Journal of Transport Geography 48 (2015) 52–60

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrg

Modeling monthly flows of global air travel passengers: An open-access data resource

Liang Mao [a,*], Xiao Wu [b], Zhuojie Huang [c,d], Andrew J. Tatem [e,f,g]

[a] Department of Geography, University of Florida, Gainesville, FL, USA
[b] Department of Statistics, University of Florida, Gainesville, FL, USA
[c] GeoVISTA Center, Pennsylvania State University, University Park, PA, USA
[d] Division of Infectious Disease, Key Laboratory of Surveillance and Early-warning on Infectious Disease, Chinese Center for Disease Control and Prevention, Beijing, China
[e] Department of Geography and Environment, University of Southampton, Southampton, UK
[f] Fogarty International Center, National Institutes of Health, Bethesda, USA
[g] Flowminder Foundation, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 8 May 2015
Received in revised form 10 August 2015
Accepted 15 August 2015
Available online xxxx

Keywords:
Worldwide airline network (WAN)
Air passenger flow
Monthly dynamics
Spatio-temporal modeling

ABSTRACT

The global flow of air travel passengers varies over time and space, but analyses of these dynamics and their integration into applications in the fields of economics, epidemiology and migration, for example, have been constrained by a lack of data, given that air passenger flow data are often difficult and expensive to obtain. Here, these dynamics are modeled at a monthly scale to provide an open-access spatio-temporally resolved data source for research purposes (www.vbd-air.com/data). By refining an annual-scale model of Huang et al. (2013), we developed a set of Poisson regression models to predict monthly passenger volumes between directly connected airports during 2010. The models not only performed well in the United States with an overall accuracy of 93%, but also showed a reasonable confidence in estimating air passenger volumes in other regions of the world. Using the model outcomes, this research studied the spatio-temporal dynamics in the world airline network (WAN) that previous analyses were unable to capture. Findings on the monthly variation of WAN offer new knowledge for dynamic planning and strategy design to address global issues, such as disease pandemics and climate change.

1. Introduction

The worldwide airline network (WAN) has played a critical role in contracting human societies into a global village through the rapid transport of people, commodities and information over long distances. Every year, on average 700 million passengers and $6.4 trillion goods are carried by air (Guimerà et al., 2005; Tyler, 2013). Its tremendous impacts on the global socio-economy have drawn increasing attention from a variety of research fields, such as regional studies, international trade, and transportation management (Derudder and Witlox, 2008; Mahutga et al., 2010; O'Kelly and Miller, 1994). Recent reports have also shown that the WAN is both directly and indirectly responsible for inter- and intra-continental spread of diseases, such as the severe acute respiratory syndrome (SARS), dengue fever, and novel H1N1 influenza (CIESIN, 2014; Khan et al., 2009; Lemey et al., 2014; Mangili and Gendreau, 2005; Tatem et al., 2012), as well as the spread of invasive species (Liebhold et al., 2006; Tatem, 2009). As the WAN continues to expand at an exceptional rate, knowledge about its characteristics and evolution is crucial for global economic development and disease control (Bogoch et al., 2015; Derudder et al., 2008; Millard-Ball and Schipper, 2011; O'Connor, 2003), among other factors.

As research interest in the WAN continues to grow, the availability and completeness of air passenger flow data have become key obstacles. To date, the available data sources can be summarized into three categories. The first category refers to commercial providers of worldwide aviation data, such as the International Air Transport Association (IATA) and the Official Airline Guide (OAG). Both the OAG and IATA have complete passenger origin and destination records for sale, but the price can be amounted to tens of thousands in US dollars, along with rigorous restrictions for users. The researchers may need to spend a fortune to obtain these data and be prohibited from sharing them with others. As an alternative, the second category of data sources follow the recent movement on open data, which is a new idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. There are a few number of open access data sources concerning the global airline network, but all of them have limitations. The Skyscanner for Business (http://business.skyscanner.net/portal/en-GB) offers online API services to access live pricing and airfare search history, but these data are not directly related to the real passenger origins and

[*] Corresponding author.

http://dx.doi.org/10.1016/j.jtrangeo.2015.08.017
0966-6923/© 2015 Elsevier Ltd. All rights reserved.

destinations. The Openflights organization (Openflight.org) offers free downloadable datasets, but they are only limited to airports and routes with no details for passenger flows. The Sky Explore (http://geog-cura-osgeo.asc.ohio-state.edu/t100/web/main.html) presents a WebGIS interface to map real passenger flows from multiple data sources, but its data coverage is confined in the North America and Europe.

Rather than releasing real data, the third category of data sources comes from statistical models of air passenger flows (Grosche et al., 2007; Johansson et al., 2011; Long, 1970; Wei and Hansen, 2006), and the modeled flows are published for open access (Huang et al., 2013). These existing models, however, are limited to predicting annual aggregates of air passenger volumes, and thus represent the WAN as a static structure during a year. In reality, the WAN is dynamic over time and space given that passenger volumes fluctuate by month and flight routes open/close by season (Feuerberg, 2008; Grubesic et al., 2009). The data aggregation in current models hides the network dynamics, which could otherwise provide insights into the spatio-temporal patterns of various global processes, such as disease spread and labor migration. For example, many diseases, such as the flu and dengue fever, have seasonal dissemination patterns, and the annually summarized air traffic data is not appropriate to predict the global disease dispersion. The temporary labor migration also varies to fit fluctuating job markets such as the tourism and agriculture, and hence monthly air traffic would offer a more reliable estimation than the annual aggregates. To date, few studies have been devoted to modeling temporally-resolved air passenger flows over the WAN. As a result, little analysis has been conducted on the fine-scale spatio-temporal variation of the WAN within a year.

The purpose of this article is two-fold. First, we refined existing models developed by Huang et al.(2013) to a finer temporal scale and predicted the monthly air passenger flows between directly connected airports worldwide. We also release the modeled monthly flows online for open access. Second, we attempt to understand the monthly WAN as a dynamic by measuring the variation of air passenger flows by month, by route, and by airport.

