix V. For every search, a log file is created and saved to a KLM database. This database has been made available for this research. It contains over 15 billion searches for the years 2017 and 2018 combined. This comes down to approximately 25 million searches per day. To give an indication of the magnitude of the data, this is 17,500 searches each minute. Each search creates a log file that is saved in JSON format. An example of such a log file is shown in Appendix VI.

The first available logged data is from the 30th of November 2016, however the system needed an installation period and not all data was logged correctly in these early days. Therefore, it was decided to use data from the 1st of January 2017 onwards. The data is stored in an Elasticsearch database. A wide selection of variables are available in the database, such as: outbound origin, destination and date, inbound origin, destination and date, cabin type, single flight or return, currency, country of the MSP, maximum number of transfer flights and number of adults, children and infants. The raw MS log is fairly extensive. For our purpose, we are most interested in how many times an O&D combination is searched for to serve as a measure for demand. Therefore, the fields that are most important for this research are origin and destination. Furthermore we are interested in the flight date, the search date and the cabin class (economy or business). The logged searches are worldwide, with a small note that for some channels a filter is applied to exclude searches for domestic flights in the USA, since it is by law not allowed to fly domestically in the USA for a non-USA airline.

By the use of a Python script and Elasticsearch queries it became possible to obtain data aggregations on O&D level per flight day. However, there was a limitation of a 120 second time out per query, which means only a few variables could be extracted within that time frame. In addition, the data was divided over two database clusters since they were in the middle of a data migration, which made the extraction more complex. In the end extractions were made on a daily level from 01-01-2017 until 31-12-2018. For extracting the data from the database, aggregating to daily level and saving it as a *.csv file Python was used with the Elasticsearch query language for aggregation per O&D on a daily level. For all other data cleaning and aggregating R was used.

Search date vs flight date
The search logs contain a field for search date and flight date. For O&D demand purpose, we are interested in the demand per flight date and not the search date. In order to create a comparable measure across days we need to introduce a (hyper) parameter for the number of days searched before the flight. In the airline industry this is generally called days before departure (DBD). Due to the limited time-out of 120 seconds at database host system a DBD of 30 days is feasible in a single query. For example, when someone searches on the 10^{th} of May for a flight on 25^{th} of May, there is a window of 15 days before departure. This approach gives us three advantages: (1) the run-time of the aggregations is workable, (2) we need only 30 days of "warm up data" and (3) the bookings made closer to the flight date are generally worth more for the airline, since prices tend to increase towards the flight date. Two disadvantages are that (1) we might introduce a bias since some markets typically book (and thus) search earlier compared to others. Secondly, (2) since people are able to search for flights up to a year in advance we do not use all data available. However, since we compare all O&D combinations on the same DBD the actual MS count does not matter that much as long as it is a (valid) proxy for the demand.

Aggregation level: time interval
The data is extracted from the database and saved in daily files. In order to perform an analysis we need to merge the files and aggregate to a higher level. Several options are available for the time period of the aggregation level for the data: daily, weekly, monthly, seasonal or yearly. Airline data typically exhibits a degree of seasonality. Especially tourist destinations are very seasonally dependent, with high demand in the summer and low demand in the winter. Airlines react to this in a way that some destinations are only flown in the summer, while others are only flown in the winter, or at a higher capacity in the winter. Airlines typically determine their flight schedule and decided upon new destinations twice a year, i.e. on a seasonal basis. Therefore the time interval we use here is on seasonal level. The aggregation level of the data set can be changed in the future researches, for example if an airline wishes to increase the frequency of the flight schedule releases. From the available data of the years 2017 and 2018, three different seasons can be extracted:

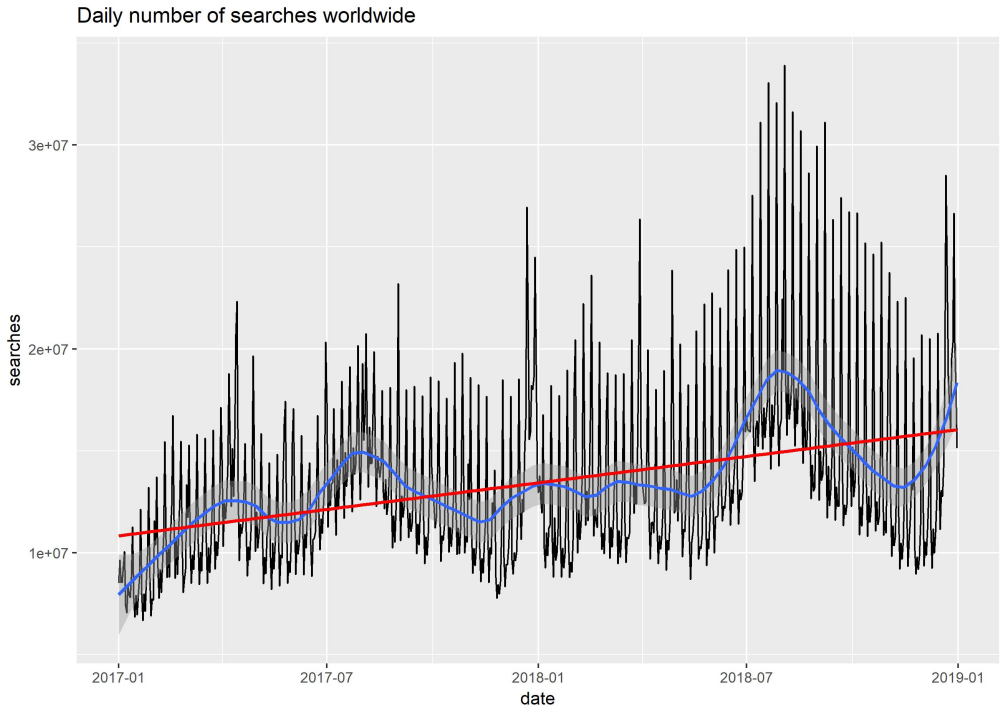
- Summer 2017: from 26 March 2017 to 28 October 2017;
- Winter 2017: from 29 October 2017 to 24 March 2018;
- Summer 2018: from 25 March 2018 to 29 October 2018.

Grosche et al. (2007) created a cross-sectional data set in a similar way by aggregating several months to one data set.

Aggregation level: airport level vs city level
Each origin (ORG) and destination (DEST) in the meta search data is indicated with a three-letter code. These origins and destinations could be airports, but also cities or train stations. We are interested in city-pair demand since, people are generally interested to travel to a city, instead of an airport as final destination. Sometimes one cities has multiple airports. The searches for individual airports or train station are summed to a city total, including the searches that were already on city level.

Statistics on meta search data
After data extraction we find 3,424,158 different O&D combinations where an actual Meta Search is logged on at least once on the 730 days. This data set contains 17,041 unique origins/destinations. To investigate what the data looks like we plot Figure 7. It shows the worldwide number of searches from 2017 and 2018 aggregated to a daily count per flight day with a DBD of 30 days. There are some patterns and trends to observe in this figure. First of all there is a weekly pattern with the most searches typically on a Friday with a dip in on the day after. Furthermore seasonal variation is present with an increase in the number of searches during the European summer months and towards Christmas (blue smoothed line). The red line shows the overall trend of the increasing number of searches.

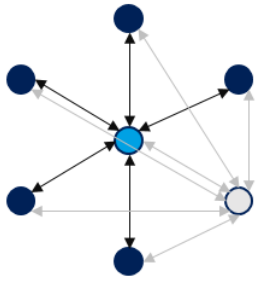
Figure 7: Meta Search daily aggregation 30 DBD



Interesting is that only $3,424,158/17,041^2 = 1.2\%$ out of all the possible direct combinations is actually searched for. A large fraction of the searches consists of O&D combinations that are only searches a hand-full number of times throughout the two years. An arbitrary threshold of 100 searches per season is set in order to filter these destinations out of the data set, while keeping the valuable information. After cleaning and aggregating to city level there are approximately 6.23 billion relevant searches left for 176,544 unique O&D combinations.

When we further subset the data for only origin Amsterdam (AMS), we find searches for 5613 unique direct destinations. Nonetheless, having the search demand insights for the entire world has a great benefit. This makes it possible to estimate demand on all one-stop transfer markets, besides the direct connections. This key contribution to the analysis for new destinations is displayed in Figure 8.

Figure 8: Meta Search insights for the transfer potential of a new destination: *the hub airport is represented in light blue, in dark blue the destinations already in the network and in light gray the potential new destination. The gray arrows represent the number of meta searches from people who want to travel between these cities.*



From the meta search data two variables are created: one for direct demand and one for indirect demand. The first variable is the number of direct meta searches per O&D summed. Secondly, the transfer potential is calculated by summing all meta searches for all potential destinations to all existing destinations in the KLM network. Only valid transfer options are considered, such as a transfer within Europe, or from one continent to another.

4.2 Traditional gravity model variables and its data sources

Besides the innovation of adding the meta search data to the gravity model, the traditional variables deserve their fair share of attention as well. For each O&D in the meta search data we want to add gravity variables that hold information about the specific origin and destination. Inspiration is taken from the literature review in Chapter 2 to come up with a selection of variables. Most of the variables are either typically required in a gravity model, other are introduced by Verleger Jr et al. (1972); Grosche et al. (2007); Hazledine (2017). This section explains how various data sources are used to create the traditional gravity model variables and why specifically these variables are chosen. Besides geo-economic variables, some aviation related variables are added that are expected to control or explain additional variation in the model.

This section explains the data that forms the foundation of the gravity model. A small paragraph is devoted to each potential variable in the model. Note that some of these variables provide information on the node in the network, i.e. on city level, where other variables provide information on the link between the notes, i.e. on the level of the O&D combination.

Origin - Destination (O&D)
In the data set, each O&D combination is a row. In total we have 176,544 O&Ds or city pairs. Information accompanying that O&D such as the geographic location (coordinates: longitude and latitude), the country, continent, and sometimes municipality is all included the data set. These additional variables are not inserted directly in the model as variables, but used to create other variables as listed in this section.

MIDT data
The current industry standard for insight in demand on O&D level is MIDT data, which is based on historic bookings. MIDT stands for Market Information Data Tapes. MIDT data is created by the commercial company OAG which incorporates the registered bookings from all large GDSs into one data set. Subsequently OAG cleans and aggregates the data to a higher level (e.g. weekly or monthly). In the end the data is sold back to airlines and airports. For this research the monthly number of passengers that traveled on an O&D is available from this source. Grosche et al. (2007) estimates the gravity model on data originating from the same MIDT data source. Regardless the limitations discussed in the previous chapter, it is the best data source that reflects large-scale true demand. For the time scope of this research about 942 million historic passenger bookings available on O&D level. The MIDT data has a skewed distribution, in the data it is observed that the top 1% most populated routes contain 40% of the worldwide traffic.

Income: GDP per capita
A measure for income is one of the cornerstone variables in the gravity model (Verleger Jr et al., 1972; Grosche et al., 2007; Hazledine, 2017). Income serves as a measure to indicate a country's economic magnitude/size, welfare level or purchasing power. The income of a city is approximated by the Gross Domestic Product (GDP) per capita of the country. We expect that GDP per capita is positively correlated with the demand for flying. The World Bank provides an overview of most country's GDP per capita, where the latest known value is considered. GDP on city level would be ideal, but this data is not available on a worldwide scale. For countries missing in the World Bank data set the GDP per capita of neighbouring countries is taken as replacement, if the countries have a similar wealth level. A very number small of cities, often from third world or very small countries, have NA values left, which are imputed with the median GDP per capita of 5806 USD per year. This only affects 1.4% of the O&D's.

Airport characteristics
An extensive open source data set lists over 8000 airports and its characteristics.[6] The variables used from the airport data are described here. Each airport is referenced to by a three-letter airport code. The data contains information on the municipality, country and continent an airport lays in. Each airport contains a size indicator (large, medium or small) and coordinates (longitude and latitude). In the previous section is explained how the data was aggregated from airport level to city level. Some cities have multiple airports. Variables for number of large/medium/small airports per city are created from this data.

