Revealing Route Bias in Air Transport Data: The Case of the Bureau of Transport Statistics (BTS), Origin-Destination Survey (DB1B)

Filipe Marques Teixeira *, Ben Derudder
Geography Department, Ghent University, Krijgslaan 281/S8, Ghent, B9000, Belgium

1. Introduction

Air transport research in its various guises analyzes the structure of air transport connections, both on their own terms and in their role as a catalyst of wider economic and social developments at various scales (Button & Yuan, 2013; Lin, 2014; O'Connor & Fuellhart, 2012; Taaffe, 1956). The increasing relevance of the research field at large has been fueled by the observation that both the size and the impact of air transport has been increasing over the past decades (Air Transport Action Group, 2008; Ishutkina & Hansman, 2008). For example, passenger numbers have been soaring, more than tripling from 1.025 billion in 1990 to 3.227 billion passengers in 2017 worldwide[1], with IATA forecasting these numbers to double again by 2036[2]. Against this backdrop, researchers have sought to better understand the impact and evolution of air transport, including new and better ways of modeling air transport networks and their many effects.

One of the most critical challenges in air transport research is the uneven availability and formatting of data (e.g. Derudder and Witlox, 2005). While there has been a growth and diversification in data collection strategies over the past few years (Poorthuis & Zook, 2017), air transport researchers still face a limited choice of data sources that may or may not provide information in the desired format and/or detail. When zooming in on researching air transport networks in and of themselves, there are two options. The first option is to build a database from scratch, which in this day and age most often entails collecting web-based air travel data. This can either be done by (1) using or developing a web Application Programming Interface (API) to access a meta-search engine or online route planner and/or by (2) 'screen-scraping' online route planners or meta-search engines (e.g. Grubesic and Zook, 2007). The second option is to use primary datasets and the tools provided to access them. This varies from freely available raw data to sometimes quite expensive databases with bespoke analytical and visualization tools as well as structured APIs (e.g. Google Flights and the Official Airline Guide (OAG) data).

In this paper, our focus will be on a publicly accessible and arguably one of the most widely used primary air transport datasets: the data provided by the United States Bureau of Transport Statistics (BTS), and its Origin Destination Survey (DB1B) in particular. Analyses based in the DB1B datasets are geographically circumscribed in that the data are restricted to information on domestic flights and international flights departing from/arriving at United States airports. In addition, the DB1B dataset is a 10% sample of reported tickets rather than a full dataset. In spite of this focus on a sample of US-centered flights, the importance of the DB1B dataset in air transport research cannot be underestimated (Mao, Wu, Huang, & Tatem, 2015; Neal, 2010, 2014; Seshadri, Baik, & Trani, 2007). There are two reasons for this. First, the data 'feeds' parts of some of the other well-known datasets. For example, in the case of the OAG, the DB1B data is used by drawing on the following simple method: "DB1B is a 10% sample of an airline's tickets, then 'adjusted' to estimate 100% of the market by multiplying the data by a factor of 10" (OAG, 2015). Second, the detailed info and the consistent way in which data are gathered make the BTS datasets ideal for air transport research. For example, Neal (2010) uses BTS data to provide an overview of the use of air traffic networks for urban research by building models of urban-economic flows. Later, he extended this work by differentiating air traffic

[1] https://data.worldbank.org/indicator/IS.AIR.PSGR?end=2017&start=1990
[2] https://www.iata.org/pressroom/pr/Pages/2017-10-24-01.aspx

networks by scale, species and season (Neal, 2014). Fuellhart (2013) and Brueckner (2014) use BTS data to analyze multi-airport regions (MARs) in the US, disentangling the geographies of offer and demand in these MARs. Brownstein et al. (2006) and Colizza et al. (2007) use BTS data to study epidemiological networks as these are increasingly undergirded by air transport movements.

Irrespective of the diversity of the topics addressed in these and in many other papers, it is clear that one of the drawing cards of the BTS sample as well as other major primary datasets is their alleged 'ground truth'. However, few of the data providers offer detailed info about how their data are sampled and treated. The data quality is therefore sometimes taken to be self-evident. The OAG (2019), for example, self-advertises as "the world's most comprehensive and accurate real-time travel data" without disclosing further details on data collection, treatment and validation. Nonetheless, to date there has been little scrutiny of the alleged quality of these datasets (Boyd & Crawford, 2012; Poorthuis & Zook, 2017). Furthermore, the few data quality reports that have been provided tend to be published by the data providers themselves (Strohmeier, Martinovic, Fuchs, Schäfer, & Lenders, 2015; Zaveri et al., 2012), with a noteworthy example being a BTS 2005 report on aviation data modernization (US Department of Transportation & Office of the Secretary, 2005). The report determined that 69% of city pairs reported by the DB1B did not meet the Department's accuracy criteria when using enplanement statistics as a validation benchmark. Although this does not necessarily imply that there is a data integrity or structural bias problem with the DB1B data, it does raise questions about how accurate the data are and what the nature of possible biases might be. Because the implications of using inaccurate data may be profound, it is of key importance for air transport researchers to map and understand such possible biases in these datasets.

The purpose of this paper is to explore how potential biases in air transport datasets can be detailed. To this end, we develop a methodology that allows identifying possibly biased routes in datasets in an automated manner. Although our focus will be on validating the DB1B data, we present our methodology as a more generic approach that can be used in different contexts and applied to different datasets. To this end, the remainder of this paper is organized as follows. We start by outlining four important factors to consider when working with both the DB1B database and the Air Carrier Statistics (T-100) database (which we use to validate the DB1B database): collection, representation, inter-mutability and nomenclature. We then use descriptive statistics to describe these DB1B and T-100 databases, and propose a Jaccard-like index to identify biased routes. We conclude by demonstrating how route/database biases can impact research by means of a number of straightforward case studies, focusing on our understanding of the position of routes/airports in air transport networks. In a concluding section, we explain how this approach can be adapted to serve as a more generic tool for assessing route bias in air transport datasets.

2. The Bureau of Transport Statistics (BTS) datasets

2.1. The Origin Destination Survey (DB1B)

The DB1B survey is conducted continuously by all certified US carriers involved in domestic passenger operations. The database covers a "two-tiered" stratified 10% sample, following the 14 CFR part 241 guidelines from the Department of Transport. Data coverage started in 1993, has been updated since, and groups data on a quarterly basis. One of the interesting aspects of this dataset is that it purports to show 'real' passenger movements rather than separate flights. In addition, as the DB1B database displays ticket information, it is possible to look at transfers, fare paid, booking class, intermediate stops and other passenger-related data (Goetz & Vowles, 2000; Vowles, 2001).

The DB1B database comprises three sets of info related with a single entry: Coupon, Ticket and Market (Figure 1). For example, consider the case of a passenger booking a return *Ticket* from Boston to Atlanta. The first leg of this trip involves a transfer at JFK, the second leg of the trip does not involve a transfer. In this case, BOS-ATL and ATL-BOS would be considered to be the *Markets* as the Boston-based passenger effectively travels to Atlanta. Meanwhile, the BOS–JFK, JFK–ATL and ATL–BOS segments would be considered to be three individual *Coupons*. This concept derives from the classic coupon/ticket concept, where tickets were printed rather than available on a digital medium.