2. Methodology

2.1. Model design

The WAN was conceptualized as a collection of nodes and links, where nodes represent airports and links represent flight routes between airports. Many empirical studies have shown that air passenger volume is proportional to the population size of the origin and destination cities, and inversely proportional to the geographic distance between origin and destination cities, similar to Isaac Newton's gravitational interaction law (Grosche et al., 2007; Long, 1970; Matsumoto, 2007). A gravity model, thus, can be utilized to estimate the air passenger volume between any pair of nodes. Our model views the air passenger flow as an outcome of spatial interactions between a pair of origin and destination airports, which can be formulated into a multiplicative function of node and link characteristics, as shown in Eq. (1):

$$P_{ij}(t) = S(t) \prod_{k=1}^{n} Node_{i,k}(t)^{\alpha_k(t)} \prod_{k=1}^{n} Node_{j,k}(t)^{\beta_k(t)} \prod_{l=1}^{m} Route_{ij,l}(t)^{\gamma_l(t)}. \tag{1}$$

Where $P_{ij}(t)$ denotes the number of air passengers from airport i to j during month t ($t = 1, 2, \ldots 12$). $S(t)$ is a scaling constant. $Node_{i,k}(t)$ and $Node_{j,k}(t)$ represent the kth characteristic during month t regarding origin airport i and destination airport j respectively, for instance, a socio-economic, demographic, meteorological and network characteristic (Table 1). $Route_{ij,l}(t)$ is the lth measurement of the linkage between airport i and j during month t, for example, the great circle distance, seat capacity, flight frequency, and link type (Table 1). The $\alpha_k(t)$, $\beta_k(t)$, $\gamma_l(t)$ are corresponding coefficients to be estimated.

Table 1
Predictors for the monthly air passenger flows.

Predictors	Descriptions
Node characteristics ($Node_{i,k}$ or $Node_{j,k}$)	
Pop_i	The population size of airport i
PPP_i	The purchasing power index where airport i serves
$In\text{-}degree_{\ i}$	The number of incoming links airport i has in the airport network
$Out\text{-}degree_i$	The number of outgoing links airport i has in the airport network
$Capacity\_In_i$	Total incoming capacity of airport i
$Capacity\_Out_i$	Total outgoing capacity of airport i
[a]$Betweenness\ Centrality_i$	The number of shortest paths going through airport i
$Temperature_i$	Monthly average temperature of airport i.
$Humidity_i$	Monthly average humidity of airport i.
$Precipitation_i$	Monthly average precipitation of airport i.
Route characteristics ($Route_{ij,l}$)	
$Inversed\_Distance_{ij}$	The inverse of the great circle distance between airport i and j.
$Capacity_{ij}$	The total seat capacity of routes between airport i and j.
$Country_{ij}$	Whether the airports i and j are in the same country or not.

[a] More detailed calculation of betweenness centrality can be referred to the Supplementary Material.

2.2. Data collection

2.2.1. Air passenger volume

For model development, we obtained monthly air passenger numbers from the US Air Carrier Statistics T-100 domestic and international segments (www.transtats.bts.gov). These datasets contains market data reported by both US and foreign air carriers, including carrier, origin, and destination for enplaned passengers, freight and mail when at least one point of service is in the US or one of its territories. The dataset of the year 2010 was selected to match the collection time of other data sources, such as the census data. The market data were tabulated into 141,182 records with information concerning the origin airport, destination airport, actual passenger number, and month. In addition, the 2010 total passenger numbers for the top 100 international airports across the world by passenger volume were downloaded from the ACI website (Airports Council International, http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2010-final). This dataset was used as an extra independent source for model validation.

2.2.2. Airport locations and flight routes

Information on a total of 3416 airports across the world were obtained from the 2010 Flightstats database (www.fightstats.com), including names, codes, and geographic coordinates (latitudes and longitudes). To connect these airports into a network, flight routes were further derived from the 2010 scheduled flight capacity dataset purchased from the OAG (www.oag.com). This dataset provides information on direct links (if a commercial flight is scheduled) of origin and destination airports, flight distances, and seat capacity by month for 2010. The airport and route datasets were utilized to compute geographic distances between airports, construct network graphs for each month in 2010, and derive network measurements, such as the in-degree, out-degree, and betweenness centrality.

2.2.3. Population size and economic index

The population data was obtained from the most recent Gridded Population of the World, Version 4 (GPWv4), released by the Center for International Earth Science Information Network (CIESIN, 2014).

The GPWv4 is a minimally-modeled gridded population data set (30 arc-second resolution) that incorporates census population data from the 2010 round of censuses. To extract the population size served by airport, a 200 km buffer was created to reflect the upper distance limit of the catchment area by two hour ground travel to the airport (Lieshout, 2012). The buffer zone was superimposed onto the gridded population dataset and a zonal aggregation was performed to extract the potential serviceable population in the catchment area of the airport.

For the economic development around an airport, a gridded map with a cell resolution of 1° × 1° was obtained from the Geographically based Economic database (G-Econ, http://gecon.yale.edu/). Each grid cell shows the purchasing power parity (PPP) at the cell location (Nordhaus, 2006). The PPP value closest to an airport was extracted and divided by the population in the grid cell to estimate the PPP value per capita for that airport.

2.2.4. Meteorological characteristics

Considering local climate as a driver of air travel, for example through tourism (Bogoch et al., 2015), three climatic variables were selected as airport characteristics, namely, the monthly average precipitation, temperature, and humidity. The data were downloaded from the WorldClim website (www.worldclim.org) as gridded surfaces of 1 × 1 km spatial resolution (Hijmans et al., 2005). Airports were superimposed onto the grids to extract the three climatic variables.