Population: catchment area
Population is expected to be a very important variable in explaining demand (Grosche et al., 2007; Kopsch, 2012; Hazledine, 2017). The larger a city is, hypothetically the more people demand flights from and to that city. Population is easily available on country level, but to require that data on city level for on a world-wide scale is a challenge. Another open source database on population per city is discovered and consulted.[7]. Some larger airports serve more than just a city it is directly associated with, but also cities in the area around it. The area an airport attracts passengers from is called the catchment area Lieshout (2012). Based on these catchment areas a variable for the gravity model is created. The open source data set includes coordinates for all the cites, similarly to the airport data set. According to the size measure of the airport, circles are drawn around it and the cities within that circle are assumed to be in the catchment area. Radii of 100 km, 150km and 200 km are arbitrarily chosen for respectively for small, medium and large airports. This calculation creates the catchment area variable.

[6]Source: https://datahub.io/core/airport-codes
[7]Source: https://simplemaps.com/data/world-cities

Airport popularity
From Verleger Jr et al. (1972) and Grosche et al. (2007) we learned that a variable for airport popularity might prove beneficial (as already seen in Chapter 2). The number of seats offered to/from a certain destination are summed for the season and could explain the geo-economic attraction aspect of the gravity model.

Distance
Distance is determined using the longitude and latitude of the O&D and calculated according to the great circle distance. The Haversine formula, presented in Equation 3 is used to find the exact distance (Chopde and Nichat, 2013).

$$d = 2r \cdot \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\psi_2 - \psi_1}{2}\right)}\right) \quad (3)$$

d is the distance between to points, ψ is the longitude, ϕ is the latitude and r is the earth's (average) radius. When travel distances are small, we should control for strong competition from other modes of transport (Grosche et al., 2007; Hazledine, 2017). Therefore, dummy variables are created for distances under 300 and under 500 kilometers.

Long Haul / Medium Haul / Short Haul
Based on the continent a city is based on and the calculated flight path distance we can create a variable that indicates whether a flight fits in the European schedule (EU) or in the Intercontinental schedule (ICA) for the airline. This is of importance since airlines generally operate routes in these schedules with different aircraft types that are not interchangeable. When both the origin and destination is in Europe a dummy variable for EU is created. Another word for the typical range of a European schedule is medium-haul. When the continent of the origin and destination is not the same, and the distance is larger than 3500 km we indicate this with a ICA dummy variable (also called long-haul). There is a third category (short-haul) which applies mainly to domestic flights, which we do not have in the Netherlands.

Number of Flights and seats
The company OAG analytics provides digital flight information[8]. They claim to track 96% of the commercial flights. Through KLM we have access to this data source via an internal application called *Skyline*. The data on the number of flights per O&D is collected on city level together with a daily time stamp. The data is aggregated to a seasonal level and the scope of the data is worldwide. The variables of interest are the number of flights per O&D per month, the number of seats (capacity), the operating airline and the alliance the airline is associated with. The data is already cleaned by OAG.

The extractions are done in batches of a month, since the system could not handle more data in a single excel export at a time. The data is merged into one data file using R. The data ranges from January 2017 to December 2018. From the OAG data source, several variables are created. The number of flights on O&D level per season, the number of seats on O&D level per season, the available capacity per route.

Direct service
A dummy variable indicating whether or not there is a direct service offered by any of the airlines is also included in the OAG data set. This variable is used to subset the dataset for routes where a direct air service exists.

Partners / competitors direct service
Multiple dummy variables are created for specific airlines of interest: either a partner or a competitor. An airline wants to prevent competing with its partners, but rather build the network in collaboration. Partner airlines for which a dummy variable is created are: Air France (AF), Transavia (HV) and Delta (DL). Competitors are also of interest for the decision making on new destinations. The typical competitors of KLM are the other European network carriers. Therefore we create dummies for Lufthansa Group, that including Lufthansa (LH), Swiss International Air Lines (LX) and Austrian Airlines (OS) and the International Airline Group consisting of British Airways (BA) and Iberia (IB).

Low Cost Carrier active
Furthermore, from the OAG data source we extract whether or not a LCC is active on the O&D. This is particularly important to control for, because of missing information on LCC in the MIDT data. A dummy variable is created to capture if the two largest European LCCs (easyJet (U2) or RyanAir (FR)) have this route as direct service in their network.

[8]Official Airline Guide - https://www.oag.com/about-oag

Historical colonial ties
Two of KLM's most successful intercontinental flights are Paramaribo in Suriname and one of the Dutch Antilles Curacao. The success of these routes and associated traffic flows can be explained by former colonial ties. Therefore it is expected that a dummy variable that indicates whether or not countries have colonial ties can control for these effects in the econometric model. The data set originates from research by Wimmer and Min (2006) and is available through Github.[9]

Hub variable
At last we observe a lot of traffic flows through the worlds largest hubs. Large network carriers rely heavily on transfer traffic.

For KLM the mix of local traffic versus transfer passengers is about 30% versus 70%. We expect that the demand to cities with hub airports is higher than what the typical gravity model variables would explain. Therefore we like to have a variable indicating whether the origin and/or destination is a hub airport. There is not one definition of when an airport is a hub. Therefore it is decided to add the worlds 30 largest airports in terms of yearly passenger numbers[10]. The variable is a factor with possible three values: zero for no hubs, one if one of the O&Ds is a hub and two if both O&Ds are considered as hubs.

4.3 Data joining process and data sub-setting

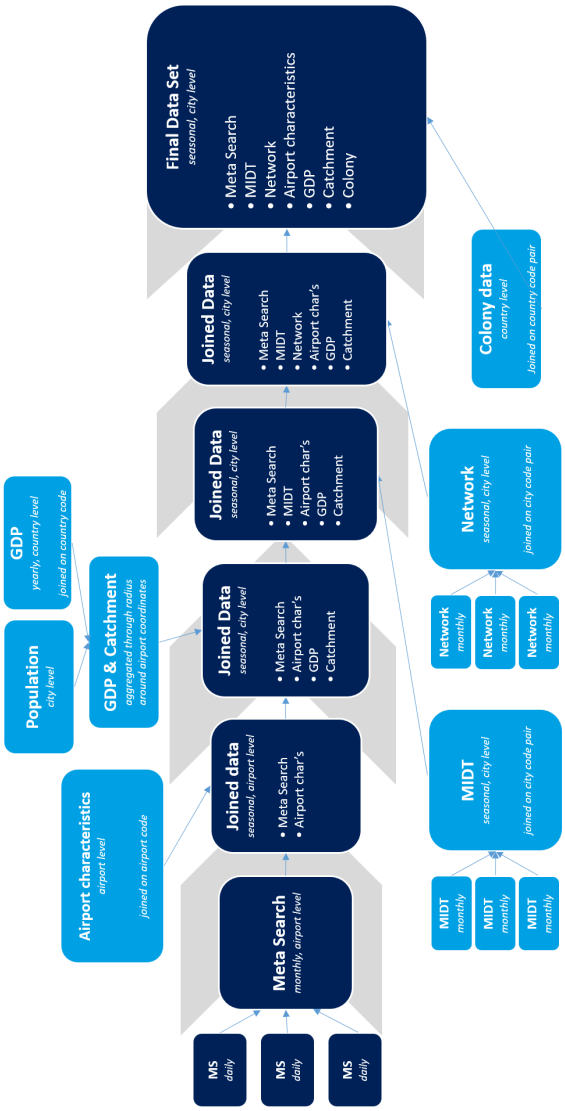
The variables explained above originate from seven different data sources. These variables are all joined together based on key variable pairs such as airport code, city code or country code. Finally we end up with one data set that includes all (world wide) information and is prepared for modelling. We subset this data such that we have the relevant information left for the KLM case. We show some summary statistics on the filtered data, followed by an exploratory data analysis in the next section.

Combining data from all data sources
The pipeline on how these data sources are joined together into one final data set is shown in Figure 9. Each of the steps are visible in the figure and are explained in the text as follows. We start with the daily meta search extractions, joined together and aggregated to a seasonal level. This data is combined with the airports characteristics data by joining it on the three letter airport code. We aggregate the data from airport to city level, by summing the meta searches for the individual airports within a city. Subsequently we create the catchment variable from the population

[9]Source: https://github.com/owid/owid-datasets/tree/master/datasets/Colonial%20Regimes%20-%20Minner%20and%20Wim%20(2006)
[10]Source: https://www.world-airport-codes.com/world-top-30-airports.html

data. Then, we join the GDP data from the World Bank to the catchment data using the ISO 3166 country codes, to join it to the main data by the use of city codes for the origin and destination. The next step is adding the MIDT data. We have monthly extractions that we aggregate to a seasonal level. Subsequently we join this MIDT data to the main data by the use of O&Ds city pairs. Hereafter we add the network data, that includes the network schedules of KLM, its partners and its competitors. Similar to the MIDT data we aggregate the monthly extractions to a seasonal level before joining the data to the main data by the key O&D city pairs. Lastly we add data about former colonial ties to the main data by again using both O&D keys. In the end we have the final data set that includes all seven data sources for 153,666 O&Ds worldwide for three seasons: Summer 2017, Winter 2018 and Summer 2018. Note that we count a return trip as two separate O&Ds here. In total the data set contains 451,563 rows and 69 columns.

Figure 9: Data Joining Flow Chart



Data sub setting and summary

The main data set that we have acquired covers the whole world. However for our case study we want to investigate which connections would be interesting to add (worldwide) from the point of view that Amsterdam is the origin or destination. So we subsetted the data to have Amsterdam as the origin or destination. The Amsterdam filtered data set contains 1861 unique true O&D pairs (return counted as two). We are left with 5517 observations and 69 columns containing variables. Now it is time to explore this data.

4.4 Descriptive statistics and exploratory data analysis

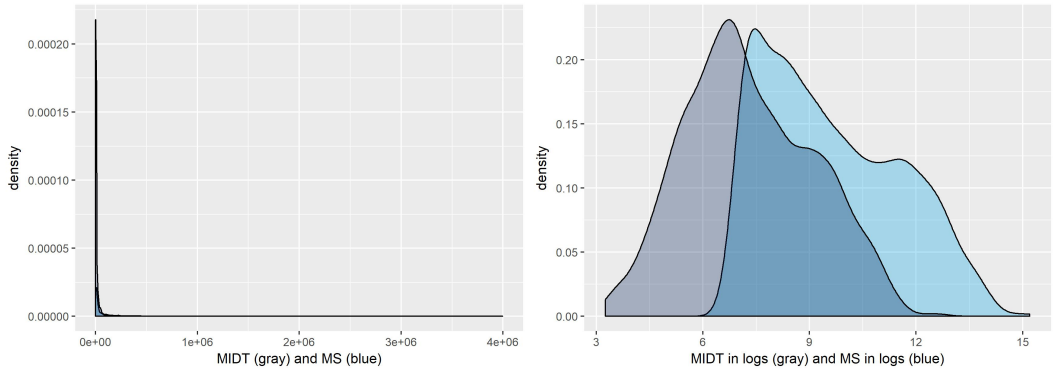
Descriptive statistics

We start by investigating the properties of the data. Table 1 shows a summary statistics of all major variables in the data set, for both the numeric and non-numeric variables. One of the first things to notice is the large difference between the median and mean for both the meta search variable and the MIDT variable. This hints to a skewed distribution. Furthermore, we find that about 26% of the O&Ds in the data set already have a direct service. On the other hand, 74% of the O&Ds are opportunities if there is enough demand. Only a very small percentages of O&Ds have LCC active or former colonial ties.

Table 1: Summary statistics AMS data set

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Meta Search	100	678	2734	56820	23168	3996637
MIDT	1	112	451	4579	2311	323404
GDP	286	17891	45638	36066	45638	100739
Airport Popularity Catchment	0	1740278	10881688	7130658	10881688	31816422
Distance	126	1788	6078	5385	7934	18700
Number of Flights	0	0	0	137.9	17	11311
Number of Seats	0	0	0	23333	2486	1657902
Direct Service	0	0	0	0.259	1	1
LCC active	0	0	0	0.0375	0	1
Colony	0	0	0	0.0161	0	1
Hub	0	0	0	0.1343	1	2

We expect that the MS and MIDT data are extremely skewed to the right. It makes sense that a lot of small O&Ds are searched and booked only very rarely while some major O&Ds are very popular. Figure 10a shows the distribution of the non-transformed data, in which this skewness is very visible. In Figure 10b the same data is log-transformed. The log-transformation helps tremendously and two somewhat similar looking distributions are uncovered. Since there are generally more searches than bookings, the MS distribution is positioned more to the right.