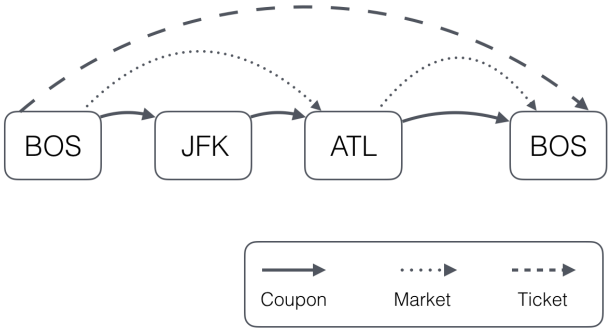

Figure 1 – The concept of coupon, market and ticket in the DB1B database.

According to the Department of Transportation (DOT), the data to be recorded centers on *"lifted ticket flight coupons"* (i.e. tickets issued by a travel agent, including online ticketing), and includes the following variables: "*Point of origin, carrier on each flight-coupon stage, fare-basis code for each flight-coupon stage, points of stopover or connection (interline and intraline), point of destination, number of passengers, and total dollar value of ticket (fare plus tax)*" (US Department of Transportation, 2012). As mentioned above, the DB1B dataset is reported as a 10% sample. The randomization is done by carriers by selecting tickets with a serial number ending in zero so that the data collection procedure effectively involves a two-tiered stratified sample.

However, the randomization methods used to build this sample are questioned by the DOT in its report (US Department of Transportation & Office of the Secretary, 2005). For example, its mentioned that *"(s)ince ticket numbers are now assigned by a computer program, the possibility that ticket numbers are assigned for reasons other than randomness arises (…) (A) tour operator might use its block of ticket numbers to issue all the ticket numbers that end in the same digit to members of a particular tour, resulting in all those tickets being selected for the sample or excluded from the sample depending on which tour was assigned ticket numbers ending in zero"*. While the deviation from the 10% seemed to be small (i.e. in the 9-11% range) the same report does raise concerns regarding the representation and reliability of information of smaller markets in particular.

Another potential issue associated with the DB1B database can be found in its definition of an Origin Destination (OD) flight. The common definition of an OD flight refers to a passenger traveling from the origin of the trip to the final destination of the trip, including all the intermediate flight stages. However, since the very beginning of the OD Destination Survey, the DOT has used a methodology called Directional Passenger Construction, which uses continuous direction of travel as its definition of 'true' OD. According to this methodology, a passenger is considered to be on a trip as long as the passenger continues in the same direction (i.e. North – South and East – West constitute different direction pairs). For example, on a trip from Albuquerque to Las Vegas with a stopover in Denver, the DOT would break the trip up into two individual trips due to the geographical position of Las Vegas in relation to

Albuquerque (US Department of Transportation & Office of the Secretary, 2005). Given the abundance of hub-and-spoke configurations in the organization of many air transport carriers' networks (Campbell & O'Kelly, 2012; O'Kelly, 2016; O'Kelly & Miller, 1994; Park & O'Kelly, 2016), this definition may potentially render more commonsensical notions of a trip, transfer and OD flight inaccurate. A final challenge associated with this database is associated with the concept of passengers versus passenger trips. As reported by the DOT (US Department of Transportation & Office of the Secretary, 2005), passenger counts are taken to represent passengers scheduled to fly in that quarter. However, the DB1B bundles all travel on a ticketed itinerary in a single quarter (i.e. if a passenger leaves in December and returns in January, the ticket will be reported as if it took place in December and no passenger will be reported for the first quarter of the year). Collectively, these issues may affect how well the sample of observed tickets represent (our conceptions of) the characteristics of actual air travel, even though there are no data regarding how frequent or profound any of these effects are.

2.2. The Air Carrier statistics dataset (T-100)

In principle, directly validating sampled data relies on having a full dataset. This full dataset is not available, which implies that we have to resort to another dataset: the BTS Air Carrier statistics dataset (T-100). The T-100 dataset contains domestic and US-related international airline market and segment data. Certificated US air carriers have to report air carrier traffic information on a monthly basis, using the so-called Form T-100 to the Office of Airline Information, Bureau of Transportation Statistics (Bureau of Transport Statistics, 2019).

The T-100 differs from the DB1B in several respects, starting from its time structure. In contrast to the DB1B, the T-100 database is grouped per month, while it represents a full sample and reports flights rather than individual or group tickets. Although it is less fine-grained than the DB1B dataset on a number of fronts, T-100 data is also often used in air transport research. By including all *boarded* passengers, it can help answering questions related with airport and carrier competition (Button & Lall, 1999; Dobruszkes & Van Hamme, 2011; O'Kelly, 2016; Song & Yeo, 2017), and it has also been used for better understanding route level variations within multi-airport regions (Fuellhart, 2007; Fuellhart et al., 2013).

The differences between the DB1B and T-100 datasets imply that using the latter full dataset to validate the former sampled dataset is not a straightforward exercise, a situation that is further complicated by the subtle differences in the terminology used to identify a Segment and a Market pair. The T-100 database identifies a Segment as a non-stop flight, including diversions, flag stops, tech-stops, emergency landings, etc. This data is not flight number driven and referred to as "transported data". However, Market data are flight number driven, which implies that if the flight number changes the market stops. Market data are often in research referred to as "enplanement data". Figure 2, using fictional flight numbers, shows how a flight can be both part of the Market data and the Segment data. Flight, UA01 (BOS – ATL with a stop in JFK) and UA02 (ATL – BOS), both represent a market, as the flight number stays the same. However, it is important to note that someone could take flight UA01 for the first segment only (BOS – JFK), which would not appear in the T-100 Market dataset as the flight retains the same number. Flight UA02 (ATL– BOS), flight UA03 (BOS – JFK) and flight UA04 (JFK – ATL) all represent individual segments as they show take-off and landing. While this could pose a challenge in counting the number of passengers, by merging both segment and market databases routes, we can have insight into the available routes in the US.

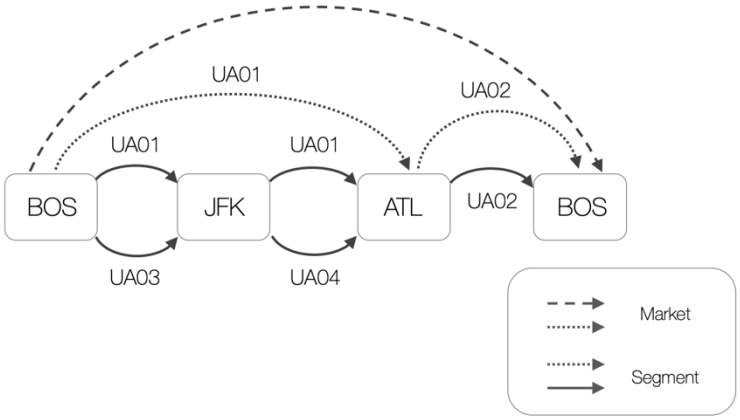

Figure 2 - T-100 concept of Market and Segment (flight numbers are fictional).

Taken together, while the data collection is broadly similar to the one used for DB1B, there are clearly also some relevant differences regarding the collection and segmentation of the data. For example, the T-100 database, rather than being a 10% sample, includes all available data. Other major differences relate to what is reflected in the data: whereas the DB1B reflects tickets that were effectively bought (i.e. the data is collected before a flight occurred), the T-100 reflects the exact number of passengers who boarded a flight (i.e. the data is collected after a flight occurred). However, as individual information is not reported in the T-100, it is difficult to infer information regarding transfers or ticket fares.

