Facultad de Economía y Negocios
Business and Economics School
Working Paper

A methodology for calculating the unmet passenger demand in the air transportation industry

Rafael Bernardo Carmona Benítez[1], *Maria Rosa Nieto*[2]

Universidad Anáhuac México

A methodology to estimate the unmet demand is developed using machine learning algorithms. The unmet demand in an origin-destination airports pair (OD pair) is the unattended number of passengers that could not flight because of economic conditions of supply and demand. The forecast of the unmet demand is important for strategic decisions of new planning such as opening new routes, increasing/decreasing number of services, and aircraft choice. The first contribution of this paper is to develop a single-class methodology to unconstraint or detruncate pax demand to estimate the market size of an OD pair. This methodology mixes time-series methods with the bootstrap distribution function and machine learning algorithms. This methodology considers socioeconomic variables at community zone and airport levels to forecast the market size of an OD pair. The second contribution of this paper is to design a methodology that estimates the unmet demand of an OD pair. The advantage is its ability to simulate the unmet demand based on statistical analysis with a confidence level of (1-α)%. The calculations are evaluated by describing the distribution of the market size historical data because distribution functions give the possibility to calculate pax demand without knowing the parameters that have an influence on it. Finally, the third contribution of this paper is to develop an approach to identify new airline OD pairs which could be considered as potential airline markets based in the calculation of the OD pair unmet demand. The proposed methodology is applied to the US air pax industry as case study. The results indicate that hubs airports are under extreme competition. Small and primary airports located in big cities are not under competition in some quarters of year meaning that socioeconomic factors among airports change according with the seasonality of year.

1. Introduction

Airlines and airports around the world are continuously looking to open new flights (non-stop and/or direct) to grow their networks and consequently increase their revenues. A challenge to airlines and airports is to calculate the true pax demand because the only data they can gather is the actual number of bookings and not the total pax demand willing to fly between an origin airport (O) and a destination airport (D). This is because customers could not book due to capacity constraints or booking limits owing to revenue management (RM) systems strategies that airlines apply (Weatherford & Pölt, 2002), and/or simple because the origin-destination pairs of airports (OD pairs) has not been operated by any airline yet. Research on developing models and methodologies for estimating potential markets allows airlines and airports to decide in what OD pairs to open new flights or increase the

1 Professor and Researcher at Universidad Anáhuac México, Facultad de Economía y Negocios. PhD in Transport Engineering and Logistics at Delft University of Technology. Address: Av. Universidad Anáhuac 46, Huixquilucan, Estado de México, 52786, México. Email: rafael.carmona@anahuac.mx ; https://orcid.org/0000-0002-8369-2748

2 Professor and Researcher at Universidad Anáhuac México, Facultad de Economía y Negocios. PhD in Cuantitative Methods at Universidad Carlos III de Madrid. Address: Av. Universidad Anáhuac 46, Huixquilucan, Estado de México, 52786, México. Email: maria.nieto@anahuac.mx ; https://orcid.org/0000-0002-9325-248X

number of them. An OD pair is a connection between two airports served either by non-stop flights or by direct flights operated by one, or more airlines or not being served by any airline yet. In an OD pair, the total passenger demand (total pax demand) or the potential market size (MS) (unconstrained pax demand data) can be defined as the sum of the enplaned passengers demand (Pax) (censored pax demand data or constrained pax demand data), which is the actual number of bookings, plus the unattended passenger demand (unmet) which is the unobserved pax or unattended pax of an OD pair. The MS of an OD pair must be extrapolated from censored booking data (Pax) in a process called demand unconstraining or detruncating to be forecasted (Guo, Xiao & Li, 2012) since unconstraining demand is not easy to measure (Wickham, 1995). In fact, demand unconstrained is normally unknown, reason why, it is considered the "Holy Grail" of forecasting because, for example, in the airline industry, it is known that unconstrained pax demand data calculates better forecasts and improve revenues gains (Guo, Xiao & Li, 2012) in 2-12% (Weatherford & Pölt, 2002). Knowing the advantages of forecasting the MS, this paper makes the next contributions. First, it proposes a single-class methodology to unconstraint or detruncate Pax to estimate the MS of an OD pair modelling the bootstrap distribution function of their O-D relation pax demand rather than assuming the normal distribution function. This methodology mixes time-series methods with the bootstrap distribution function and machine learning algorithms. Second, it proposes a methodology to estimate the unmet demand of an OD pair; and third, we develop an approach to identify new airline OD pairs which could be considered as potential airline new markets. This approach demonstrates an applicability of the proposed methodologies. The case study is The United States (U.S.) air passenger transportation system.