2.3. Model implementation

To identify the best fit model, the general gravity model (Eq. (1)) was transformed into three types of model specifications. The first model was a log-normal model proposed by Balcan et al. (2009), assuming that the natural log of the monthly air passenger volume follows a normal distribution. A logarithm transformation was performed on each numerical variable to maintain the configuration of the gravity model. To improve performance, the model also considered interaction terms between origin and destination nodes, denoted as $Interaction_{ij}$, for example, the product between the total populations of the origin and destination nodes. The first model took a form of Eq. (2):

$$\begin{aligned}\ln\left[P_{ij}(t)\right] = \beta_0(t) + \sum_{k=1}^{n}\alpha_k(t)\cdot\ \ln\left[Node_{i,k}(t)\right] + \sum_{k=1}^{n}\beta_k(t)\cdot\ \ln\left[Node_{j,k}(t)\right] \\ + \sum_{l=1}^{m}\gamma_l(t)\cdot\ \ln\left[Route_{ij,l}(t)\right] \\ + \sum_{p=1}^{q}\theta_p(t)\cdot\ \ln\left[Interaction_{ij,p}(t)\right] + \varepsilon_{ij}\end{aligned} \tag{2}$$

$$\text{where} \quad \ln\left[p_{ij}(t)\right] \sim Normal.$$

The second model assumed that the monthly air passenger volume follows a Poisson distribution, as it is a count. Eq. (1) was transformed into a simple Poisson regression model (Eq. (3)), which has been used in the previous model developed by Johansson et al. (2011). Since the passenger numbers on a flight can never exceed the seat capacity, it is more appropriate to set the seat capacity as an offset, rather than a regular covariate, with its coefficient constrained to 1. Further, a dispersion parameter was added to account for the potential over-dispersion problem. This problem arises from the Poisson distribution, which confines its variance to be equal to its mean. For count data, the observed variance could in fact be greater than the mean, known as over-dispersion. As formulated in Eq. (3), adding a dispersion parameter ϕ allows the variance to vary from the mean value μ, which may produce a better fit. If the estimated ϕ is close to 1, there is probably no over-dispersion problem and vice versa.

$$\begin{aligned}\ln\left[\boldsymbol{E}\{P_{ij}(t)\}\right] = \beta_0(t) +\ \ln\left[Capacity_{ij}(t)\right] + \sum_{k=1}^{n}\alpha_k(t)\cdot\ \ln\left[Node_{i,k}(t)\right] \\ + \sum_{k=1}^{n}\beta_k(t)\cdot\ \ln\left[Node_{j,k}(t)\right] \\ + \sum_{l=1}^{m}\gamma_l(t)\cdot\ \ln\left[Route_{ij,l}(t)\right] \\ + \sum_{p=1}^{q}\theta_p(t)\cdot\ \ln\left[Interaction_{ij,p}(t)\right]\end{aligned} \tag{3}$$

$$\text{where } P_{ij}(t) \sim \text{Poisson}(\mu,\phi), \text{ i.e., } Pr\{P_{ij}(t) = y\} = \frac{\mu^y e^{-\mu}}{y!}, \ y = 0, 1, 2, \cdots$$

$$E\{P_{ij}(t)\} = \mu \text{ and } Var\{P_{ij}(t)\} = \phi\mu$$

The third model is a negative binomial normal model, which is often chosen when the Poisson regression has a poor fit. The model takes the same form as Eq. (3) except that the $P_{ij}(t)$ follows a negative binomial distribution instead of the Poisson distribution. This model can also accommodate the over-dispersion problem for count data under some circumstances.

2.4. Model selection and evaluation

For each model and each month, the US air passenger data and all covariates were input into SAS 9.3 to estimate the model coefficients. The evaluation of model performance included a cross-validation within the US to select the best model, and a validation beyond the US to gauge to model accuracy. The cross-validation was performed as follows: one tenth of observations were randomly selected and held as a testing dataset; the rest of the observations were treated as a training set for model fitting; after the model was built, predictions were made with the testing dataset and then errors were further obtained by computing the differences between the predicted and observed values. This process was repeated 10 times, and the root mean squared error (RMSE) and the mean absolute error (MAE) were computed as evaluation criteria. Lower values of RMSE and MAE indicate a better fit of the model. The model with the smallest RMSEs and MAEs was chosen for subsequent spatio-temporal analysis.

To investigate the model prediction power beyond the US, the best-fit model was expanded to estimate the worldwide passenger flows. The model predicted passenger volumes were compared with those observed at the top 100 world airports reported by the Airports Council International (ACI). We only considered the top 100 world

Table 2
Summary of sample data and cross validation results for each model.

Month	Observed number of flight routes	Mean passengers per route	Lognormal (RMSE, MAE)	Poisson (RMSE, MAE)	Negative binomial (RMSE, MAE)
January	2677	8088.5	1690, 900	1545, 830	1666, 878
February	2562	7803.4	1666, 886	1343, 771	1533, 836
March	2622	9656.2	1993, 983	1509, 807	1797, 917
April	2588	9373.8	2038, 997	1564, 851	1889, 936
May	2678	9348.7	2025, 1051	1730, 915	1952, 1030
June	2770	9578.2	2466, 1224	1996, 1040	2263, 1142
July	2754	10163.5	2649, 1312	2278, 1174	2517, 1255
August	2648	10144.6	2946, 1479	2604, 1321	2794, 1388
September	2563	9060.4	2853, 1460	2578, 1380	2819, 1444
October	2580	9742.3	3181, 1569	2787, 1461	3233, 1603
November	2595	9040.4	3245, 1652	2956, 1567	3171, 1641
December	2707	8861.7	3471, 1709	3078, 1585	3859, 1901