2.3. Comparing the DB1B and T-100 datasets

It is clear that, in part due to the differences in data collection procedure, the structure of the data, and the terminology the DB1B and T-100 databases are difficult to *directly* compare. Or, when cast in the context of this paper: the less detailed but full T-100 data cannot be directly used to validate the more detailed but sampled DB1B data. Fortunately, the key challenge in establishing connections between the two datasets is restricted to comparing *passenger volumes* on the same route. Although the terminology and data structure does not simply translate between these two datasets, it is possible to develop a method measuring the presence of a route throughout time in the DB1B dataset and comparing it with its presence in the T-100 dataset. However, this requires establishing a terminological comparison of the Segment/Market concepts in the datasets.

Figure 3 shows a fictional passenger traveling between BOS and LAX. This passenger took two flights (i.e. UA01 and UA06), and transferred in ATL. Meanwhile, the first flight (i.e. UA01) had a stop in JFK, albeit that the passenger did not disembark the plane there. In the DB1B dataset, this passenger – if part of the sample – would be represented by two segments (i.e. BOS-ATL and ATL-LAX) and one market (i.e. BOS-LAX). However, in the T-100 dataset, the same passenger would be represented by three segments (i.e. BOS-JFK, JFK-ATL and ATL-LAX), and two markets (i.e. BOS-ATL and ATL-LAX). It would be tempting but incorrect to infer that the DB1B Segment and the T-100 Market are simply interchangeable, as for example the T-100 Market does not consider passengers boarded in JFK as they are not part of a Market. Given this, we can infer that the only pattern that can be *directly* compared between both datasets are the routes represented rather than the number of passengers. At the same time, it is important to understand that such a comparison is only possible when both the T-100 Market and Segment data are combined. This comparison relies on the principle that the T-100 Segment shows point-to-point routes, whereas the T-100 Market shows market segments. By merging the two

datasets, we capture all available routing options, and can compare these with information derived from the DB1B sample.

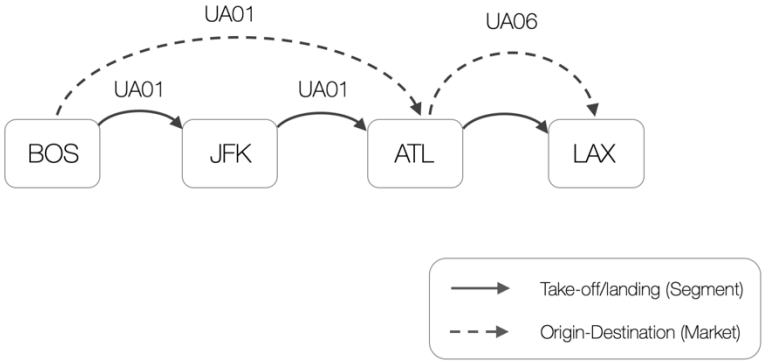

Figure 3 - Segment and market concept in the DB1B and T-100 datasets

3. Testing the randomness of the DB1B 10% sample

To compare both datasets, we work with 11-year samples between 2005 and 2015 on a quarterly basis. The data are processed into networks using the R package SKYNET (Teixeira & Derudder, 2018). Both networks were transformed so that they are comparable in terms of their nodes and edges: we merged data as to produce matrices with undirected edges featuring the exact same sets of origins and destinations, carriers, and timeframes. Because the T-100 generated network includes the full sample of airports and routes, we intersected both networks to generate a T-100 adapted sample by retaining the largest connected component of the network. Due to the 10% sample characteristics of the DB1B, we multiplied the number of passengers by 10 to have a number that is comparable with the T-100. As routes represented in the DB1B database will in theory amount to 10% of the T-100 values, T-100 routes with less than 10 passengers per quarter are not likely enough to appear in the DB1B to make a meaningful comparison, and these were therefore excluded from the analysis.

3.1. Descriptive Statistics

Figure 4 shows the number of edges (routes and passenger volumes) and nodes (airports) over time. The picture emerging here suggests a broadly consistent pattern for both databases, with the number of airports slightly decreasing and the number of passengers slightly increasing over time, except for the number of routes which remain stable in the T-100 database and decrease in the DB1B database. As can be expected, these longer-term trends are interspersed with cyclical year-long patterns with the first and fourth quarters having lower number of passengers (Bureau of Transport Statistics, 2017). Nonetheless, the number of routes is considerably higher and the number of passengers considerably lower in the DB1B dataset than in the T-100.

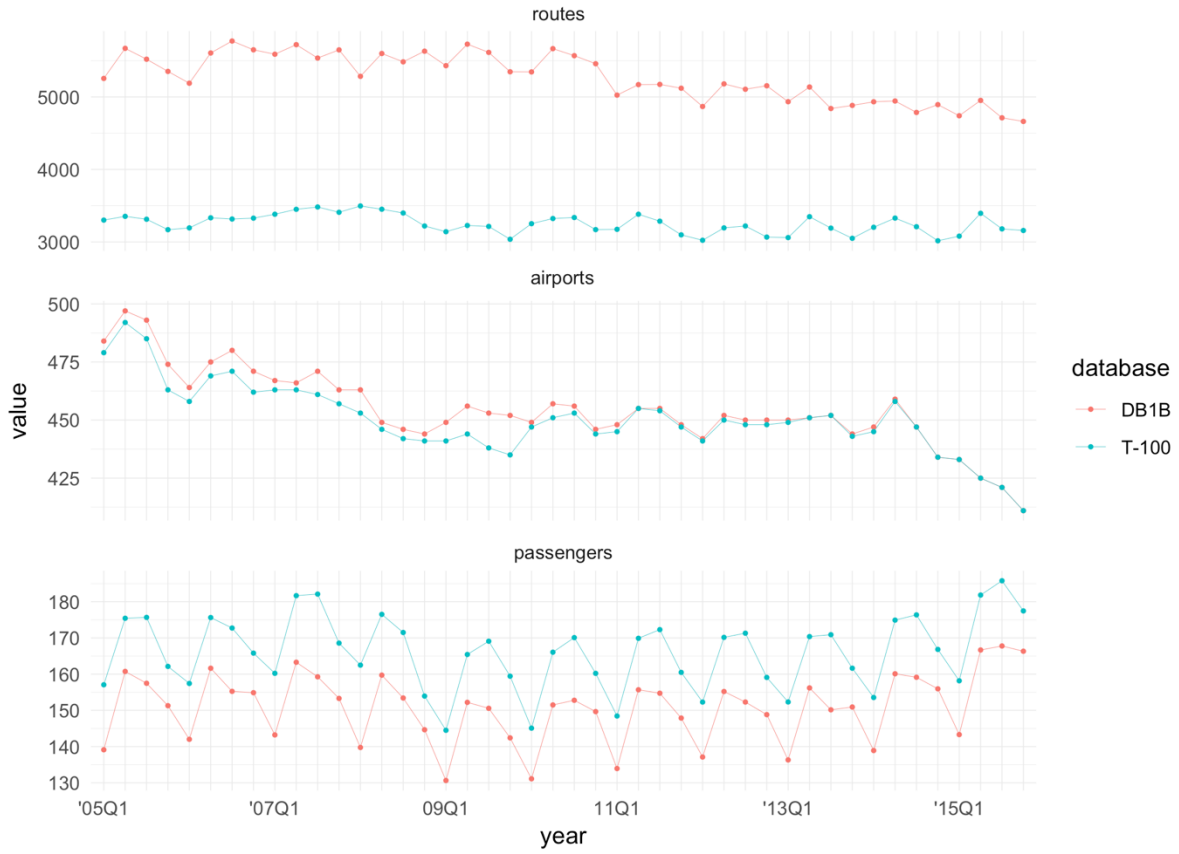

Figure 4 – Quarterly number of routes, airports and passengers (thousands) between 2005-Q1 and 2015-Q4 in the DB1B- and T-100 datasets. Note that for the airports graph, data between 2014 Q1 and 2015 Q4 is similar for the DB1B and T-100 databases.