(a) Non-transformed (b) Log-transformed

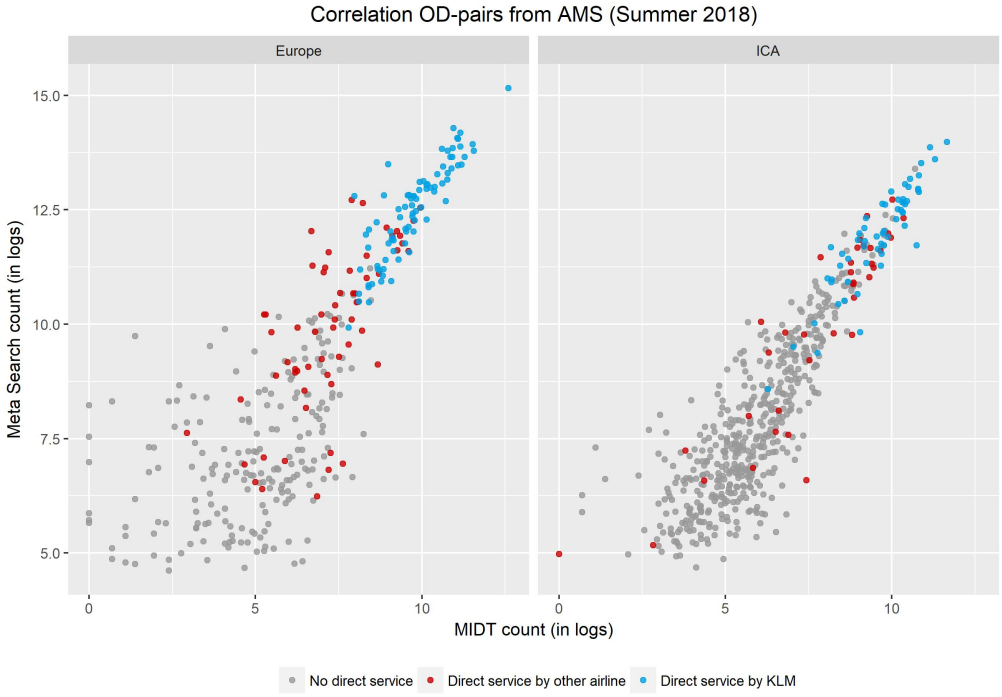
Figure 10: Density of MIDT (gray) and MS (light blue)

Next, we look at the correlation between the new meta search data and the traditional MIDT data, which is shown in Figure 11. It shows a scatter plot with MIDT count on the x-axis versus the MS count on the y-axis. Interesting to see is the strong positive correlation of 0.94 (for 739 O&Ds). This indicates that MS seems to be a strong explanatory variable for MIDT. For the extensive worldwide data set the correlation is a lot lower at 0.54 if we take all O&Ds into account (153666/2 O&Ds).

We do not show the graphs here but we find that the correlation between the MIDT data and the traditional gravity variables is a lot lower. The correlation between MIDT and GDP is only 0.11 and the correlation between MIDT and catchment is 0.30. Scatter plots for a visual reflection of these correlations are available in Appendix VII. This suggests that the meta search variable will outperform the traditional gravity model. The correlation of the number of seats and number of flights is a higher than expected, since it basically shows the supply of direct flights. The correlation between the number of seats and MIDT and the number of seats and MS is respectively are 0.92 and 0.94. However, note that the flights/seats data only shows direct routes flown, while the MIDT and MS data shows the demand for true O&D, which may or may not require a connecting flight depending on the supply.

Since the correlation between MIDT and MS is very high, we are naturally interested in the most searched destinations from the raw data. Potentially this data already provides valuable insight for the destination selection procedure. In Appendix VIII an overview is made with the top 5 destinations for the European and intercontinental schedules. Similar as in Figure 11 three categories are distinguished: the destination where KLM is already active, the destinations where other airlines are active and the destinations where no airline is active.

Figure 11: Scatter MS vs MIDT: *Each dot is a destination with Amsterdam as origin. In the plot we show Europe in the left part and intercontinental in the right part. Furthermore we observe three colors: blue for flights operated directly by KLM, red for flights operated directly by another airline, which can be a partner or a competitor of KLM and gray for O&Ds where currently no direct service exists.*



5 MODELLING, ESTIMATION AND VALIDATION

This chapter introduces the theoretical model. In Section 5.1 is explained how the model is created and there will be explained the estimation method and the associated assumptions and restrictions. In Section 5.2 the comparison metrics are introduced which help with selecting which model is 'best'. Finally, Section 5.3 elaborates on the validation method(s) in order to make a statement about the real world value and reliability of the model.

5.1 Model and estimation

The initial idea of the thesis was to create a gravity model for demand with the addition of using meta search data. The basic model for this is given by:

$$D_{i,j} = f(G_{i,j}) + M_{i,j} + \nu_{i,j}, \tag{4}$$

where $D_{i,j}$ is demand at origin i and destination j, f a function of gravity variables $G_{i,j}$, $M_{i,j}$ the meta search variable and $\nu_{i,j}$ the error term. However, there must be dealt with the issue that 'true' demand $D_{i,j}$ is unobserved. Therefore the demand $D_{i,j}$ is substituted with observed data: booking data from the MIDT data source. Since the MIDT data is incomplete, there will be introduced an extra measurement error in the following way:

$$D_{i,j} = y_{i,j} + \epsilon_{i,j}, \tag{5}$$

where $y_{i,j}$ is the observed demand (MIDT data) and $\epsilon_{i,j}$ the measurement error. This is then substitute $D_{i,j} = y_{i,j} + \epsilon_{i,j}$ into equation (4) which gives:

$$\begin{aligned} y_{i,j} &= f(G_{i,j}) + M_{i,j} + (\nu_{i,j} - \epsilon_{i,j}) \\ &= f(G_{i,j}) + M_{i,j} + \mu_{i,j}, \end{aligned} \tag{6}$$

with $\mu_{i,j} = (\nu_{i,j} - \epsilon_{i,j})$. The uncertainty effect of the model and the extra uncertainty about the incomplete MIDT data are now captured in μ_t. Adding the meta search data into the gravity model is supposed to fix part of the extra uncertainty introduced by the MIDT data. It 'fixes' this measurement error to a certain extend.

In order for equation (6) to give reliable results, a check for exogeneity of the regressors is a necessity. The following assumption must be checked:

$$E(D_{i,j}|X_{i,j}) = E(y_{i,j}|X_{i,j}) + E(\epsilon_{i,j}|X_{i,j}), \tag{7}$$

where $X_{i,j}$ are all explanatory variables. In other words, the measurement error $\epsilon_{i,j}$ should not be correlated to the regressors $X_{i,j}$. If there are other variables in $\epsilon_{i,j}$ that affect $X_{i,j}$ there will be a change of over- or under predict $y_{i,j}$. There are reasons to believe this might be the case in our model as well, since the MIDT data is incomplete.

It could be that the low cost travelers are not represented properly and that the users of meta search are likely to be highly price sensitive and biased towards the leisure segment. Also there might be an indication that the meta search variable is not exogenous. The expectation is that when a low MIDT is observed on a low cost route, a high count of meta searches are observed, which results in underestimating the β coefficient on the MS variable. Adding a control variable to the equation can help to remove this relationship. Therefore a dummy variable is added to routes where the largest European low cost carriers RyanAir and EasyJet are active.

Despite its shortcomings, MIDT is chosen as dependent variable and the MS variable is used as one of the explanatory variables to add information that is not captured by the MIDT variable. This makes it possible to check to what extent the meta-search data can explain the current observed demand. It also allows us to see where MS and MIDT give different results which is interesting. If there are a lot of searches but not that many bookings, this could indicate that this route is overlooked when only considering MIDT data. In this way, destinations can be fined that are currently predicted above or below.

Model formulation
The formulation of the gravity model within aviation, in the style of Grosche et al. (2007), is given in equation (8). This equation shows travel demand between the cities i and j.

$$V_{i,j} = k \cdot \frac{(A_i A_j)^{\alpha}}{d_{i,j}^{\gamma}}, \tag{8}$$

where $V_{i,j}$ entails the passenger volume between i and j, for $i \neq j$. The variables A_i and A_j are attraction factors of respectively i and j. Here $d_{i,j}$ represents the great circle distance between i and j, and k is a constant. The parameter α indicates the influence of the attraction variables, while the parameter γ indicates the effect of the distance on travel demand. In order to estimate the model in a linear form, the equation is transformed by taking the logarithm on both sides of the equation, which gives us:

$$log(V_{i,j}) = k + \alpha \cdot log(A_i A_j) - \gamma \cdot log(d_{i,j}) \tag{9}$$

The equations (9) and (6) can now be combined to find our preferred model: the gravity model enriched with the meta search variables and some control variables.

$$log(y_{i,j}) = k + \alpha \cdot log(A_i A_j) - \gamma \cdot log(d_{i,j}) + \beta \cdot log(m_{i,j}) + \rho \cdot log(l_{i,j}), \tag{10}$$

where $y_{i,j}$ is the MIDT known demand, k is the intercept, $log(A_i A_j)$ are all the attraction variables, $d_{i,j}$ is the distance, $m_{i,j}$ is the meta search variable and $l_{i,j}$ are all control variables. An overview of all specific variables used in the different model specifications are given in the results chapter.

Now that the theoretical model has been completed, the decision can be made as to how the model should be estimated: the way of finding coefficients for k, γ and all α's and β's. The estimation of this model is done by Ordinary Least Squares (OLS), a commonly used method for estimating parameters in a linear regression model. See also Gómez-Herrera (2013) and Grosche et al. (2007). OLS minimizes the sum of the squares of the differences between the observed dependent variable and those predicted by the linear function. The OLS estimation method comes with a set of assumptions that must hold to produce valid results when looking to find a causal relationship. However that is not our aim here, we are trying to find the model that produces the best prediction. So even if not all assumptions are satisfied, the model still has value. However, it will be checked whether the errors exhibit heteroscedasticity (non-constant variance over time). If it is determined that the errors are heteroscedastic (by examining this graphically in the results chapter), this can be dealt with by using robust standard errors with heteroskedasticity, such as the standard errors Heteroskedasticity and Autocorrelation Consistent (HAC) represented by (Newey and West, 1986).

5.2 Model comparison

The different model specifications are tested in the next chapter. To choose which model specification yields the best estimate of the demand, comparison statistics are needed to evaluate the different models.

The first metric is the R^2, see equation (11). The R^2 measures how good the line produced by the linear model captures your data points, so the 'fit' of the model on the data. So it indicates how much of the variance in the data is explained by our model.

$$R^2 = 1 - \frac{SS_R}{SS_T}, \tag{11}$$

where $SS_R = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ is the squared sum of the regression error and $SS_T = \sum_{i=1}^{N}(y_i - \bar{y}_i)^2$ is the squared sum of the total error. Now to compare (slightly) different models to each other the *adjusted* R^2 can be used instead of the R^2. The R^2 increases with every added variable, even if this variable has no explanatory power and only introduces extra (estimation) uncertainty. The *adjusted* R^2 corrects for adding extra variables to the model. Otherwise adding extra variables to the model always seems like a good idea, which makes comparison between models impossible. The formula for the *adjusted* R^2 is given by:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}, \tag{12}$$

where n is the number of observations and k is the number of regressors.