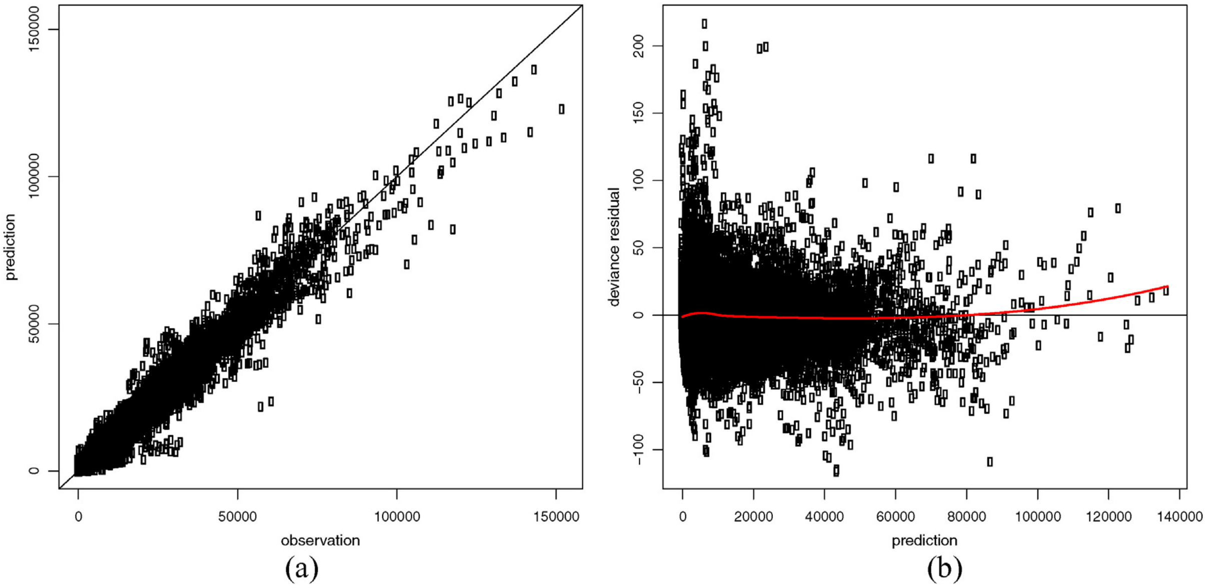


Fig. 1. Diagnostic plots from the best fit model: a) the scatter plot for monthly observed air passengers versus the predicted; b) the residual plot against predictions with a fitted smoothing curve.

airports here, because the ACI only releases these data freely. The Pearson correlation coefficient and the Spearman's rank correlation coefficient were employed for validation.

3. Results and discussion

3.1. Model selection and validation

For each month, the log-normal model produced the greatest RMSE and MAE, followed by the negative binomial model and then the Poisson model (Table 2). For this reason, the Poisson model was selected as the best-fit model for subsequent analysis. Given that the average number of transported passengers was 9239 per month per route, the MAEs suggest that the average prediction error was roughly ±7% and hence the average model accuracy was 93%.

The diagnostic plots (Fig. 1) compare the estimated monthly passenger volumes with corresponding observations. A majority of paired values fall close to the 45° line, indicating a good fit of the model (Fig. 1a). With regard to the error distribution (Fig. 1b), the fitted curve almost coincides with the horizontal line ($y = 0$) showing that the errors scatter randomly with no obvious biases or unusual patterns, and hence the model is appropriate. The spatial distribution of prediction uncertainty by route was also estimated and reported as a map in the supplementary material (Fig. S1).

In addition to the cross validation, the monthly predictions were also aggregated into an annual total for each airport, and compared with the observed annual passengers reported by the Airport Council International (ACI, 2011). Fig. 2 shows a high level of consistency between our predictions and actual observations (correlation coefficients >0.95), in terms of both the magnitude and the ranking order. Although the model was built based on the US data, it can reasonably predict air passenger volumes in other regions.

3.2. Model structure

The estimated coefficients of the Poisson model by month are provided in the supplementary material. In general, most variables were statistically significant, with a few being significant throughout all 12

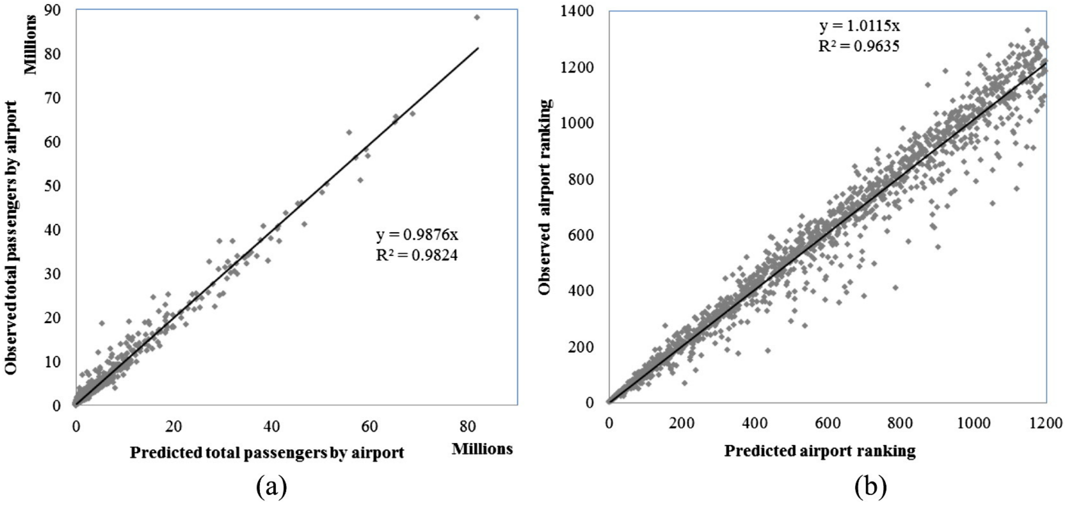
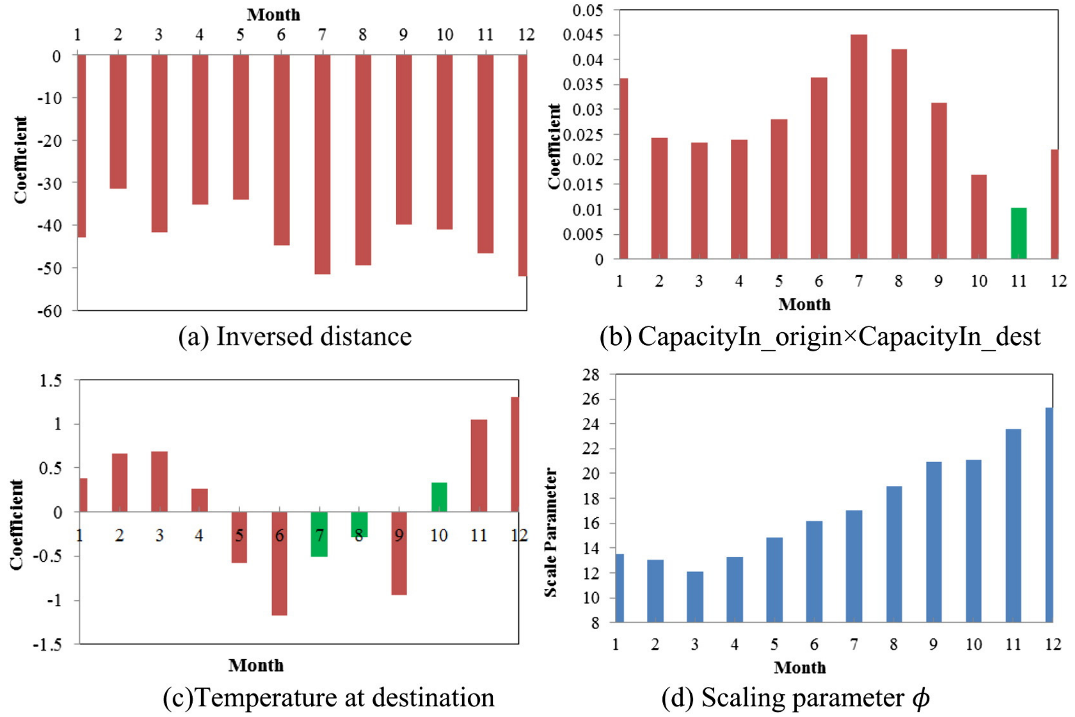