After testing for the presence of normal distributions by running a Shapiro-Wilks test, we calculated Pearson's correlation coefficients between the same variable in both databases (e.g. number of airports per quarter in the DB1B against the same value for the same period in the T-100). We observe a strong correlation at the level of the airports ($r = 0.9695$, $p < 0.001$, Figure 4b) and passengers ($r = 0.950$, $p < 0.001$, Figure 4a). However, this does not hold at the route level ($r = 0.547$, $p < 0.01$, Figure 4c).

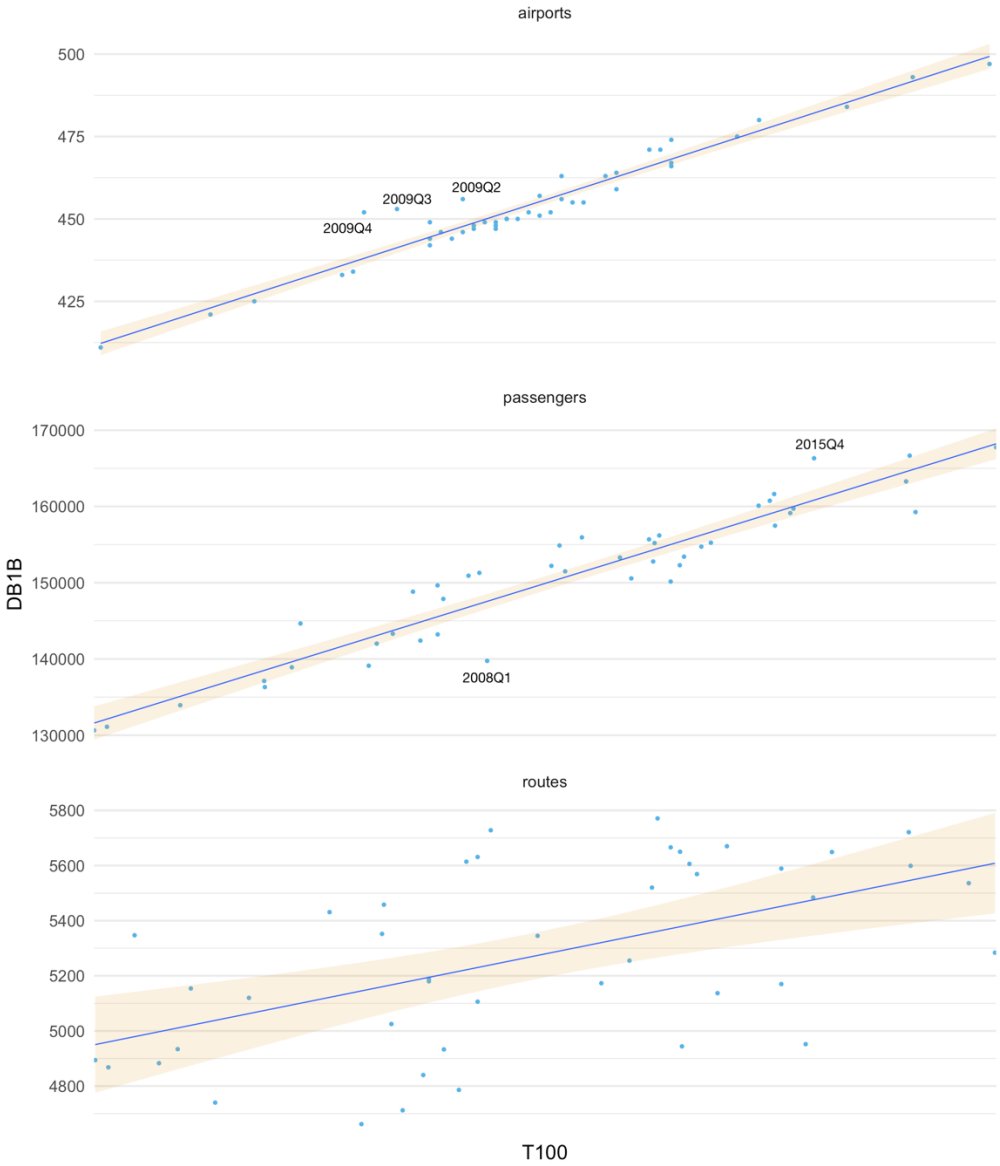

Figure 5 – Correlation analysis of airports, passengers (thousands) and routes per quarter for T-100 and DB1B (CI = 95%).

Taken together, it is clear that while the number of passengers and airports are roughly consistent in the DB1B and T-100 datasets, this is not the case for the number of routes. In addition to the low correlation, the number of routes in the DB1B dataset is consistently higher, which is implausible for a random sample (i.e. the number of routes remains consistent in the T-100 while it decreases with about 10% in the DB1B during the period under analysis). Further analysis shows that when intersecting the two databases at the route level, we see that an average of 38% (min 33%-max 43%) of the routes found in the T-100 database cannot be found in the DB1B database in the corresponding period (Figure 6). This pattern becomes even more compelling when selecting only routes with a number of passengers higher than the 75% quintile (which can be hypothesized to be more consistent and/or less prone to random effects) because even in that case an average of 15% (min 9% - max 22%) of the DB1B routes are missing from the T-100 database (Figure 7).

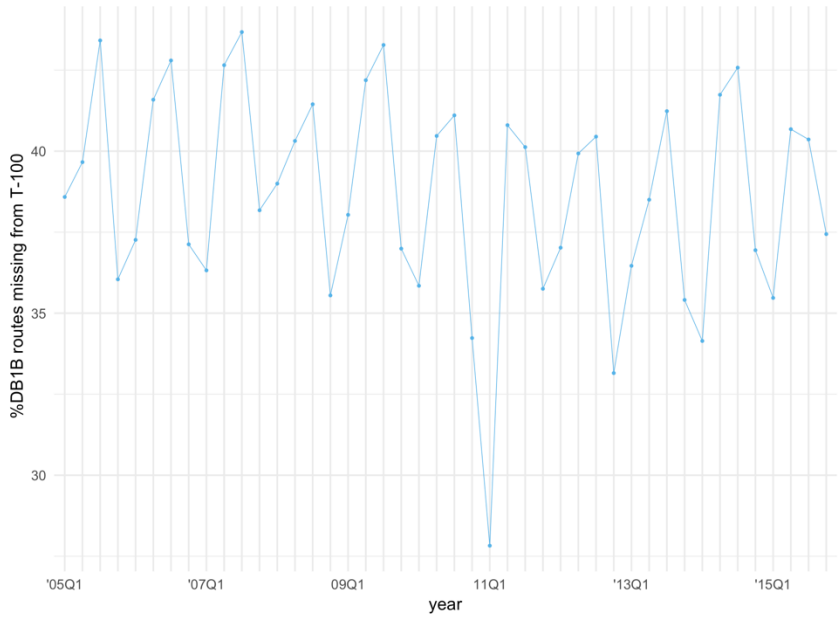

Figure 6 – Quarterly percentage of DB1B routes not present in the intersection between the DB1B and the T-100 Segment.