Another way to compare models is the Akaike Information Criterion (AIC). It is a measure used in Maximum Likelihood Estimation (MLE). MLE using a normally distributed log-likelihood function results in exactly the same results as OLS estimation. That is why the AIC can be used here. The AIC formula is shown in equation (13). Similar to the *adjusted* R^2 the AIC penalizes the use of more variables in the model. It makes a trade-off between a better log likelihood $(\ln \hat{L})$ and adding more parameters (k). The lower the AIC value, the better the model performs, since the log-likelihood is subtracted from the $2k$ term.

$$AIC = 2k - 2ln(\hat{L}) \tag{13}$$

The previous two measures describe the in-sample fit of the model with respect to the data and which variables enhance this fit. The whole premise of our model is that meta search data has been added to 'fix' the incomplete MIDT data. Therefore we are more interested in the predictive power of our model. Ultimately, the model that performs 'the best' is chosen, based on the case study where empirical data are used. Better performance will be measured in various ways as well. One way to evaluate the performance of the different models is to look at error measures, such as the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE), which take the form:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|e_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}e_i^2}, \tag{14}$$

with e_i as error term per observation. The error measures depend on both the prediction and the realization, meaning they depend on the prediction error $(\hat{y} - y)$ only and they are symmetric. Of course in both cases lower values indicate more accuracy. Both measures have their advantages. The MAE is a linear score which means that all observations are weighted equally. The RMSE gives relatively high weight to large errors, since the errors are squared. They do not necessarily always indicate the same model as the 'winner' and we will therefore use both to check which of our models does better.

5.3 Validation and sensitivity analysis

The main purpose of validating a model is to check if the model is reliable and if the results have real world value. This is also related to the performance of the model. As explained, the interest lies with the model that has the best predictive power. The model will be validated in two ways: cross validation to investigate the differences in estimation outcomes when leaving out a part of the data (sensitivity) and predicting back removed KLM destinations.
If a model is very sensitive to small changes, such as using slightly less or more data, the validity of the model must be questioned. Whether the outcomes are robust, is tested by doing k-fold cross-validation. This method creates k equivalent

folds and uses k-1 folds to estimate the model, and testing the models performance on the last k set. For example if k = 5 is used, each fold contains $\frac{1}{5}$ of the data. In a stable model the coefficients and model fit are not very sensitive and will not change that much using the different folds for estimation.

The most important part is to check if our model has real world value and thus if it can be used to accurately estimate demand for destinations that are not currently in the network (and signal under and over-prediction of existing destinations). This can be tested by eliminating successful routes (which therefore have high demand) that are operated by KLM from Amsterdam airport and see if it is possible to predict them back.

If the model does this accurately, the conclusion can be drawn that it has real world value. We have therefore implemented the following iterative validation procedure:

1. Remove a current KLM destination from the dataset.
2. Estimate the model on the remaining data (so all observations minus the removed destination).
3. Predict the fitted values for all observations (including the removed destination).
4. Rank all destinations that are not currently in the network according to predicted demand (based on the fitted values).
5. Report the rank of the removed destination.

We do this for all (151) direct destinations, one by one, and create a vector containing all ranks. The the mean, median and mode of the rank vector are reported for each model and the percentage of removed KLM destinations that the model predicts in the top 10 destinations. This shows the performance of the different model specifications. The intercontinental flights and the European flights are compared separately and the summer of 2018 is used as the preferred period because it contains the most data points.

The results of the sensitivity analysis and the validation check are presented in the next chapter. This helps in making a decision about the final model.

6 RESULTS

Now that the model is described conceptually it can be applied to the KLM case and look how it performs. The econometric results are elaborated on in Section 6.1. This includes an overview of parameter estimates, a comparison between several different model specifications and the decision on the preferred model. Subsequently the model's assumptions are tested with diagnostic tests and the validity of the preferred model is evaluated in Section 6.2. In the last section, Section 6.3, the results for the KLM case study are presented. These include the demand predictions for O&Ds where currently no direct air service exists. The recommended destinations based on these demand estimations are presented alongside an explanation why. Furthermore, the real world implications validity of the method is checked by looking at KLM's recent additions her flight network.

6.1 Econometric results

The conceptual model is explained with clusters of variables such as gravity variables and control variables. An overview of the specific variables used in the model is shown in Table 2.

Table 2: List of variables

Letter	Variable	Transformation	Data source	Unit	Variable type
y	observed demand	log(y)	MIDT	passengers	Dependent
m	meta search count	log(m)	Meta Search	searches	Meta search
t	transfer potential	log(t)	Meta Search	searches	Meta search
i	income	$i_i \cdot i_j$ / mean(i)	GDP per capita	USD per person	Gravity
p	catchment	$p_i \cdot p_j$ / mean(p)	population	persons	Gravity
d	distance	log(d)	coordinates	km	Gravity
c	colony	-	colony	0, 1	Control
h	hub	-	airport data	0, 1, 2	Control
ds	direct service	-	OAG	0, 1	Control
lcc	low cost carrier	-	OAG	0, 1	Control
sd1	distance < 300km	-	coordinates	0, 1	Control
sd2	distance < 500km	-	coordinates	0, 1	Control
al	nr. of large airports	-	airport data	count	Control
am	nr. of medium airports	-	airport data	count	Control
as	nr. of small airports	-	airport data	count	Control
ss	season	-	KLM schedule	S17, W17, S18	Control
s	seat capacity	-	OAG	seats	Aviation
f	number of flights	-	OAG	flights	Aviation
a	airport popularity	-	OAG	passengers	Aviation
airlines	KL, AF, DL, HV, LH, BA	-	OAG	passengers	Aviation

Many variables are available from the data set, this means that many possible combinations can be made to find the best model. Nine different model specifications are created to test how the model performs with certain clusters of variables in- or excluded from the model. For example, the pure gravity model can be compared to the gravity model enriched with meta search data. An overview of the model specifications that will be estimated is provided in Table 3.

Table 3: Model specifications

#	Specification	Independent variables
01	Gravity	i, p, d
02	Gravity + control	i, p, d, c, h, ds, lcc, sd1, sd2, al, am, as, ss
03	Gravity + control + country	i, p, d, c, h, ds, lcc, sd1, sd2, al, am, as, ss, cy
04	MS (excl. MS transfer)	m
05	MS (incl. MS transfer)	m, t
06	MS + gravity + control	m, t, i, p, d, c, h, ds, lcc, sd1, sd2, al, am, as, ss
07	MS * country + gravity + control	m * cy, t, i, p, d, c, h, ds, lcc, sd1, sd2, al, am, as, ss
08	All variables	m * cy, t, i, p, d, h, s, f, ds, lcc, sd1, sd2, al, am, as, ss, s, f, a, airlines
09	Only significant variables	tbd

Parameter estimation
The models are estimated using OLS and parameter estimates are found. These are all presented in Table 4. The continuous variables are presented with their parameter estimates on the top line, followed by the standard error between parentheses and the statistical significance displayed in stars. Three stars (***) means significant at a 0.001 level, two stars (**) at 0.01, one star (*) at 0.05 and a dot (.) at 0.1. For the control variables, only the statistical significance is reported in the table. The bottom four rows shows the various performance measures of the models.

A lot of information can be digested from this extensive table. The most important findings will be discussed. The first model that is estimated is the pure gravity model without any control variables added. All parameters are statistically significant in the direction that is expected. However, the model performs poorly with an adjusted R^2 of only 0.2. Adding the set of control variables helps tremendously at the cost of loosing significance of the income and distance variables. The third model introduces dummy variables for each country. This means that for every country a different intercept is estimated. This addition again increases the model fit to a level of 0.70 adjusted R^2.

Table 4: Parameter estimates (coefficients)

Model -> Variable	01	02	03	04	05	06	07	08	09
k	2.464	2.127	4.183	-0.786	-1.448	-1.644	2.426	1.525	-4.067
	(0.674)	(0.401)	(1.299)	(0.068)	(0.114)	(0.277)	(4.158)	(3.446)	(3.158)
	***	***	**	***	***	***			
m				0.842	0.730	0.642	0.851	0.845	0.846
				(0.007)	(0.018)	(0.020)	(0.263)	(0.269)	(0.273)
				***	***	***	**	**	**
t					0.146	0.166	0.179	0.186	0.172
					(0.021)	(0.021)	(0.029)	(0.027)	(0.027)
					***	***	***	***	***
i	0.148	0.018	-0.354			0.014	-0.756	-0.759	0.104
	(0.020)	(0.017)	(0.143)			(0.010)	(0.494)	(0.349)	(0.280)
	***		*						
p	0.443	0.240	0.345			0.001	0.023	0.008	0.015
	(0.032)	(0.017)	(0.023)			(0.012)	(0.017)	(0.016)	(0.016)
	***	***	***						
d	-0.413	-0.004	-0.024			0.050	0.109	0.242	0.212
	(0.031)	(0.030)	(0.071)			(0.021)	(0.067)	(0.063)	(0.063)
	***					*		***	**
s								-5.32e-7	
								(8.54e-7)	
f								6.42e-4	
								(1.51e-4)	
								***	
a								-4.18e-8	
								(6.60e-9)	
								***	
cy (intercept)			***				***	***	***
cy (interaction)							***	***	***
c		*						.	
h		***	***			***	***	*	
ds		***	***			***	***		.
lcc		***	***			**	.	.	
sd1, sd2		***	***			***	*	*	*
al, am, as		***	***			***	***	***	***
ss		***	***			***	***	***	***
airlines								***	***
R^2	0.198	0.552	0.708	0.800	0.804	0.819	0.902	**0.917**	0.915
Adj. R^2	0.197	0.551	0.693	0.800	0.804	0.818	0.892	**0.909**	0.906
AIC	13474	11424	10215	8543	8463	8209	6644	**6059**	6164
RMSE	1.61	1.20	0.973	0.805	0.796	0.765	0.563	**0.518**	0.526
MAE	1.29	0.938	0.716	0.602	0.592	0.568	0.385	**0.361**	0.365

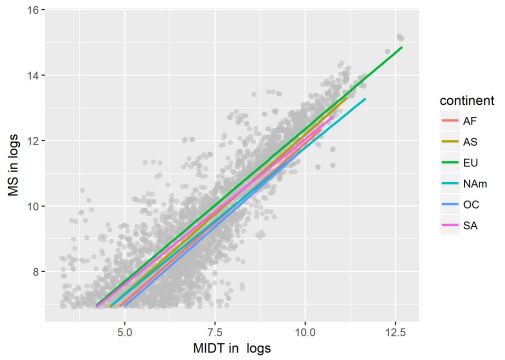


Figure 12: Model with continent interaction effects

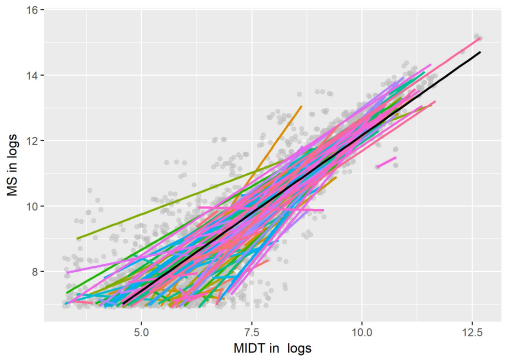


Figure 13: Model with country interaction effects

The forth model is the pure meta search model. With only a single variable, 80% of the data's variation is explained and the gravity model including control variables is outperformed. The good fit does not come as a surprise since a high correlation between MIDT and MS was already spotted in the exploratory data analysis. Since both the dependent variable as the MS variable are log-transformed the coefficients can be interpret as elasticities. This implies that a 1% increase in searches results in 0.84% more demand on average. In model five the MS transfer variables is added. Even though the model fit stays similar, the addition is highly significant. It reduces the effect previously attributed to direct meta searches. Model six is a combination of previous models, thus includes the meta search, gravity and control variables. This increases the fit slightly to 0.82. It can be preliminary concluded when MS is already in the model, the gravity and control variables describe very similar variance in the data and are only of marginal improvement. Note that both income and catchment are not significant, while the distance parameter has a small positive, significant coefficient.

Model 7 introduces an new idea. The country variable that has been added to model 3 previously, is now added as interaction variable with the meta search variable. This implies that both the intercept and slope of the regression line are slightly changed for each country. These effects are visualized for continents in Figure 12 and for countries in Figure 13. This addition improves the model fit quite a lot to 0.89 adjusted R^2. However, one should be cautious for over-fitting. The goal is not to fit the model as closely as possible to the MIDT data, but to find new destinations. The risk of fitting the data so closely to the MIDT data is that the model neglects the added value captured in the meta search data.