Fig. 2. A comparison of model predictions to the observed airport traffics reported by the ACI in 2010. a) Pearson correlation analysis for passenger volumes; b) Ranked correlation analysis for airport rankings.



(a) Inversed distance

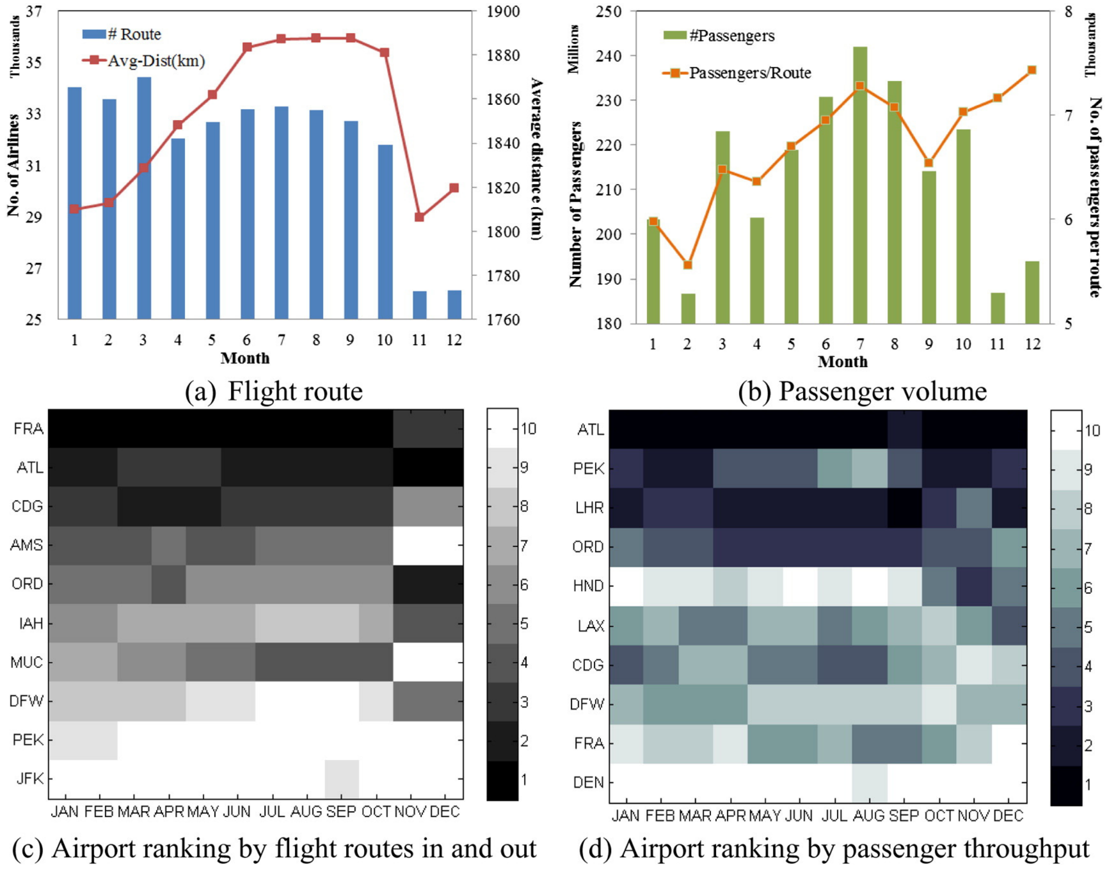
(b) CapacityIn_origin×CapacityIn_dest

(c)Temperature at destination

(d) Scaling parameter ϕ

Fig. 3. Regression coefficients by month for selected covariates. From panel a to c, the black bars indicate statistical significance at a level of 0.05, and gray bars indicate not statistically significant.

months while some were only significant in specific months in 2010. For example, the inverse distance between two airports was statistically significant in all monthly models (Fig. 3a), and shows a negative association with air passengers, reflecting the distance-decay effect of ‘gravity’. The interaction term between the incoming capacities of origin and destination airports was positively associated with the air



(a) Flight route

(b) Passenger volume

(c) Airport ranking by flight routes in and out

(d) Airport ranking by passenger throughput

Fig. 4. Estimated monthly variation of the WAN in terms of it's a) flight routes, b) passenger volume, c) airport rank by flight connections, and d) airport rank by passenger throughput.

passenger volumes throughout the 12 months (Fig. 3b), implying more passengers if there were larger airports at both ends of the journey. The temperature at the destination is another driver of air travel (Fig. 3c). During cold months in the northern hemisphere, such as from January to April and from October to December, the temperature was positively related to the number of air passengers, suggesting that people tend to fly to warmer places. However, the temperature had negative or no associations with air travelers during the summer months in the northern hemisphere from May to September. Lastly, the scale parameter, defined as the square root of the dispersion parameter ϕ in Eq. (3), was estimated to range from 13 to 26 over the 12 months. The rising trend in Fig. 3d implies that the WAN had an increasingly heterogeneous structure over the 12 months, as the passenger volume varies more widely around its mean value.

3.3. Spatio-temporal dynamics

The predicted monthly air passengers of the WAN are included in supplementary material II (data and video clip), and also published online a part of the Vector-Borne Disease Airline Importation Risk (VBD-Air) project (www.vbd-air.com/data/) for free download. Fig. 4 shows the monthly variation of the WAN in terms of its flight routes, passenger volume, and role of airports. Based on the model estimates, the monthly variations in air passenger flows can be roughly divided into three stages in 2010. The first stage spanned from January to March, characterized by the greatest number of flight routes, but the shortest average flight distance and a low passenger volume (Fig. 4a and b). These statistics suggest that stage 1 was dominated by short-range flights and a low passenger volume per route.