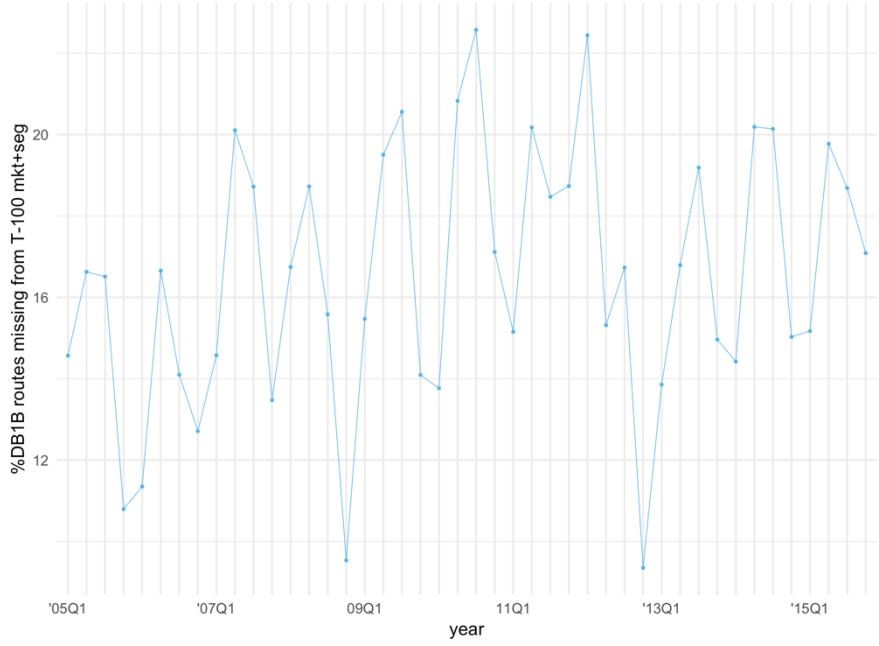

Figure 7 – Quarterly percentage of DB1B routes no present in the intersection between DB1B and T-100 (combined Market and Segment) for the passenger quintile above 75%.

4. Conceptualising a Jaccard-Like index – The Route Equality Ratio

The above exploratory assessment of the data revealed that the main differences between the T-100 and DB1B databases are route related. As the combination of the T-100 Segment with the T-100 Market database should in theory display all available routes, it can be inferred that any route exclusive to the DB1B should be considered to be inaccurate. With that in mind, we can assess and compare the presence of routes in both datasets over time in more detail. To this end, we develop a Jaccard-like index to identify how (un)evenly routes are covered in both datasets. In other words, our aim is here to capture and subsequently compare the presence of a route in the DB1B database with its presence

during the same period in the T-100 database. To this end, we develop the Equality Ratio (*EQR*) associating route frequency (*RF*) between any pair of airports in both datasets.

The calculation of the frequency of route presence *RF* consists of two complementary parts. The first parts captures yearly frequency per quarter (i.e. to capture route seasonality), while the second part captures the quarterly presence per year (i.e. to capture dispersion over years). In Equation 1, we define Ny_{ij} as equal to 1 and Nq_{ij} equal to 0.25 when a route is present for any given pair *ij* of origin destination airports in both conceptions, respectively. As most routes tend to have a seasonal dimension (Burghouwt & de Wit, 2005; Mao et al., 2015; Rocha, 2017; Sun, Wandelt, & Linke, 2015), a heavier weight is given to the yearly presence per quarter (i.e. Ny_{ij}). The frequency of route presence between airports *i* and *j* (RF_{ij}) can then be calculated as follows:

$$RF_{ij} = \frac{\sum_{q \in Q}(\sum_{y \in Y} Ny_{ij})}{|Y|} + \frac{\sum_{y \in Y}(\sum_{q \in Q} Nq_{ij})}{|Y|}$$

With:

- RF_{ij} = Route frequency for routes between airports i and j
- Y = Range of years
- Q = Range of Quarters
- $Ny_{ij} = \begin{cases} 1, & if\ the\ route\ exists \\ 0, & if\ the\ route\ does\ not\ exist \end{cases}$
- $Nq_{ij} = \begin{cases} 0.25, & if\ the\ route\ exists \\ 0, & if\ the\ route\ does\ not\ exist \end{cases}$

Equation 1 – Route frequency

RF_{ij} ranges from 0 (never present in a dataset) to 5 (consistently present in a dataset), and is used to calculate the Equality Ratio, which is given by:

$$\mathrm{EQR_{ij}} = \begin{cases} \dfrac{RF(DB1B)_{ij}}{RF(T-100)_{ij}} & for\ RF(DB1B)_{ij} < RF(T-100)_{ij} \\ -\dfrac{RF(T-100)_{ij}}{RF(DB1B)_{ij}} & for\ RF(DB1B)_{ij} > RF(T-100)_{ij} \end{cases}$$

With:

- $\mathrm{EQR_{ij}}$ = Equality Ratio for routes between airports *i and j*.

Equation 2 - Equality ratio

$\mathrm{EQR_{ij}}$ ranges from -1 to +1, with a value of 1 representing a perfect match (i.e. consistent route presence/absence in both datasets) and negative values indicating a higher frequency of route presence in the DB1B than in the T-100. If the DB1B dataset would be an unbiased sample, we would expect high positive values converging on a value of +1.

Table 1 shows some examples of routes and their associated *RF* and *EQR* values. If we take the example of ABE-CLE, which has an *RF* equal to 2.39 in the DB1B database, we can infer that the route tends to be fairly but not consistently present. In the T-100 database, this route has a *RF* equal to 3.5, which

implies that the route has been almost always present across years and seasons. For the ABE-CLE example, this produces an *EQR* equal to 0.68, which captures that the frequency for that route is not the same for the two databases. Some of the other routes (e.g. ABE-ATL) are consistently present in both datasets and have the expected value of *EQR* of 1, but it can be seen that for quite a large number of routes the *EQR* is indeed not equal to 1. A systematic appraisal of EQR across airports allows assessing how well both databases match (or don't).

Origin	Destination	DB1B frequency	T-100 frequency	Equality ratio
ABE	ATL	5	5	1
ABE	AVP	2.72	1.70	-0.62
ABE	BOS	0.1	1.1	0.1
ABE	CLE	2.39	3.5	0.68
ABE	CLT	5	5	1
ABE	CVG	1.7	1.7	1

Table 1 – Route frequency in DB1B and T-100 and the associated EQR, for first 6 alphabetically ordered routes by origin and destination.

In Figure 8, which shows the distribution of *EQR* values for all sampled routes between 2005 and 2015, we can observe that over 40% of the routes (shown in blue), have an *EQR* lower than 0.85. Even though an *EQR* of 1 is by far the single largest value, which suggests that there is indeed a fair share of routes that is consistently covered in both datasets covering 45% of all routes, there is no convergence on 1. Furthermore, there is a substantial number of routes with an *EQR* < 0. Overall, then, we find that routes are either consistently covered (the grey bar to the right in the *EQR* range) or exhibit different and seemingly almost random presences in both databases (the distribution in the remainder of the *EQR* range). As can be expected, this is slightly less the case for the busiest routes: routes with *RF* values of 5 – think: JFK-LAX or ORD-SFO – do more often result in EQR values of 1.