Model eight is the most extensive model since it contains all available variables. It includes the aviation variables such as number of seats available per O&D. Note that this data is only available for routes where a direct service is available. The fit increases slightly while the added continuous variables have really small coefficients. Two out of three added variables are significant. The added dummy variables on partner and competitor airlines are significant as well. However one should be cautious for multi-colinearity since meta search correlates highly with the number of flights (0.91), number of seats (0.94) and airport popularity (0.83). It is trivial to say that these variables also correlate highly among each other.

The ninth model contains only the significant variables and is created by backwards elimination: the least significant variable is removed until all variables are significant. The change in adjusted R^2 is almost indifferent compared to model 7 and 8, yet the $RMSE$ and MAE are improved quite a lot.

6.2 Model validation

The preferred model will be evaluated by the validation methods proposed in Section 5.3. Conclusions will be drawn regarding the real world practicality and value of this model.

Diagnostic test results
A diagnostic check for constant variance in the error term is performed. The residuals are graphically checked for heteroskedsticity. In Figure 14 the fitted values are plotted versus the actual values. Quite a good fit is observed across the diagonal. Destinations that have a direct service are plotted in light blue, while dark blue means the opposite. Figure 15 shows the variance of the residuals. A larger spread is observed on the lower end of the demand spectrum (left on the x-axis). Form this figure we suspect a slight bit of heteroskedasticity. In order to prevent for unreliable standard errors, we correct for this by making use of HAC standard errors in the estimation.

Validation results
Five-fold cross validation is performed on the model. This means that the data is split randomly in five equal folds. Subsequently the model is trained on 80% of the data and predicts on the other 20% of the data, for each combinations of folds. The total process is repeated five times, so we have a five-times-five cross validation. The parameter estimates change only very marginally. The model fit measures and error measures only change slightly as well. A weaker measure is expected since the model only uses 80% of the data in stead of the former 100% to estimate. The average R^2 is 0.87, the $RMSE$ is 0.627 and the MAE is 0.406. From these measures we can conclude that the model seems fairly stable under changing data inputs.

Figure 14: Fitted values vs Actuals

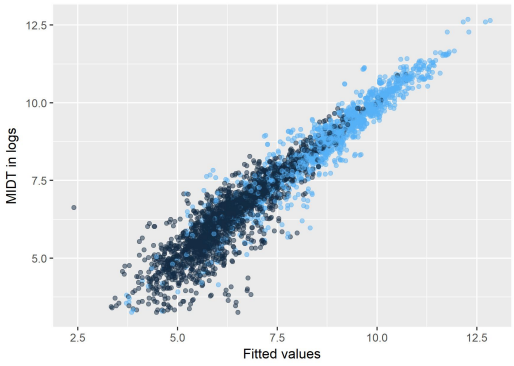
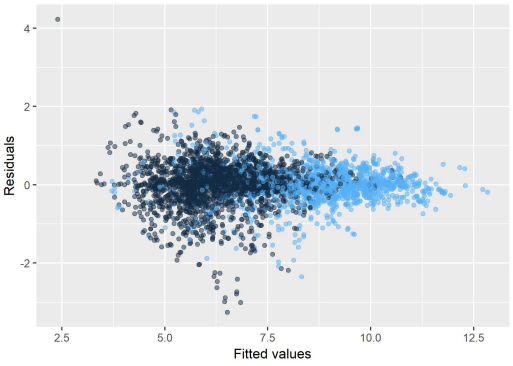


Figure 15: Residual variance



The second validation method is the back prediction of existing KLM destinations. The routine on how this validation method works exactly is explained in Section 5.3. The results of are presented in Table 5. For all nine models the mean, median and mode of the rank is presented as well as the percentage of KLM destinations that is ranked in the top 10. We observe that the first model performs quite poorly, but once the Meta Search variable is added (from model four and higher) the validation results become fairly stable. Models eight and nine score the best validation results, which is also in line with the best model fit and prediction power reported earlier.

Table 5: Validation results: predicting back destinations

Model	01	02	03	04	05	06	07	08	09
Mean rank	57.8	19.4	15.7	14.0	13.7	13.2	13.9	11.15	11.28
Median rank	30	8	7	5	5	6	6	5	6
Mode rank	3	1	2	1	1	1	1	1	1
Percentage in top 10	26.8	57.0	58.4	65.1	63.0	64.4	63.1	69.8	69.8

A visual presentation of the distribution of the ranks is shown in Figure 16 by the use of boxplots. Note that there are quite a few outliers. These are destinations that the model is not able to predict back with a high rank. This could be explained by the fact that the destinations are ranked based on predicted demand, while not all KLM destinations are selected predominantly on the largest demand. Some smaller destinations (in terms of total demand) are profitable for KLM, because of various reasons. An example is a high share of business passengers, because of an oil business. This is the case for numerous of the outliers. In order to compare the validation results between different models a histogram is made. In Figure 17 the ranks from model 2 and model 9 are compared. From the histogram it becomes clear that model 9 has many more destinations ranked at number one and performs

better overall. The results from model 8 and 9 are very similar. Overall we observe that about 70% of the KLM destinations are predicted in the top 10 with the best performing models, that the most predicted rank is number one and that the median lays around number 5 or 6. It is plausible to conclude that the model is valid for real world use.

Figure 16: Validation rank boxplots

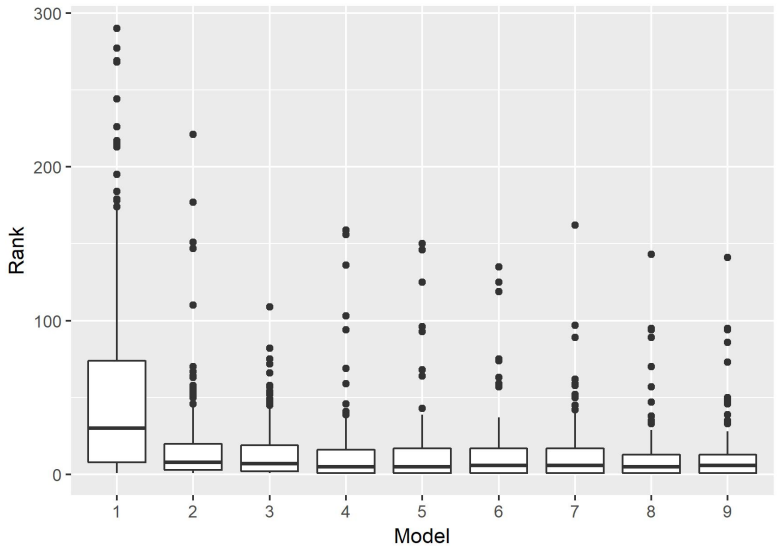
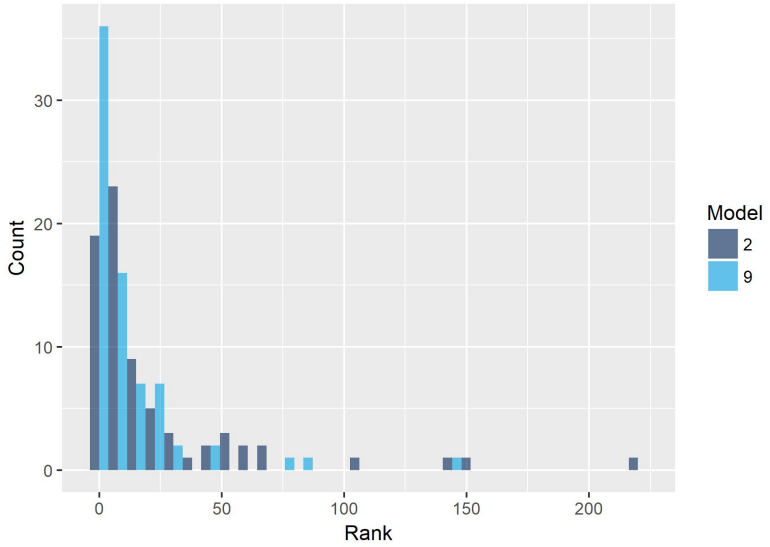


Figure 17: Validation rank histogram



Model comparison
Nine models are created whereof one will be chosen to continue the analysis with and present the final results for the case study. Four statistics are available to compare the different models among each other. Model 8 and 9 have the highest model fit measures in adjusted R^2 and AIC and the highest predictive power indicated by the error measures $RMSE$ and MAE. Besides that, they also score the best during the model validation. All statistics are very similar between model 8 and 9. Since the goal is to predict demand for city pairs where currently no direct service is available, the aviation related variables included model 8 are unavailable for those destinations. Furthermore the predictive ability is arguably more important than model fit given the goal to predict demand. For these reasons model 9 is chosen as preferred model.

6.3 Results for KLM case study

Demand predictions are created by applying the meta search enriched gravity model to the KLM case. An extensive ranking of potential destinations is provided. Furthermore the real world value of the model will be argued by looking at KLM's recent additions to the network.

Recommended destinations
Below we present the very extensive results for KLM. In order to share the results for KLM four tables are presented:

- Recommended destinations where already another direct service exists based on highest predicted demand (Table 6)
- Recommended destinations where already another direct service exists based on highest under-predicted demand from negative residuals (Table 7)
- Recommended destinations where currently no direct service exists based on highest predicted demand (Table 8)
- Recommended destinations where currently no direct service exists based on highest under-predicted demand from negative residuals (Table 9)

In each table the top 10 destinations based on predicted demand from the model are presented. The tables are split in 4 for parts, each representing a different schedule: *European Summer*, *European Winter*, *Intercontinental Summer* and *Intercontinental Winter*. Note that the destinations recommended are based on demand predictions and does not explain other aspects of the destinations decision-making. The recommendations are meant to be the starting point of a business case. Besides that if a destination does not provide enough demand to open a new route, it can be an interesting opportunity to extent code-share partnerships.

Table 6 & 7 includes only destinations where other airlines are already present. This can be a partner airline, or a competitor. When a partner is already there, KLM could either strengthen the joint position by increasing on that route capacity or leaving the partner to prevent unwanted competition. When a competitor is operating on that route KLM can either choose to compete for market share or stay away if the expected returns are too low. On the contrary, Table 8 & 9 show the top destinations for the same four schedules, this time where no other airline is active. This means that adding these destinations to the network results in achieving a monopoly on these routes. A motivation to add a destination to the network can be because of direct traffic, but it can also be because of the transfer traffic that can flow from and to this new destination. Adding a destination that is not yet connected to one of the main European hubs can be an interesting opportunity to optimally exploit economies of scale.

Table 6: Recommended destinations where already a direct service exists

#	Code	Destination	Country	Partner	Competitor
Due to confidentiality of the data, the content has been removed from the public version					

Table 7: Recommended destinations where already a direct service exists based on under-predicted demand from negative residuals

#	Code	Destination	Country	Partner	Competitor
Due to confidentiality of the data, the content has been removed from the public version					

Recommendations for intercontinental destinations where other airlines already operate and the demand is under-predicted is unavailable. The destinations Boston, Seattle and Orlando do show up, but the demand is not under-predicted. KLM is already the largest airline at Amsterdam Airport Schiphol and there are simply no intercontinental destination with under-predicted demand where KLM does not fly to, but others do. Furthermore, note that the destinations presented in grey means that the distance it too large to offer a direct service, without at least one stop for refuelling.

Table 8: Recommended destinations where currently no direct service exists

#	Code	City	Country
Due to confidentiality of the data, the content has been removed from the public version			

Table 9: Recommended destinations where currently no direct service exists based on under-predicted residuals

#	Code	City	Country
Due to confidentiality of the data, the content has been removed from the public version			

The real world value of this model becomes clear when we investigate the latest additions to the KLM network. In 2018 five new destinations are announced:

- Boston (BOS): #1 in both the summer and winter;
- Marseille (MRS): #1 in the summer and #2 in the winter;
- Wroclaw (WRO) #2 in both the summer and winter;
- Las Vegas (LAS): #3 in the winter;
- Napels (NAP): #3 in the summer and #4 in the winter from the under-predicted destinations.