The second stage (from April to October) was the peak season of the entire year due to the largest number of passengers. It was also featured by a substantial decrease in flight routes, but also an increase in average flight distance. With fewer operating routes, the number of passengers per route was larger than stage 1 implying higher seat occupancy rates or larger carriers. The third stage included the last two months of the year characterized by the fewest passengers across the year. There was another significant drop in the

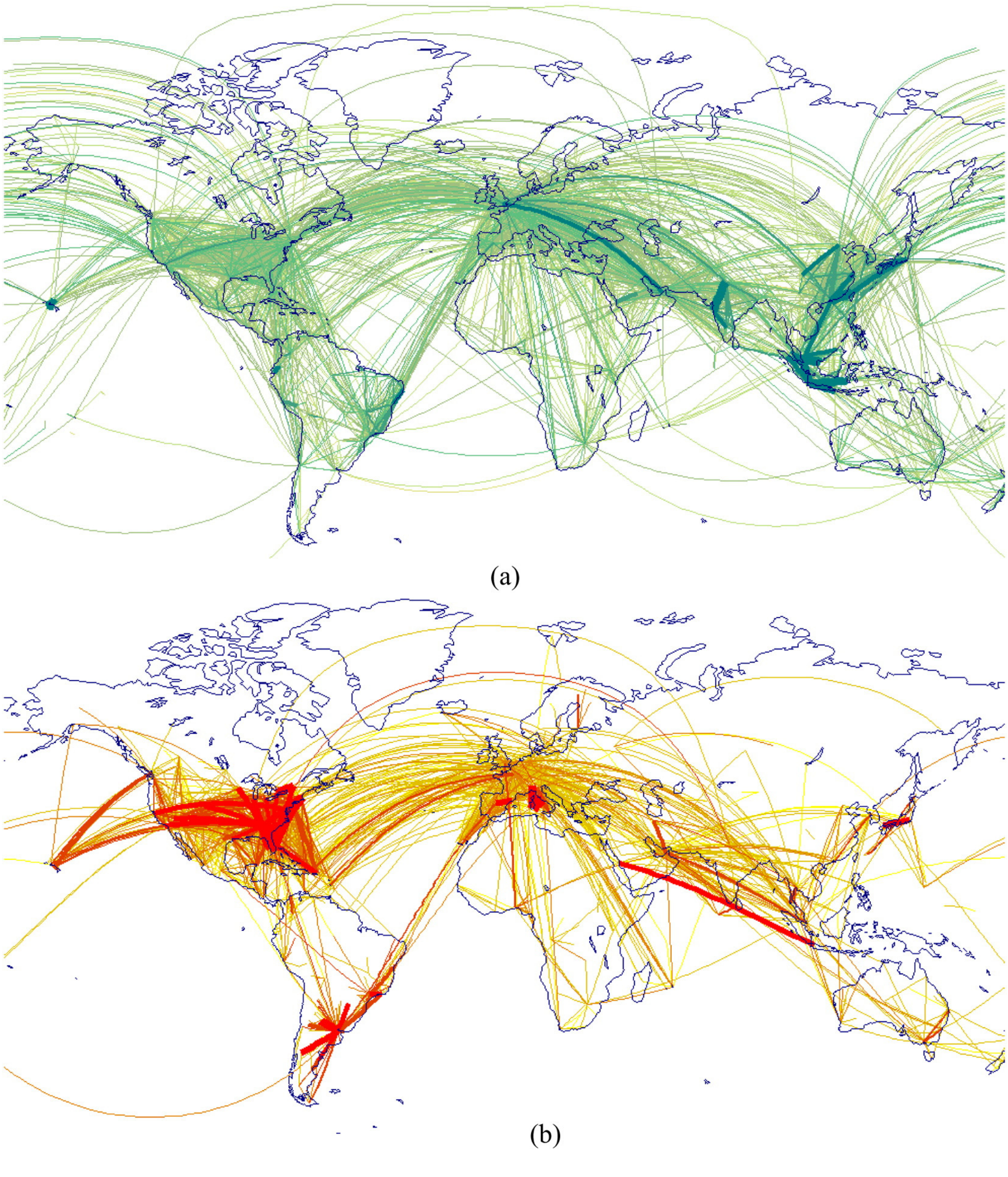


Fig. 5. The estimated decrease (cold colors) and increase (warm colors) of passenger volume from March to April in 2010 (the transition from stage 1 to 2). Darker colors indicate greater changes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

number of flight routes, but the average flight distance per route remained longer than stage 1.

The varying roles of airports over months are also of interest, as shown in Fig. 4c and d. The FRA (Frankfurt) was the most connected airport (with the greatest number of flight routes connected to) in most months, while the ATL (Atlanta) played as the busiest airport (with the greatest passenger throughput), showing a concentration of world air travel activities in Europe and North America. With regard to the flight connections, the top 10 airports in Europe (e.g., FRA and CDG) had higher ranks in the early months of the year, while in late months their roles were gradually replaced by airports in the US, such as the ATL and ORD (Chicago). This might be related to the flight route reduction shown in Fig. 4a (More discussions later). The PEK (Beijing) was the 2nd busiest airport during stage 1 and 3, but it was replaced by the LHR (London) in stage 2, suggesting that a remarkable fall and rise of passenger numbers in East Asia and Europe among stages (More discussions later).