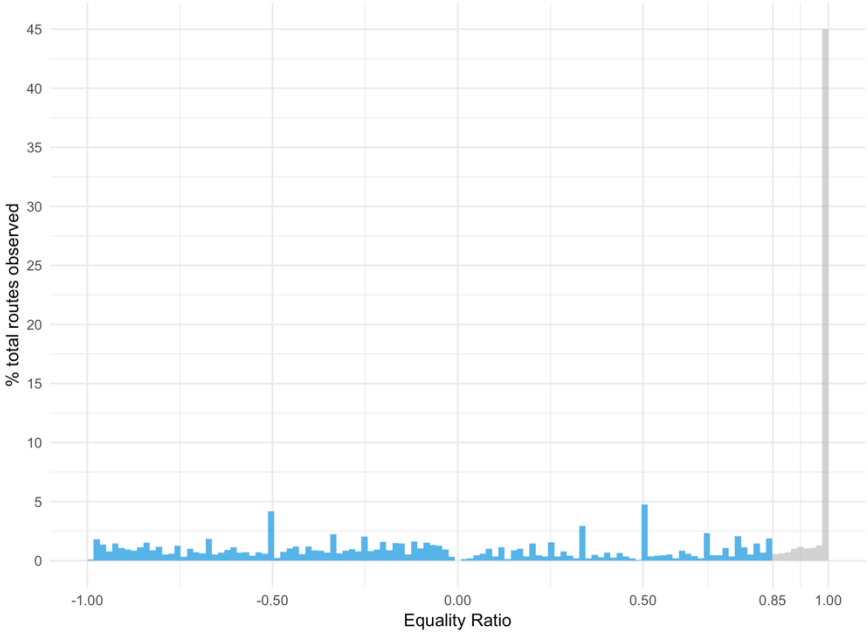

Figure 8 – Histogram of percentage of routes observed by EQR for all sampled routes between 2005 and 2015.

5. Assessing the impact of using biased data

Although almost half of the routes in the DB1B sample are indeed coherently present in time and space, this also implies that more than half of the routes in the DB1B sample do *not* coherently reflect actual routing in space and time. In addition, there is no convergence in *EQR* on 1, which suggests that some routes in the DB1B dataset have no real-world meaning. To describe the potential impact of this issue, we explore what it this may mean in practice for two types of air transport research: airport focused research and route focused research.

5.1. The potential impact on airport focused research

While it can be expected that some of the lower passenger volume routes may not be present in the DB1B due to its 10% sample characteristics, below we will demonstrate that even some of the busiest routes are not always represented in both DB1B and T100 databases. Table 2 shows the mean EQR for seven airports of varying importance alongside other information that is relevant to reflect on the potential impact of the 'ghost routes', i.e. routes present in the DB1B but not in the T-100 database. The table shows, for example, that an airport such as LAX with more than 25 million passenger departures per year in 2015 had almost 50% of its unique routes in the DB1B dataset not regularly matching the routes present in the T-100 database.

Airport	Passenger (departures for 2015 in thousands)	Equality ratio (mean)	Equality ratio (standard deviation)	% routes with EQR = 1	Total number of unique routes (incoming and outgoing)
Hartsfield-Jackson Atlanta International (ATL)	44319	0.801	0.510	72.48	447
Chicago O'Hare International (ORD)	30661	0.806	0.491	74.29	389
Dallas/Fort Worth International (DFW)	27914	0.761	0.563	71.39	395
Los Angeles International (LAX)	26854	0.514	0.699	48.61	323
Denver International (DEN)	25521	0.684	0.628	64.07	437
Phoenix Sky Harbor International (PHX)	20773	0.623	0.611	55.05	287
Charlotte Douglas International (CLT)	20719	0.805	0.509	74.23	291

Table 2 – EQR for 7 busiest airports by volume of departed passengers (domestic flights only) in the US (2005-2015)

There are several papers in the literature dealing with the impact and relevance of specific routes for airports (Goetz, 1993; Goetz & Vowles, 2000; Wei & Grubesic, 2016). Analyzing an airport that has a substantial number of 'ghost routes' entails that the research question is tackled with skewed data and may subsequently affect the researcher's final conclusions. Research on Multi Airport Regions (MARs) is one research area where airport-level variations in routes are tackled (Brueckner et al., 2014; Fuellhart, 2007; Fuellhart et al., 2013; Fuellhart, Ooms, Derudder, & O'Connor, 2016). Brueckner (2014), for example, provides a methodology looking at the effects of competition on airports in MARs by drawing on DB1B data. Although it is not clear which routes were used and how these are affected by the pattern described here, and although we acknowledge that Brueckner (2014) filtering the data by removing routes shorter than 200 miles and with less than 10 passengers per each way may have tackled some the issues raised here, some of the airports analyzed do display a high percentage of

routes with a low-quality EQR. Meanwhile, Fuellhart (2013) looks at route-level passenger variations within three multi-airport regions (i.e. Boston, San Francisco and Washington). While the data he used was retrieved from the T-100 database, it is nevertheless clear that there would be an impact if DB1B data would be used to answer similar questions. For example, we can observe that all the largest airports within these three areas have a high percentage of low-quality routes (i.e. BOS – 18,7%, SFO – 32%, IAD – 23,04%).

5.2. The potential impact on route focused research

Our analysis reveals that between 2005 and 2015 there are 782 unique routes with more than 137 passengers per day (one way) where their frequency in the DB1B database is higher than in the T-100 (the threshold was calculated by selecting routes with passenger volumes above the second quintile).

At the level of the routes, there are two important elements to consider. The first one resides in the assumption that as the DB1B is a 10% sample, it is expected that some – especially: smaller – routes may end up being missing from that database. However, we can assume that the opposite is in principle not possible as the T-100 Segment and Market should include all available non-stop or flight number dependent routes. The second element is the number of passengers shown for a particular route. While some small differences are expected due to the characteristics of the sampling, it is hard to assess the number of passengers represented on a route due to the way in which the data is collected (i.e. ticket collected vs passengers flown), but most importantly due to not being able to compare the number of passengers in the 10% sample with a full dataset. The literature shows several papers studying the impact and relevance of routes on the economy, in terms of competition, or as a way of understanding people's movements (Dresner, Lin, & Windle, 1996; Neal, 2014; Vowles, 2001). In such research routes are not aggregated per airport, but scrutinized individually which makes research even more vulnerable to the effects of routes being overrepresented in the DB1B database. With this in mind, it becomes crucial to understand if a route is over-represented (i.e. its frequency is higher in the DB1B than in the T100 database). To explore this issue, we selected three routes (i.e. BPT-IAH, OTH- SFO, MSP-PSP) based on our preliminary finding that their presence was higher in the DB1B than in the T-100 database. When looking at the BPT (Beaumont/Port Arthur, TX) – IAH (Houston, TX) route, we can see that for 2012 Q4, 2013 Q1, 2014 Q1, 2014 Q3, 2015 Q2 and 2015 Q4, there were no results in the T-100 database, while the DB1B does show results for that route. While the passenger average per month is low (i.e. approximately 170 per quarter) in 2012 Q4, the route showed 640 passengers in the DB1B. In the case of the route OTH (North Bend/Coos Bay, OR) – SFO (San Francisco, CA) we can see a similar pattern. Despite having a more constant presence, it is possible to observe in both databases that this route was being served by SkyWest Airlines (OO), but it is only in the DB1B that this same route emerged as being operated by United Airlines (UA) and Continental Airlines (CO). Although it is known that SkyWest often flies for United (Wickham, 2011), this does not explain United's absence in the T-100 database. However, it is important to notice as well that 2011 Q4 shows 7297 passengers flying the OTH – SFO segment for the T-100 and 9090 passengers for the DB1B. Assuming that these are direct flights it is difficult to understand the extra carriers and the different number of passengers shown between both databases.