It is an amazing result that all KLM's new destinations are all ranked highly since the decisions to add these destinations the network completely independent of this research and the meta search data. This indicates that our method is likely to perform well in finding the next best destination to add to the network, based on this case.

Another application of the method is to reverse it and predict which O&Ds in the current network have the lowest demand. An overview of the KLM destinations with the lowest predicted demand in provided in Appendix IX.

7 CONCLUSION & DISCUSSION

Conclusion
In order to conclude on the most important insights, we should go back to the beginning where the main research question is presented:

How can air passenger demand be estimated accurately, suitable for city-pairs where currently no direct air service exists in order to assist the flight network decision-making?

In order to answer this question a new method is developed: the gravity model enriched with meta search data. The first conclusion is that the meta search data seems to be highly valuable for air passenger demand estimation. The results give the impression that adding the meta search data improves the estimation of true air passenger demand and that it partly overcomes the measurement error that is present in the MIDT data. In the gravity model enriched with meta search data, the traditional gravity variables are not significant anymore as they are outperformed by the meta search data. This might indicate that the gravity model is not as relevant today anymore as when it was created. On the contrast, the meta search data seems to uncover patterns that were not visible before in historic booking data as the deteriorating data source MIDT. The new model is able to discover new destinations where a lot of people search for online, but where no direct air service is offered yet. The model shows its real world value by accurately ranking the five latest additions to the KLM flight network highly. For KLM the model could contribute insights to allocate resources more efficiently. For Amsterdam Airport Schiphol the model might contribute to a more efficient network that helps to defend the competitive position of Schiphol on the short term while it is growth constrained. The model can be extended to other airlines and constrained airports as well.

Discussion: academic implications
Despite the promising results, the meta search data as well as the model has some limitations which one should be aware of. It is important to realize what the data can explain and what not. One should note that there are often many more meta searches than bookings, as people generally use meta search to explore travel options and compare prices. Furthermore it is the price-sensitive leisure passengers that typically compares flights the most and thus logs the most searches. Meanwhile it is the business passenger that contributes the largest margin for an airline. Therefore the meta search data seems to be a good indicator for demand volumes, but likely fails to provide reliable insight for the distribution between leisure and business traffic, which in the end determines the profitability and success of a destination.

Furthermore, one should be cautious with over-fitting with the preferred model, as it includes many variables to correct for all country's interaction effects and intercepts. For predicting purposes it could be that a model with a little less control variables actually predicts better on new data, because otherwise the model fits the measurement error in the MIDT too closely. We tried to tackle this challenge with the validation method to predict back all KLM destinations and find that the most extensive models are the most accurate.

In search for the optimal model specification, many models were created in order to measure the effects of each cluster of variables. Several interesting findings are done. The gravity variables were not significant anymore after enriching the model with meta search data. This means that the gravity model is not able to explain any additional variation in the data after the meta search data is added. This can have various reasons (not exhaustive): 1) The gravity model might not be suitable for applying to a large scale across heterogeneous markets, even with one fixed point of origin or destination. 2) The data on the typical gravity variables might not be fine-grained enough, for example GDP per capita is measured on country level in stead of the ideal city level. 3) The effects of the typical gravity variables might be already captured implicitly inside the the meta search data. This would make the meta search data very valuable and superior to the gravity model. Two academic conclusions could be drawn from this case study:

- Meta search data seems to be highly valuable for air passenger demand estimation;
- The gravity model might not be as relevant today as when it was created.

Discussion: societal implications
The gravity model enriched with meta search data seems to shows real world value as it accurately ranks the most recent KLM destinations highly. The model contributes to the network decision-making of the airline by providing a prediction for the unobserved true demand. Therefore it might uncover destinations where previously no attention or interest had been based on the deteriorating MIDT data. The advanced method for demand estimation might lead to a more efficient network in terms of resource allocation implying higher load factors and lower emissions per passenger.

While an improvement in resource allocation might be attributed to more advanced demand estimation, the tension in the aviation industry will stay. The number of flight movements is not expected to decrease with this method, since airlines try to allocate their fleet as optimal as possible. Rather, in a more optimal network, the same number of flights is more likely to lead to higher profitability levels and more passengers transported while keeping the CO_2 emissions and noise pollution constant.

As the public debate on the capacity limit of Schiphol continues, the involved parties keep their conflicting interests. A well developed and more efficient KLM network helps Amsterdam Airport Schiphol to offer better connectivity within its capacity constrains. A more efficient network has economic benefits as well as reduction in emission per passenger. While Amsterdam Airport Schiphol is not able to grow until at least to 2020, this model might be able to contribute partly to defend Schiphol's competitive position towards growing airports in the short run. In the long run other solutions and technical innovations are required to make the aviation industry sustainable from an environmental point of view. Nonetheless, the method keeps value as airlines and airports are always interested in a network that allocates the resources the more efficient. During this research the case study is applied to the KLM network on the constrained Amsterdam Airport Schiphol. However, the developed method can be applied to other airlines and airports as well.

Further research
The analysis showed that meta search data mainly reflects the low yield part of the market, i.e. the price sensitive leisure passengers. The model for air passenger demand estimation on big data might be improved in the future if more big data sources come available that also reflects other yields of the market. If so, this method might sketch a more complete image of the true unobserved demand.

In spite of that, this research is just one application of meta search data. The potential of this source of big data does not stop here. Further research could be done to add meta search data to other air passenger demand models besides the gravity model. Moreover, the time dimension in meta search data is left largely unexplored. Applications to forecast demand into the future or models that explain (changing) prices are on the horizon.

BIBLIOGRAPHY

Agarwal, R. and Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for is research.

Alders, H. (2008). Alders akkoord 2008 (kamerbrief).

Cetin, T. and Eryigit, K. Y. (2018). Estimating the economic effects of airline deregulation. *Journal of Transport Economics and Policy (JTEP)*, 52(4):404–426.

Chaney, T. (2018). The gravity equation in international trade: An explanation. *Journal of Political Economy*, 126(1):150–177.

Chopde, N. R. and Nichat, M. (2013). Landmark based shortest path detection by using a* and haversine formula. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2):298–302.

Daft, J. and Albers, S. (2015). An empirical analysis of airline business model convergence. *Journal of Air Transport Management*, 46:3–11.

Doganis, R. (1966). Traffic forecasting and the gravity model. *Flight International*, 29:547–549.

Fernandez de la Torre, P. E. (1999). *Airline alliances: The airline perspective*. PhD thesis, Massachusetts Institute of Technology.

Gallup, J. L., Sachs, J. D., and Mellinger, A. D. (1999). Geography and economic development. *International regional science review*, 22(2):179–232.

Gómez-Herrera, E. (2013). Comparing alternative methods to estimate gravity models of bilateral trade. *Empirical Economics*, 44(3):1087–1111.

Grosche, T., Rothlauf, F., and Heinzl, A. (2007). Gravity models for airline passenger volume estimation. *Journal of Air Transport Management*, 13(4):175–183.

Group, R. S. (2018). Royal schiphol group facts and figures.

Harvey, D. (1951). Airline passenger traffic pattern within the united states. *J. Air L. & Com.*, 18:157.

Hazledine, T. (2017). An augmented gravity model for forecasting passenger air traffic on city-pair routes. *Journal of Transport Economics and Policy (JTEP)*, 51(3):208–224.

Howell, C. (2016). Big data: Does it add to the bottom line? *Journal of Airport Management*, 10(4):326–333.

Isard, W. (1954). Location theory and trade theory: short-run analysis. *The Quarterly Journal of Economics*, pages 305–320.

Jorge-Calderón, J. (1997). A demand model for scheduled airline services on international european routes. *Journal of Air Transport Management*, 3(1):23–35.

Kim, S. et al. (2016). Forecasting short-term air passenger demand using big data from search engine queries. *Automation in Construction*, 70:98–108.

Kopsch, F. (2012). A demand model for domestic air travel in sweden. *Journal of Air Transport Management*, 20:46–48.

Lieshout, R. (2012). Measuring the size of an airport's catchment area. *Journal of Transport Geography*, 25:27–34.

Lordan, O. and Klophaus, R. (2017). Measuring the vulnerability of global airline alliances to member exits. *Transportation Research Procedia*, 25:7–16.

Luchtvaartnota (2009). Luchtvaartnota: Concurrerende en duurzame luchtvaart voor een sterke economie.

Michael, K. and Miller, K. W. (2013). Big data: New opportunities and new challenges [guest editors' introduction]. *Computer*, 46(6):22–24.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Newton, I. (1687). *Philosophiae naturalis principia mathematica*, volume 1.

Redondi, R., Malighetti, P., and Paleari, S. (2012). De-hubbing of airports and their recovery patterns. *Journal of Air Transport Management*, 18(1):1–4.

Russon, M. and Riley, N. (1993). Airport substitution in a short haul model of air transportation. *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, pages 157–174.

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

Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70:263–286.

Tacke, G. and Schleusener, M. (2003). Bargain airline pricing–how should the majors respond. *Travel and Tourism White Paper Series, Simon, Kucher and Partners, Bonn*.

Verleger Jr, P. K. et al. (1972). Models of the demand for air transportation. *Bell Journal of Economics*, 3(2):437–457.

Wang, M. and Song, H. (2010). Air travel demand studies: A review: . *Journal of China Tourism Research*, 6(1):29–49.

Weatherford, L. R. and Pölt, S. (2002). Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. *Journal of Revenue and Pricing Management*, 1(3):234–254.

Wimmer, A. and Min, B. (2006). From empire to nation-state: Explaining wars in the modern world, 1816–2001. *American sociological review*, 71(6):867–897.

Zhao, Y., Zhang, H., An, L., and Liu, Q. (2018). Improving the approaches of traffic demand forecasting in the big data era. *Cities*.

APPENDICES

Appendix I: Implications of Schiphol's capacity constrains

Amsterdam Schiphol Airport, Europe's third largest hub-airport has reached its capacity limit. This harms the growth opportunities of the airport and the main airlines operating at Schiphol, which subsequently threatens their competitive position. This is a threat to the development of the international connectivity of the Netherlands and the economy depending on it (Luchtvaartnota, 2009). The societal relevance of this problem and the urgency to take action is explained in Section 6.3. Since the future capacity at Schiphol airport is highly uncertain, the only way to develop the connectivity is not in absolute number of flight movements, but to plan these flights more efficiently: optimize the network of destinations. In order to find new destinations, we need to understand the prerequisites that make a destination viable. The most important factor is the demand: the number of passengers willing to go travel that destination. When the demand is a lot smaller than the number of seats in the aircraft, the route is bound to fail. Therefore not all routes are served with a direct flight, but require a transfer at a hub airport. Some background information on different types of airports is provided in Section 7. How the hub function of an airport works and how this impacts the decision making on destinations is explained as well. Specifically the hub function of airports that accommodate and stimulate transfer passengers is important to understand before we dive into air passenger demand estimation.

The problem: Amsterdam Airport Schiphol is full
Amsterdam Airport Schiphol has been growing a lot in the last decades and has reached its current capacity limit. The *Omgevingsraad Schiphol* is a comity consisting of a variety of actors that tries to balance the development of aviation and improve the quality of the environment and the area around the airport. In 2008 these parties agreed to have a maximum number of flights at Schiphol of 500,000 a year until the year 2020 (Alders Agreement, 2008). The capacity limit is enforced not due to physical capacity limitations, but due to noise constraints. In the year 2017, 497,747 flight movements are registered.[11] In the year 2018, the maximum of 500,000 flight movements was reached. In the figures we see strong growth in the last few years, with the threshold displayed by the red line.

While Schiphol has reached its maximum operational capacity, the overall aviation industry is growing, especially in the upcoming economies such as China and India.[12] Amsterdam Airport Schiphol is not able to accommodate the growth of the industry because of the capacity limit. The problem is relevant and urgent because

[11] https://www.nrc.nl/nieuws/2018/01/08/schiphol-net-onder-maximum-aantal-vliegbewegingen-a1587513
[12] https://www.iata.org/publications/store/Pages/20-year-passenger-forecast.aspx

there is no clear cut solution for it right now. The limit of 500,000 flight movements per year will be maintained until 2020. Initiatives on re-negotiating the current limit are ongoing, but an expansion of the capacity limit is highly uncertain due to the public resistance on the topic.[13] Especially concerns regarding noise pollution and CO_2 emissions are raised by residents or the area.