To further explain the temporal variations in Fig. 4, the air passenger flows were mapped geographically to depict where the increase and decrease took place between the three stages of 2010. As shown in Figs. 5 and 6, most changes were concentrated in the northern hemisphere due to the majority of the world's population being distributed there. From stage 1 (January to March) to stage 2 (April to October), significant decreases in travelers were clustered in short-haul flights within East and Southeast Asia (Fig. 5a). Meanwhile, a noticeable increase of long-distance travelers was seen within the US and between Europe and other continents (Fig. 5b). A possible reason is that the stage 2 spanned the late spring to fall in northern hemisphere, which had longer daytimes and warmer weather, and thus encouraged human activities (e.g., tourism) as well as long distance travels. To increase profits, many short-haul flight routes were likely closed temporarily so that the air carriers could be redistributed to serve medium or long-haul routes, for example, the seasonal air routes between Philadelphia (USA) and Barcelona (Spain), and between Atlanta (USA) and Athens (Greece). In addition, the declining economic trends of the time of the data used to construct the model were likely another reason for the reduced number of passengers, given its tremendous impact on airline industry. The global economy was contracting in 2010, and therefore it is

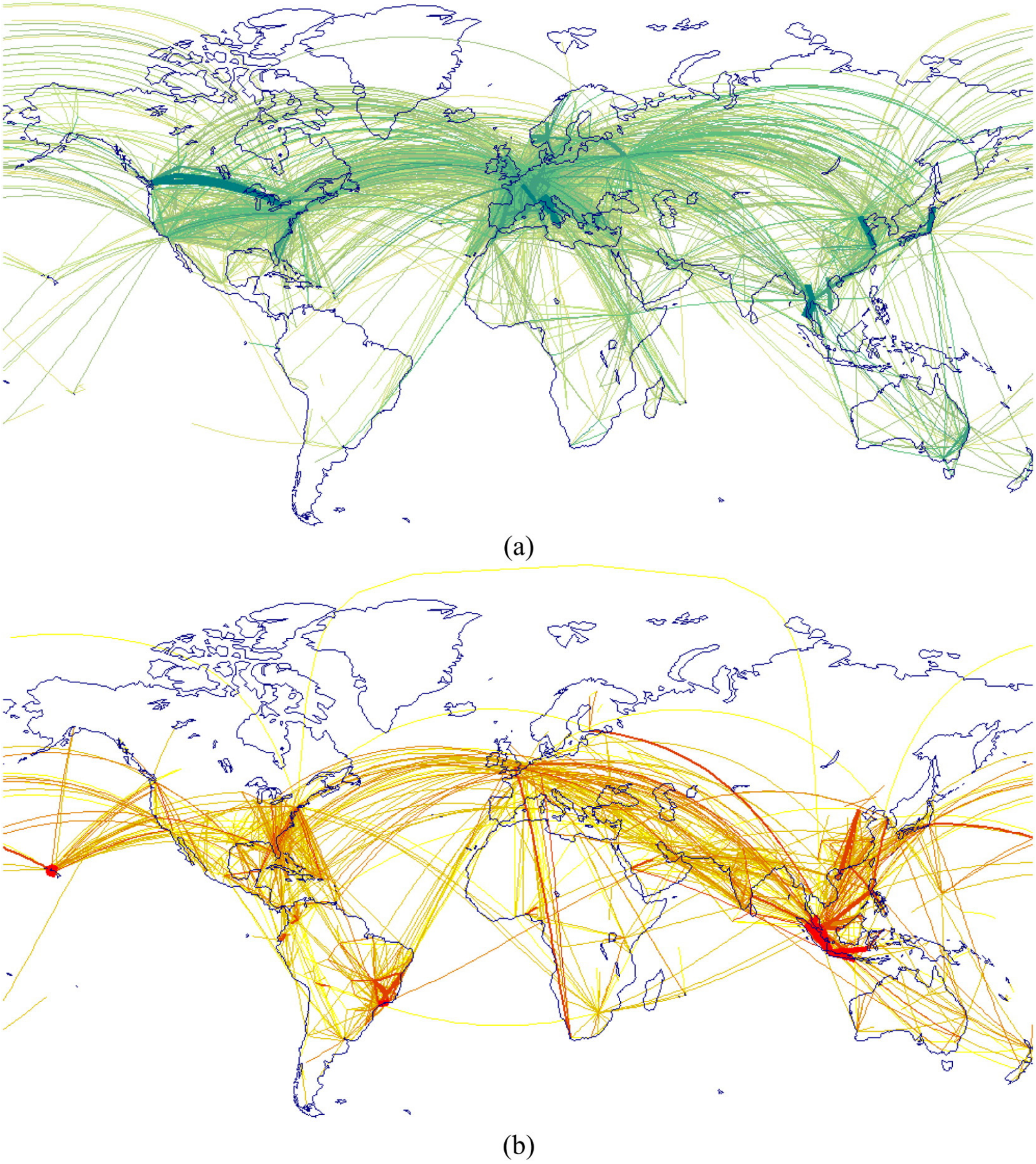


Fig. 6. The estimated decrease (cold colors) and increase (warm colors) of passenger volume from October to November in 2010 (the transition from stage 2 to 3). Darker colors indicate greater changes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

very possible that some variations were attributable to the economic decline.

From stage 2 (April to October) to stage 3 (November to December), there was an apparent shrink of passenger volumes within Europe, between Europe and North America, and between Europe and Asia (Fig. 6a). Those routes with reduced passengers were primarily long-haul flights, possibly because the winter weather in the northern hemisphere discouraged the demands for trans-continental travels. Only a few flight routes had an increased passenger volume, and a majority of them were medium haul flights toward tropical regions, for example, Caribbean islands and Southeast Asia.

3.4. Limitations

There were several limitations of this modeling work that may introduce uncertainties in estimation and bias in interpretation. First, the flight datasets used in model building only record passenger information on direct links between airports. Although the model is capable of estimating the flows between airports separated by 2 or more stops, the estimates could be biased to a certain degree. To address this issue, additional datasets with passenger transfer information, such as flight ticket databases, could be further included in the model. Second, a simple 200 km catchment area was set to estimate the population size served by an airport. The realistic travel time/distance to an airport could vary by airport size, local economic size, population density, travel means, etc. More sophistication should be introduced to delineate the catchment area and better reflect the population sizes at the origin and destination airports. Third, the number of air passengers was assumed to be independent with one another between flight routes to satisfy the assumption of linear regression models. This route independence could be problematic, because the passenger numbers could be related when one route is connected to the other. In future research, the Airline Origin and Destination Survey (DB1B) data, which records itinerary samples, can be further included in our model to account for multiple-stop travels and dependence between flights. At last, the selection of months as the basic temporal unit for modeling is an ad hoc criterion that may impose new problems. For instance, the air traffic often peaks during long holiday periods and school breaks that may cross months. Such fine-grained peaks might be evened in the monthly model and maps.