The issue becomes even more compelling but complex, when analyzing our last example: the route between MSP (Minneapolis/St. Paul, MN) and PSP (Palm Springs, CA). Table 3 shows a rough comparison between the three datasets. The presence of Northwest Airlines (NW) in the DB1B database and its absence in the T-100 can easily be explained by its end of operations on the 31st of January 2010 and its subsequent merger with Delta Airlines (DL) (Luo, 2014). However, Sun Country Airlines (SY), with its main hub in MSP, displays more passengers for a segment which is similar in number of passengers for both the T-100 Segment and Market datasets. On Sun Country Airlines website (Sun Country Airlines, 2018), it can be read that they offer non-stop flights between both

airports except for the earlier-mentioned "roll-over" behavior where a passenger can see its ticket being changed to another carrier.

Operating Carriers	AA	DL	NW	SY	UA
DB1B	140	15850	430	9930	-
T-100 Segment	-	17063	-	9633	-
T-100 Market	163	16677	-	9633	291

Table 3 - passengers for route MSP - PSP, 2010 Q1

The uneven presence of routes can, for example, have an impact in research that looks at route competition in the US (Bania, Bauer, & Zlatoper, 1998; Borenstein & Rose, 2002; Morrison & Winston, 1990). While airline mergers and "roll-over" effects can impact the observed data (e.g. the already mentioned NW merger with DL), some of the overrepresented values in the DB1B database cannot be directly and easily explained. Nonetheless, tagging routes based on their EQR can be used as a means of understanding if they can be directly used in research or if further scrutiny is needed.

6. Discussion and final remarks

The primary focus of this paper has been to explore potential biases in the DB1B database, the source of some of the most commonly used datasets in air transport research. To this end, we developed a methodology that allows identifying possibly biased routes: a Jaccard-like index was proposed to compare route presence in the DB1B data against a route presence database derived from T-100 data. Importantly, this approach implies that our methodology has broader purchase in that it can be cast as a more generic approach to validate route presence in different contexts and applied to different datasets: our proposed *EQR* algorithm can – in this or an amended form – be used to identify potentially biased routes in other databases as well.

To the best of our knowledge, the DB1B database has not yet been validated with the partial exception of an earlier and equally critical BTS self-assessment. One reason for this is that validating the DB1B database is not straightforward because of a range of differences with other BTS datasets. As we have discussed in this paper, there are four important factors to consider when comparing the BTS databases: *Collection* – how are the tickets collected and sampled? *Representation* – does the data reflect the final results (i.e. does it represent a passenger boarding a flight or the ticket bought by a passenger)? *Inter-mutability* – can data be transferred between the DB1B and T-100 datasets? And *nomenclature* – do variables with the same name mean the same thing in both datasets? By coherently considering the above factors, it became possible to assess validate the DB1B database using T-100 data.

Our main findings are that (1) although roughly half of the routes in the DB1B dataset are consistently present in the T-100 dataset there are also many inconsistent routes, and (2) that although this issue is present across routes and airports this is somewhat less pronounced for important routes and airports. Our research is able to identify some of the issues with the DB1B data, but is if course not able to shed light on the *source(s)* of the issues, even though the BTS self-assessment and some of the route level bias examples discussed in the previous section offer some suggestions. That said, the purpose of this paper has not been to invalidate the DB1B database or its potential relevance: its level of detail, consistency, wide longitudinal range, and free availability alone imply that it remains a premier data source for air transport researchers. Other data sources often costs tens of thousands of dollars or are locked behind confidentiality agreements for data usage (Huang, Wu, Garcia, Fik, & Tatem, 2013), making it almost impossible to unlock these for research purposes.

Follow-up research can focus on potential solutions to the issues raised here. One preliminary suggestion is to use our method to isolate potentially biased routes and the validate these using secondary sources. For example, researchers could flag routes with an *EQR* < 0.85, and then use airlines' websites or other secondary sources to cross-check the actual presence of the route. Another possibility is to add our method to existing frameworks (e.g. TensorFlow, Keras, Torch), mostly in the domain of Machine Learning, to build a full dataset (e.g. Abadi et al., 2016; Chollet François, 2015; Collobert et al., 2016). As our method looks at route presence and potentially frequency rather than passenger volume, *EQR* scores can be used to label routes in order to be later used by machine learning algorithms to "fill in" missing values by using different imputation methods (Batista & Monard, 2003; Biessmann, Salinas, Schelter, Schmidt, & Lange, 2018; Schafer & Olsen, 1998).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2016). TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning. *Proc 12th USENIX Conference on Operating Systems Design and Implementation*.

Air Transport Action Group. (2008). *The economic and social benefits of air transport 2008*. Geneva, Switzerland.

Bania, N., Bauer, P. W., & Zlatoper, T. J. (1998). U.S. air passenger service: A taxonomy of route networks, hub locations, and competition. *Transportation Research Part E: Logistics and Transportation Review*, *34*(1), 53–74. https://doi.org/10.1016/S1366-5545(97)00037-9

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*. https://doi.org/10.1080/713827181

Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). "Deep" Learning for Missing Value Imputationin Tables with Non-Numerical Data (pp. 2017–2025). Torino, Italy: CIKM '18 Proceedings of the 27th ACM International Conference on Information and Knowledge Management. https://doi.org/10.1145/3269206.3272005

Borenstein, S., & Rose, N. L. (2002). Competition and Price Dispersion in the U.S. Airline Industry. *Journal of Political Economy*. https://doi.org/10.1086/261950

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*. https://doi.org/10.1080/1369118X.2012.678878

Brownstein, J. S., Wolfe, C. J., & Mandl, K. D. (2006). Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.0030401

Brueckner, J. K., Lee, D., & Singer, E. (2014). City-Pairs Versus Airport-Pairs: A Market-Definition Methodology for the Airline Industry. *Review of Industrial Organization*. https://doi.org/10.1007/s11151-012-9371-7

Bureau of Transport Statistics. (2017). A Time Series Analysis of Domestic Air Seat and Passenger Miles. Retrieved from https://www.bts.gov/archive/publications/transportation_indicators/october_2002/Special/A_Time_Series_Analysis_of_Domestic_Air_Seat_and_Passenger_Miles

Bureau of Transport Statistics. (2019). Bureau of Transport Statistics.

Burghouwt, G., & de Wit, J. (2005). Temporal configurations of European airline networks. *Journal of Air Transport Management*, *11*(3), 185–198. https://doi.org/10.1016/j.jairtraman.2004.08.003

Button, K., & Lall, S. (1999). The economics of being an airport hub city. *Research in Transportation Economics*, *5*(C), 75–105. https://doi.org/10.1016/S0739-8859(99)80005-5

Button, K., & Yuan, J. (2013). Airfreight Transport and Economic Development: An Examination of Causality. *Urban Studies*, *50*(2), 329–340. https://doi.org/10.1177/0042098012446999