This makes the problem we are addressing a typical multi-actor problem where the perspectives of the involved parties are unaligned. Several actors with their corresponding interests and involvement are discussed. First we have Royal Schiphol Group facilitating the airport infrastructure and acting as a hub. A hub airport is defined as an airport that one or more airlines use as a transfer point for passengers to switch planes to continue to their final destination. More on how an hub airports works exactly is explained in the next Section 7. The airport is used by the airlines transporting passengers and cargo over their network of destinations. Most of these airlines operate in collaboration with partner airlines, while competing with other airlines. Partners and competitors are often, but not always, linked to the alliances airlines operate in. More about the collaboration between airlines is described in section 3.2. Moreover, there are local residents around Schiphol airport which heavily oppose expansion of Schiphol airport. The nation's secondary airports Rotterdam-The Hague Airport, Eindhoven Airport and Lelystad Airport try to capture some of Schiphol's spill over traffic now that the number of flight movements is constrained. There is on-going discussion on a topic of opening Lelystad airport for medium-haul 'holiday' flights to relieve Schiphol from some of her flight movements such that Schiphol can focus on developing its hub connectivity further. There is a lot of political and public debate on the extra emission and noise pollution of additional air traffic, which delays the opening of Lelystad airport for commercial traveling.[14] The last key actor is the government that underlines the importance of the Dutch aviation sector and specifically the network of Schiphol and KLM (Luchtvaartnota, Alders agreement). The goal in government is stated in the Luchtvaartnota (translated): *"to further develop the optimal connectivity network while maintaining the competitive position and improving sustainability of the dutch aviation.* The Luchtvaartnota (2009) is a Dutch policy document about the development of the aviation industry.

Implications of Schiphol's capacity limit: the relevance and urgency
Schiphol has reached a respectable number of 71 million annual passengers in 2018, whereof 36% is transfer passengers (Group, 2018). The home market, also referred to as catchment area, of Schiphol is rather limited which means that the airport heavily relies on transfer traffic. Schiphol Airport is currently ranked 3rd largest

[13] https://nos.nl/artikel/2260816-milieueffectrapport-schiphol-kan-doorgroeien-naar-540-000-vluchten.html

[14] https://nltimes.nl/2018/02/21/lelystad-airport-opening-delay-blow-schiphol-budget-airlines

airport of Europe, and wants to maintain this competitive position at least (Luchtvaartnota, 2009). Furthermore the Luchtvaartnota states that a high quality network is crucial to achieve the country's ambitions to be a top international business location. Moreover the government values that the airport stimulates the economy directly and indirectly by creating employment. They address that not only the number of flights, but especially the quality of the network is crucial to support international business activity. The quality of the network is expressed in number of (unique/relevant) destinations, flight frequencies and valid transfer options with reasonable connection times. In the Luchtvaartnota (2009) the Dutch Government explicitly states that further developing the worldwide connectivity is key for being an interesting international business location and being an important driver of the Dutch economy.

In a world with a growing aviation industry, an increasing number of relevant destinations and a very competitive European aviation landscape, Schiphol is at risk to loose its competitive position to other highly ranked and fastly growing European airports such as Istanbul or Frankfurt. When Schiphol is not able to grow it finds difficulty to add new destinations in growing economies. Airlines will be more likely to choose for growing airport offering more transfer options for the airline and her partners. Redondi et al. (2012) analyzed the phenomenon de-hubbing or airports. This occurs when a large airline leaves an hub airport an the number of transfer flights is significantly reduced. They conclude that airports experiencing de-hubbing did not recover their original traffic and that de-hubbing is likely to be irreversible once happening. While it is unlikely that large airlines will abandon Schiphol immediately, the capacity limit puts a serious thread to the competitive position of Schiphol airport. Schiphol is not able to accommodate that growth, which harms the development and competitive position of Schiphol as mainport and economic engine of the Netherlands and north-west Europe.

However, it is not the first time an airport is constrained in its capacity. London Heathrow is an example of a congested and physically constrained airport for years. Here we see that flights are scattered over multiple airports in and around London, harming the transfer capabilities of airlines and her passengers. However we also observe that while London Heathrow's number of yearly flight movements is lower compared to Schiphol (23.000 less), the number of passengers is higher (10 million more) (Royal Schiphol Group Facts and Figures, 2018). This indicates that average number of people per aircraft is higher, thus the usage of the flight movements is more efficient.

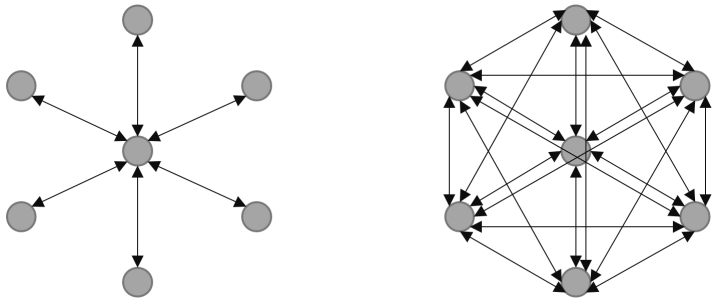
Under the likely scenario that Schiphol cannot grow in the near future, we need similar efficiency upgrades at Schiphol. The hub function of Schiphol is essential to offer transfer passengers a large network of destinations. In order for Schiphol

to keep its top position as one of the largest and relevant hub airports of Europe, it needs to keep developing its network. KLM Royal Dutch Airlines is with 48 percent of the 2018 annual passengers by far the largest and most important network carrier at Schiphol to perform the transfer flights providing the extensive connectivity network for Schiphol (Royal Schiphol Group Fact and Figures, 2018). Because KLM cannot grow in absolute terms, it needs more efficient network development to react to the growth constrains. Therefore we introduce a KLM case study in this research to apply our models to and find the most promising new destinations to add to the existing network. More information about the case study and specific KLM information is provided in Chapter 3.

Appendix II: Airport hubs

Now that we have the main motive for the research we further provide background information on how an airport works, specifically the hub function of airports that accommodate and stimulate transfer passengers. Two main types of network design for airlines are distinguished: the hub-and-spoke network and the point-to-point network. A symbolic representation is shown in Figure 18. The network design an airline chooses to adopt says a lot about the strategy and characteristics of an airline. The advantage of a hub-and-spoke network is that an airline needs less flight movements to connect each node, compared to a fully connected point-to-point network. Equation 15 shows the formula to calculate the number of links for the different network types. In a network with seven cities, the hub-and-spoke network has consists of six links, while the fully connected point-to-point network requires 21 links to directly connect the same number of cities.

Figure 18: Network design: hub-and-spoke (left) vs point-to-point (right)



$$links_{hub} = n - 1 \qquad links_{p2p} = \frac{n(n-1)}{2} \tag{15}$$

The largest airports in the world facilitate a hub-and-spoke network design for airlines. The largest airports in Europe in terms of annual passenger numbers are London Heathrow Airport (80.1 million), Paris Charles de Gaulle (72.2 million), Amsterdam Airport Schiphol (71.1 million), Frankfurt Am Main Airport (69.4 million) and Instanbul (68.2 million) all rely on their hub function (Schiphol Facts and Figures, 2018). These airports facilitate Europe's largest network airlines that transport passengers by bringing them to their hub, the central node in their network, and from there bring them to their final destination via another flight. By this strategy Amsterdam Schiphol Airport and its largest airline KLM have been able to grow far beyond the size of their own catchment area. Last year, 25 million passengers (37%) were transfer passengers (Schiphol Group Facts and Figures, 2018).

Associated with the airport layout is the airline's strategy. In general two strategies are associated with the network types: the full service carrier, or also called

the network airline, who operates in a hub-and-spoke network type and the low cost carrier (LLC), who operates in a point-to-point network. This approach is a bit simplified, because in reality there is a full spectrum in between these two typical strategies. Recently there is the trend that airlines become more hybrid in their strategies (Daft and Albers, 2015).

Full service carrier is an airline that operates in a hub-and-spoke network design and transports passengers using transfer flights. The operations of network carriers are typically centered at one hub or home base, but multi-hub systems do exist as well. The entire network schedule is designed to offer as many transfer connections as possible, prioritized on the importance of all the specific markets. This system providing transfer possibilities for passengers throughout the whole network. Typically the network is split in an intercontinental part (long haul), and an short- and medium-haul part to bring passengers to the hub for a transfer flight. This is called a *feeder function* and has as goal to exploit economies of density on the more profitable (intercontinental) routes.

The hub system is used extensively in the airline industry. Therefore is crucial to understand airline networks when estimating city-pair demand. The demand estimations are positivally biased towards the hub if we only look at direct passenger flows. In a hub-and-spoke system, we measure the passenger flows on the links in the left figure, while we are interested in the true O&D demand, i.e. all the links in the right figure. However, in reality the world is partially connected by multiple hub-and-spoke networks. This makes measuring the underlying true demand very complex. A lot of traditional airlines exploit a hub operation, while we observer a growing number of direct point-to-point connections are added to the global network.

In contrast to network carriers, there are the low cost carriers that operates in a point-to-point network design. The point-to-point network is partially connected, because operating on all links increases quadratically with the number of nodes in the network, which is hardly ever feasible from business point of view. The airline selects the links where she expects the most profit and operates a service between these two cities. By making use of a partially connected point-to-point network, the carrier has a simpler network schedule to design, since it does not offer transfer flights. Low cost airlines often open a base on a medium size or regional airport, and position a small number of aircraft there. From this base the profitability of destinations is explored and changes to the network can be made easily and rather quickly, since each link is a service on its own (a direct flight). This is the opposite of a full service airline which accommodates transfers, so that a change in one destination results in a change for the whole network.

Appendix III: Collaboration in the airline industry

In 1978 the US aviation market changed because of the Airline Deregulation Act, subsequently followed by the liberation of the European aviation market in the 1980s and 1990s (Cetin and Eryigit, 2018). Formerly one (national) airline had the rights to fly between two countries. Real competition between airlines started after the liberalization. Today, aviation is one of the most competitive industries and the profit margins are low. Ironically, because of the competition, collaboration is required to make routes profitable and to exploit economies of scale. There are multiple levels of collaboration between airlines, which will be discussed in this section arranged from the most intense form to the least intense form of collaboration.

The most intense form of collaboration is a merger as Air France and KLM have done. As said before, AF-KLM operate in a dual-hub system, which grands them benefits from a network strategy point of view. Since both airlines focus on connecting passengers, the network schedule should offer as much feasible transfer options as possible. AF-KLM are synchronizing their intercontinental network to offer even more options to their shared customers.

The second most intense form of a collaboration between two or more (separate) airlines is a joint venture (JV). A joint venture is an agreement between two airlines to share the revenues, costs, and profits, on a specific (set of) routes. Because the revenues, costs and risks are shared by multiple parties, it is possible to offer a more efficient, cheaper and larger network. In a joint venture, decisions on schedules and frequencies on selected routes will be made in a way that the two airlines operate as one, to maximize the profit. Currently, KLM has multiple joint ventures, with Delta Air Lines, Alitalia, Kenya Airways, China Eastern and China Southern.

Lastly there are airline alliances. Due to the increasing competition, major airlines formed alliances. One of the main benefits of an alliance is to create code shares. A code share is a less intense form of collaboration than a JV. When two airlines have a code share agreement, the airlines can sell there own tickets on a flight performed by the partner airline. Code sharing is widely used to connect the networks of partner airlines to provide many more transfer options to their customers. Currently KLM has approximately twenty code share partners (excluding her JV partners). The main benefits of an alliance is offering a larger network through code sharing to increase market share and to lower the costs. Besides the extended network, the benefits of an alliance include the reduction of costs by sharing sales offices, airport lounges, passenger services, maintenance facilities, operational facilities such as catering or computer systems, operational staff such as ground handling personnel and investments and purchases by negotiating discounts on e.g. aircraft and fuel due to extra volume (Fernandez de la Torre, 1999). Note that airlines are

not obligated to be in the same alliance to have code share agreements, but alliance partners make natural code share partners (Lordan and Klophaus, 2017). Worldwide there are three main alliances: Skyteam, Star Alliance and Oneworld. KLM is a member of Skyteam.