4. Conclusions

The global flow of air travel passengers varies over time and space, but analyses of these dynamics and their integration into applications in the fields of regional studies, epidemiology and migration, for example, have been constrained by a lack of data, given that air passenger flow data are often difficult and expensive to obtain. Here, these dynamics are modeled at a monthly scale to provide an open-access spatio-temporally resolved data source for research purposes.

The contributions of this research have covered two aspects. First, Poisson regression models were developed to predict monthly passenger volumes in the WAN, which refine the previous annual scale models. The models not only performed well in the United States, but also showed good confidence in estimating air passenger volumes in other regions. The proposed modeling approach can be extended to other years too, if data of those years are available.

Second, the models and estimates are all shared online for researchers to further reveal the monthly characteristics of WAN that previous analyses were unable to capture. Existing studies have devised various tools to understand yearly or quarterly evolution of WAN (Feuerberg, 2008; Grubesic et al., 2009; O'Connor, 2003). These tools help identify global roles of cities, evaluate population's accessibility to air travel, and predict future geographic patterns. It would be interesting to reexamine these topics with the same tools, but a closer lens at the monthly scale. For instance, cities that play important roles in some months, but not the entire year, can be revealed. The effects of seasonal flight routes on the WAN structure can be investigated. The accessibility to air travel can be assessed in a spatial and temporal manner. Such fine-grained analyses on WAN would offer new knowledge for regional planning or dynamic strategy design. For example, those cities that are temporarily important for months in the WAN could be fast growing nodes in the future regional development and are worth attention from urban planners. The monthly assessment of accessibility to air travel may suggest dynamic airfare strategies to mitigate local and regional biases in time and costs. The World Health Organization can also identify possible high-risk routes for the next (few) month(s) according to the monthly WAN structure and disease prevalence, and then optimally focus its control efforts (e.g., airport surveillance) to these routes.

Acknowledgment

This research was funded by the National Academies National Academy of Sciences Airport Cooperative Research Program Board (#ACRP A02-20).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jtrangeo.2015.08.017.

References

ACI, 2011. World Airport Traffic Report 2010, Geneva.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A., 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc. Natl. Acad. Sci. 106, 21484–21489.

Bogoch, I.I., Creatore, M.I., Cetron, M.S., Brownstein, J.S., Pesik, N., Miniota, J., Tam, T., Hu, W., Nicolucci, A., Ahmed, S., 2015. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. Lancet 385, 29–35.

CIESIN, 2014. Gridded Population of the World, Version 4 (GPWv4), Preliminary Release 2 (2010). Columbia University.

Derudder, B., Witlox, F., 2008. Mapping world city networks through airline flows: context, relevance, and problems. J. Transp. Geogr. 16, 305–312.

Derudder, B., Witlox, F., Faulconbridge, J., Beaverstock, J., 2008. Airline data for global city network research: reviewing and refining existing approaches. GeoJournal 71, 5–18.

Feuerberg, G., 2008. Uncovering Trends in Seasonal Transportation Data: New Tool Analyzes Airline Revenue Passenger-Mile Trend Data. Bureau of Transportation Statistics.

Grosche, T., Rothlauf, F., Heinzl, A., 2007. Gravity models for airline passenger volume estimation. J. Air Transp. Manag. 13, 175–183.

Grubesic, T.H., Matisziw, T.C., Zook, M.A., 2009. Spatio-temporal fluctuations in the global airport hierarchies. J. Transp. Geogr. 17, 264–275.

Guimerà, R., Mossa, S., Turtschi, A., Amaral, L.N., 2005. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc. Natl. Acad. Sci. 102, 7794–7799.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

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, e64317.

Johansson, M.A., Arana-Vizcarrondo, N., Biggerstaff, B.J., Staples, J.E., Gallagher, N., Marano, N., 2011. On the treatment of airline travelers in mathematical models. PLoS ONE 6, e22151.

Khan, K., Arino, J., Hu, W., Raposo, P., Sears, J., Calderon, F., Heidebrecht, C., Macdonald, M., Liauw, J., Chan, A., 2009. Spread of a novel influenza A (H1N1) virus via global airline transportation. N. Engl. J. Med. 361, 212–214.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C.A., Smith, D.J., Pybus, O.G., Brockmann, D., 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 10, e1003932.

Liebhold, A.M., Work, T.T., McCullough, D.G., Cavey, J.F., 2006. Airline baggage as a pathway for alien insect species invading the United States. Am. Entomol. 52, 48–54.

Lieshout, R., 2012. Measuring the size of an airport's catchment area. J. Transp. Geogr. 25, 27–34.

Long, W.H., 1970. Air travel, spatial structure, and gravity models. Ann. Reg. Sci. 4, 97–107.

Mahutga, M.C., Ma, X., Smith, D.A., Timberlake, M., 2010. Economic globalisation and the structure of the world city system: the case of airline passenger data. Urban Stud. 47, 1925–1947.

Mangili, A., Gendreau, M.A., 2005. Transmission of infectious diseases during commercial air travel. Lancet 365, 989–996.

Matsumoto, H., 2007. International air network structures and air traffic density of world cities. Transp. Res. E Logist. Transp. Rev. 43, 269–282.

Millard-Ball, A., Schipper, L., 2011. Are we reaching peak travel? Trends in passenger transport in eight industrialized countries. Transp. Rev. 31, 357–378.

Nordhaus, W.D., 2006. Geography and macroeconomics: new data and new findings. Proc. Natl. Acad. Sci. U. S. A. 103, 3510–3517.

O'Connor, K., 2003. Global air travel: toward concentration or dispersal? J. Transp. Geogr. 11, 83–92.

O'Kelly, M.E., Miller, H.J., 1994. The hub network design problem: a review and synthesis. J. Transp. Geogr. 2, 31–40.

Tatem, A.J., 2009. The worldwide airline network and the dispersal of exotic species: 2007–2010. Ecography 32, 94–102.

Tatem, A., Huang, Z., Das, A., Qi, Q., Roth, J., Qiu, Y., 2012. Air travel and vector-borne disease movement. Parasitology 139, 1816–1830.

Tyler, T., 2013. IATA Annual Review 2013. International Air Transport Association, Cape Town, pp. 38–39.

Wei, W., Hansen, M., 2006. An aggregate demand model for air passenger traffic in the hub-and-spoke network. Transp. Res. A Policy Pract. 40, 841–851.