Campbell, J. F., & O'Kelly, M. E. (2012). Twenty-Five Years of Hub Location Research. *Transportation Science*, *46*(2), 153–169. https://doi.org/10.1287/trsc.1120.0410
Chollet François. (2018). Keras: The Python Deep Learning library. *Keras.Io*. https://doi.org/10.1086/316861
Colizza, V., Barthélemy, M., Barrat, A., & Vespignani, A. (2007). Epidemic modeling in complex realities. *Comptes Rendus - Biologies*. https://doi.org/10.1016/j.crvi.2007.02.014
Collobert, R., Van Der Maaten, L., & Joulin, A. (2016). Torchnet: An OpenSource Platform for (Deep) Learning Research. In *Proceedings of the 33rd International Conference on Machine Learning*.
Dobruszkes, F., & Van Hamme, G. (2011). The impact of the current economic crisis on the geography of air traffic volumes: An empirical analysis. *Journal of Transport Geography*, *19*(6), 1387–1398. https://doi.org/10.1016/j.jtrangeo.2011.07.015
Dresner, M., Lin, J. S. C., & Windle, R. (1996). The impact of low-cost carriers on airport and route competition. *Journal of Transport Economics and Policy*, *30*(3), 309–328. https://doi.org/10.2307/20053709
Fuellhart, K. (2007). Airport catchment and leakage in a multi-airport region: The case of Harrisburg International. *Journal of Transport Geography*, *15*(4), 231–244. https://doi.org/10.1016/j.jtrangeo.2006.08.001
Fuellhart, K., O'Connor, K., & Woltemade, C. (2013). Route-level passenger variation within three multi-airport regions in the USA. *Journal of Transport Geography*, *31*, 171–180. https://doi.org/10.1016/j.jtrangeo.2013.06.012
Fuellhart, K., Ooms, K., Derudder, B., & O'Connor, K. (2016). Patterns of US air transport across the economic unevenness of 2003–2013. *Journal of Maps*, *12*(5), 1253–1257. https://doi.org/10.1080/17445647.2016.1152917
Goetz, A. R. (1993). Geographic patterns of air service frequencies and and pricing at US cities. *Journal of the Transportation Research Forum*, *33*, 56–72.
Goetz, A. R., & Vowles, T. M. (2000). 'Pockets of Pain' across the deregulated landscape: the geography of US airline fares and service in the, 1990s. *Paper Presented at the Annual Meeting of the Association of American Geographers, Pittsburg*.
Grubesic, T., & Zook, M. (2007). A ticket to ride: Evolving landscapes of air travel accessibility in the United States. *Journal of Transport Geography*, *15*(6), 417–430. https://doi.org/10.1016/j.jtrangeo.2006.12.002
Huang, Z., Wu, X., Garcia, A. J., Fik, T. J., & Tatem, A. J. (2013). An Open-Access Modeled Passenger Flow Matrix for the Global Air Network in 2010. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064317
Ishutkina, M., & Hansman, R. J. (2008). Analysis of Interaction between Air Transportation and Economic Activity. In *The 26th Congress of ICAS and 8th AIAA ATIO*. https://doi.org/10.2514/6.2008-8888
Lin, W. (2014). The politics of flying: Aeromobile frictions in a mobile city. *Journal of Transport Geography*, *38*, 92–99. https://doi.org/10.1016/j.jtrangeo.2014.06.002
Luo, D. (2014). The Price Effects of the Delta/Northwest Airline Merger. *Review of Industrial Organization*, *44*(1), 27–48. https://doi.org/10.1007/s11151-013-9380-1
Mao, L., Wu, X., Huang, Z., & Tatem, A. J. (2015). Modeling monthly flows of global air travel passengers: An open-access data resource. *Journal of Transport Geography*, *48*, 52–60. https://doi.org/10.1016/j.jtrangeo.2015.08.017
Morrison, S. A., & Winston, C. (1990). The dynamics of airline pricing and competition. (Deregulated airline markets). *American Economic Review*, *80*(2), 389.
Neal, Z. (2010). Refining the air traffic approach to city networks. *Urban Studies*, *47*(10), 2195–2215. https://doi.org/10.1177/0042098009357352
Neal, Z. (2014). The devil is in the details: Differences in air traffic networks by scale, species, and season. *Social Networks*, *38*(1), 63–73. https://doi.org/10.1016/j.socnet.2014.03.003
O'Connor, K., & Fuellhart, K. (2012). Cities and air services: The influence of the airline industry.

Journal of Transport Geography, *22*(April 2011), 46–52. https://doi.org/10.1016/j.jtrangeo.2011.10.007
O'Kelly, M. E. (2016). Global Airline Networks: Comparative Nodal Access Measures. *Spatial Economic Analysis*, *11*(3), 253–275. https://doi.org/10.1080/17421772.2016.1177262
O'Kelly, M. E., & Miller, H. J. (1994). The hub network design problem. A review and synthesis. *Journal of Transport Geography*, *2*(1), 31–40. https://doi.org/10.1016/0966-6923(94)90032-9
OAG. (2015). DOT Analyser User Guide. Retrieved from https://www.oag.com/dot-analyser-user-guide#1.3.2
OAG. (2019). Official Aviation Guide. Retrieved from www.oag.com
Park, Y., & O'Kelly, M. E. (2016). Origin–destination synthesis for aviation network data: examining hub operations in the domestic and international US markets. *Journal of Advanced Transportation*, *50*(8), 2288–2305. https://doi.org/10.1002/atr.1459
Poorthuis, A., & Zook, M. (2017). Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. *Journal of Urban Technology*, *24*(4), 115–135. https://doi.org/10.1080/10630732.2017.1335153
Rocha, L. E. C. (2017). Dynamics of air transport networks: A review from a complex systems perspective. *Chinese Journal of Aeronautics*, *30*(2), 469–478. https://doi.org/10.1016/j.cja.2016.12.029
Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*. https://doi.org/10.1207/s15327906mbr3304_5
Seshadri, A., Baik, H., & Trani, A. (2007). A Model to Estimate Origin-Transfer-Destination Route Flows and Origin-Destination Segment Flows across the Continental United States, (540).
Song, M. G., & Yeo, G. T. (2017). Analysis of the Air Transport Network Characteristics of Major Airports. *Asian Journal of Shipping and Logistics*, *33*(3), 117–125. https://doi.org/10.1016/j.ajsl.2017.09.002
Strohmeier, M., Martinovic, I., Fuchs, M., Schäfer, M., & Lenders, V. (2015). OpenSky: A swiss army knife for air traffic security research. In *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. https://doi.org/10.1109/DASC.2015.7311411
Sun Country Airlines. (2018). Sun Country Airlines. Retrieved February 10, 2018, from https://www.suncountry.com/booking/search.html
Sun, X., Wandelt, S., & Linke, F. (2015). Temporal evolution analysis of the European air transportation system: Air navigation route network and airport network. *Transportmetrica B*. https://doi.org/10.1080/21680566.2014.960504
Taaffe, E. J. (1956). Air Transportation and United States Urban Distribution. *Geographical Review*, *46*(2), 219–238. https://doi.org/10.2307/211645
Teixeira, F., & Derudder, B. (2018). SKYNET: An R package for generating air passenger networks for urban studies. *Urban Studies*. https://doi.org/10.1177/0042098018803258
US Department of Transportation. (2012). *14 CFR 241 section 19-7 - Passenger Origin-Destination Survey*.
US Department of Transportation, & Office of the Secretary. (2005). *Aviation Data Modernization*.
Vowles, T. M. (2001). The "Southwest Effect" in multi-airport regions. *Journal of Air Transport Management*, *7*(4), 251–258. https://doi.org/10.1016/S0969-6997(01)00013-8
Wei, F., & Grubesic, T. H. (2016). The pain persists: Exploring the spatiotemporal trends in air fares and itinerary pricing in the United States, 2002–2013. *Journal of Air Transport Management*, *57*, 107–121. https://doi.org/10.1016/j.jairtraman.2016.07.018
Wickham, C. (2011). A tale of two Airports: Exploring flight Traffic at SFO and OAK. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1198/jcgs.2011.4de
Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality Assessment Methodologies for Linked Open Data. *Semantic Web – Interoperability, Usability, Applicability*.