Appendix IV: Network decision-making

Now we arrive at the question how airlines choose their destinations. Finding new destinations for an airline is a complex problem without one unique correct answer. Finding a suitable next best destination depends on a lot of factors. Let me introduce this in a general form and look at the aspects that make a destination viable. The key precondition/requirement is the existence of enough demand on a new route. For network development purposes, airlines are specifically interested in demand to destinations they do not offer, since they are able to measure demand quite effectively through the bookings they receive for the destinations they already offer. The starting point of evaluation a destination's potential is a measure for demand on that origin-destination combination.

Next to finding an accurate measure for direct demand, which is a complicated task itself, the decision-making process includes demand measures for transfer traffic. If we delineate our scope specifically to network carriers, we understand that airlines does not only want to know the demand on from/to all possible destinations to their hub. They equally interested in the direct demand from/to every possible new destination to all the relevant transfer destinations. We call this the indirect demand. Lets say an airline wants to add a new regional destination, it is interested to know the size of the demand from all the intercontinental destinations it already offers to this new destination to see you many additional transfer traffic they could attract because of this new regional destination. Since these connections are not offered already, no/limited information about historic bookings is available to create demand estimations from. The demand is partially unobserved, because there is no accurate way to measure the number of people that would have wanted to travel on a route that is not offered.

Besides the key factor demand, strategic, tactic and operational aspects determine the outcome of the network decisions as well. On the strategic level we have the type network connectivity that an airline aims to provide which should be in accordance with the fleet composition of an airline. Aircraft are ordered many years in advance to delivery. Besides that the life cycle of an aircraft is approximately 25 years. This means that airlines make decisions on orders now for aircraft that will be used in service until the year 2050 roughly. Large uncertainties such as the future capacity of an airport (Schiphol as example here) forces airlines to delay there strategic decisions on fleet renewal untill more information is available.

On a tactical level we have the multi-actor arena an airline lives in. Competition in the industry is fierce, margins are low and fluctuating oil prices have huge effects on the overall profitability. Most network carriers operate in alliances where they collaborate with other airlines to exploit economies of density. Network carriers face competition from each other and form low cost airlines expanding market

share with their low prices. Adding new destinations to the network is a multi-actor game, where one responds to the other. Emerging markets that pass an (unknown) demand threshold are often only profitable for only one airline to operate in. If two airlines compete for market share, no one will win. Therefore there is a first mover advantage for growing market that are poorly connected internationally. Moreover, airlines operate in many different countries and deal with all kinds of political environments and economic situations. The revenues are often obtained in the local currency, while the costs are made in the home currency. An unstable political environment or volatile exchange rates are valid reasons not to add a destination to their network, since the expected profitability is highly uncertain.

Lastly we have operational factors such as weather, airport infrastructures and aircraft capabilities (e.g. range). Two examples are whether or not the airport can facilitate ground handling for the aircraft types and if a suitable alternative airport is found in the region necessary by law in case of operational disruptions. Besides that, the profitability of a route is strongly dependent on the aircraft type used. Older aircraft generally use more fuel and require more maintenance, which makes them more expensive in operation. Thereby is the aircraft size (number of seats) crucial, since an airline needs to be able to sell all/the majority of the seats for an acceptable price in order to be profitable. Range and seat capacity correlates positively, which implies that the further away the destination, the more passengers an airline needs to attract.

Routes are constantly opened and closed, because new destinations fail all the time. Often destinations will be closed down after a certain period, sometimes because the political or competitive landscape changed, or sometimes when the demand turned out to be estimated incorrectly. For example, the first case applies to KLM's flight to Teheran, where demand drastically reduced as a result of the reintroduction of economic sanctions. Trail and error is accepted to some degree as this is often the only way to really test the demand/hypotheses/business case in the real world. However adding a destination is really costly in terms of resources (time, money, opportunity cost, effects of schedule changes), which means an airline only considers to try a new destination with a strong business case.

True Air passenger demand is unobserved, while the value of the historic booking data is declining. This is the case because ticket sales is becoming more decentralized such that the traditional central parties have less information on historic bookings than before (more information on this is provided in Section **??**). We are in need of a reliable method for air passenger demand estimation that is capable of estimating demand for routes that are currently not served directly. With this method we are capable of creating a more efficient network for capacity constrained airports and airlines.

Appendix V: Meta Search Channels

List here the Meta Search channels with its abbreviation, full name and share of searches. To calculate the share of searches, data is taken from dates 20-06-2017 to 20-06-2018.

Table 10: Meta Search Channels

	Abbreviation	Meta Search channel	Share
1	SC	Skyscanner	38%
2	KY	Kayak	16%
3	KM	Kayak mobile	12%
4	MM	Momondo	10%
5	AS	Avia Sales	4%
6	ML	Momondo mobile	4%
7	JC	Jetcost	2%
8	DH	DoHop	2%
9	LL	Liligo	2%
10	CH	Cheapflights	1%
11	CL	Cheapflights mobile	1%
12	SB	Skyscanner for business class	1%
13	FI	Finn	1%
14	TA	Trip Advisor	1%
15	SU	Swoodoo	1%
16	QU	Qunar	1%
17	LM	Liligo	1%
18	SW	Swoodoo mobile	< 1%
19	CF	Checkfelix	< 1%
20	WE	Wego	< 1%
21	HM	Hipmunk	< 1%
22	CM	Checkfelix mobile	< 1%
23	VU	Vuelos Baratos	< 1%
24	KO	Kelkoo	< 1%
25	TR	Trabber	< 1%
26	CY	Chase.nl	< 1%
27	BI	Biletyplus	< 1%
28	IX	Ixigo	< 1%
29	FA	Fare compare	< 1%
30	MD	Mundi	< 1%
31	JR	Jetradar	< 1%
32	EL	EuropeLowcost	< 1%
33	KI	Kite	< 1%
34	FR	Viviro	< 1%
35	PL	Peroley.com	< 1%

Appendix VI: Meta Search raw JSON example

Figure 19: Meta Search raw JSON example


```
{
  "_index":"logstash-metasearch-2019.01.06-000064",
  "_type":"stml",
  "_id":"6182f08556dd-138b-9e11-8d21-03a0b927",
  "_score":1.0,
  "_source":{
    "ApplicationId":"METASEARCH",
    "Scope":"SOAP",
    "Activity":"OTA_AirLowFareSearchRQ",
    "SessionId":"0",
    "bamInfo":{
      "responsetime":0,
      "srv":"JYY",
      "msp":"SC",
      "msp_country":"DE",
      "lang_rq":"de",
      "currency_rq":"EUR",
      "owrt":2,
      "outbound_org":"FRA",
      "outbound_dest":"PMI",
      "outbound_depdate":"2019-02-07",
      "outbound_datespan":1,
      "cabintype":"Economy",
      "maxsegments":"3",
      "inbound_org":"PMI",
      "inbound_dest":"FRA",
      "inbound_depdate":"2019-02-17",
      "inbound_datespan":1,
      "separateCalls":"False",
      "useLocalization":"False",
      "useCPD":"True",
      "nr_adt":2,
      "nr_chd":0,
      "nr_inf":0,
      "responsetype":"Warning",
      "warningcode":"4095",
      "warningtext":"No Priced Itineraries for Org=FRA and Dest=PMI"
    },
    "beaconId":"insightsH",
    "beaconVersion":"restV1",
    "beaconTime":"2019-01-08T00:00:45.651Z",
    "ActivityTime":"2019-01-08T00:00:45.651Z",
    "ActivityLatency":0,
    "UUID":"6182f08556dd-138b-9e11-8d21-03a0b927",
    "headers":{
      "X-Forwarded-Origin":"internal",
      "X-Forwarded-For":"171.21.247.158",
      "X-Forwarded-Host":"www.klm.com",
      "X-Forwarded-Proto":"https",
      "X-Cipher-Version":"TLSv1",
      "X-Cipher-Name":"ECDHE-RSA-AES256-CBC-SHA",
      "Client-Ip":"171.21.247.158"
    },
    "firstPublicIP":"171.21.247.158",
    "bamInfoSize":542,
    "geoip_city":"Amstelveen",
    "geoip_country":"NL",
    "geoip_name":"Netherlands",
    "geoip_point":"52.3097,4.8773",
    "host_param":"fluenth1",
    "stml_env":"production",
    "fluentgateway_host":"kl147dwx",
    "tag":"METASEARCH.baminsights.beacon"
  }
```

Appendix VII: Additional scatter plots for MIDT vs explanatory variables

Figure 20: Scatter MIDT vs catchment area

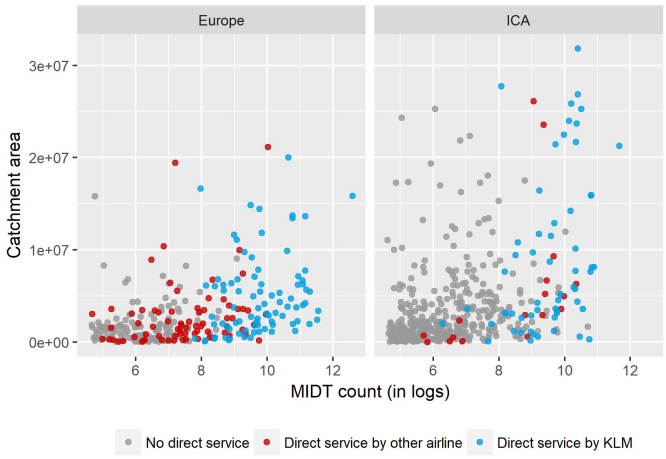
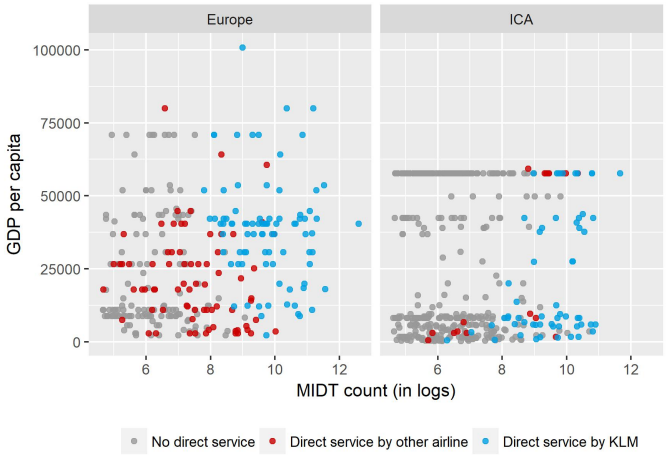


Figure 21: Scatter MIDT vs GDP per capita



Appendix VIII: Insight on destination level from the AMS raw data set

The data is arranged on Meta Search counts in the 2018 Summer season. Table 11 shows the top destinations from Amsterdam for the 2018 Summer season (blue color in Figure 11). Table 12 shows the top destinations from Amsterdam fot the

Table 11: Top destinations from Amsterdam (AMS) for summer 2018

#	Code	City	Country	Continent	Meta Search	MIDT
Due to confidentiality of the data, the content has been removed from the public version						

2018 Summer season that are not operated by KLM, but are by either their partners or competitors (red in Figure 11). Table 13 shows the top destinations from Am-

Table 12: Top destinations from Amsterdam (AMS) for summer 2018 not operated by KLM

#	Code	City	Country	Continent	Meta Search	MIDT
Due to confidentiality of the data, the content has been removed from the public version						

sterdam for the 2018 Summer seasons where currently no direct service is offered by any airline (gray color in Figure 11). These destinations can only be reached via a transfer.

Table 13: Top destinations from Amsterdam (AMS) for summer 2018 no direct service by any airline (multileg destinations removed)

#	Code	City	Country	Continent	Meta Search	MIDT
Due to confidentiality of the data, the content has been removed from the public version						

Appendix IX: KLM destinations with lowest predicted demand

Table 14: KLM destinations with lowest predicted demand

#	Code	City	Country
Due to confidentiality of the data, the content has been removed from the public version			