In an OD pair, the unmet demand exists despite the best attempts by airlines to anticipate and respond to the demand of seats, because the possibility exists that the total demand is not satisfied by the supply of seats that airlines offer because of economic conditions of supply and demand interactions (Haensel, Koole & Erdman, 2011). It means, the MS of an OD pair could be bigger than its Pax. In the case of the air pax transport industry, the unmet demand of an OD pair results when airlines do not satisfy its MS because they do not supply enough seats (Chandukala et. al., 2011). This happens when passengers cannot travel because of high-ticket fares or for the lack of seats offered by airlines. However, these reasons do not mean an OD pair might not have more Pax to attend on different conditions. So, in an OD pair, the concept of unmet demand can also be defined as the seats that passengers are willing to buy at a certain ticket price, or the number of seats that airlines do not offer to serve the MS of the OD pair. An example of this is the fact that in 1984, nobody would anticipate that Dubai International Airport (DBX) would be in the top three worldwide airports in terms of pax demand in 2017 (Neufville, 2017) and one of the airports with the highest average load factor per air transport movement (ATM). This happened mainly because Emirates Airlines identified OD pairs with enough unmet demand to open new flights at DBX increasing the total number of seats per OD pair, as a result Emirates Airways and DBX networks grew, and therefore their market shares and revenues.

The proposed methodology estimates the unmet demand of OD pairs in five steps. Step 1, the proposed methodology classifies airports into different sets of airports using clustering methods from machine learning and data mining algorithms. Clustering methods are popular in multivariate analysis which is important because airports can be classified based on different socioeconomic factors that determine pax demand in a certain market, connectivity at the airport, and competition (measured as airline market concentration and airport competition). Step 2 based on these airport clusters, OD pairs are classified into sub clusters according with the relation between the O cluster and the D cluster (O-D relation), i.e., if O is classified into cluster A and D is classified into cluster B, then, the OD pair is classified into an O-D relation or sub cluster AB. Step 3 proposes a mathematical model to calculate the MS of OD pairs by modelling the distribution of pax demand for each O-D relations or sub cluster applying Bootstrap methods and Step 3 also proposes a mathematical model to calculate the Pax of OD pairs by modelling the distribution of pax demand for each of them applying Bootstrap methods. The results are the propositions of distributions for estimating the MS of each OD pair and distributions for estimating the Pax of each OD pair. Step 4, the pax demand of OD pairs are forecasted using a time series method. In this paper, the ARIMA+GARCH+Bootstrap time series method is applied. Nieto & Carmona-Benitez (2018) validated and proved this time series method to be the most accurate to forecast Pax of OD pairs at the air pax transportation industry. The advantage of this forecasting method is its capacity to develop a forecasting models per OD pair and per O-D relation. Step 5, a mathematical model is proposed to calculate the unmet demand of an OD pair. This model uses the forecast of the O-D relations or sub clusters as the MS of OD pairs and the forecast of the total Pax of OD pairs (Step 3).

In this paper, the unmet demand methodology applies the ARIMA+GARCH+Bootstrap to forecast the potential MS of the OD pairs and their Pax. It is important to mention that, in this paper, the ARIMA+GARCH+Bootstrap forecasting method is applied because Nieto & Carmona-Benitez (2018) prove that this method is more appropriate for long-term Pax. However, the methodology is opened to use any other forecasting method that

better calculates Pax, as long as, the forecasting method is validated with the analysis of the root mean square error (RMSE).

The mathematical models of the unmet demand methodology are empirically applied and validated using data from the U.S. domestic air passenger transport market (1st quarter of 1993 to the 4th quarter of 2015 (Bureau of Transportation Statistics, BTS)). The results show the classification of airports in clusters; the forecast of MS and Pax; and finally, the applicability and use of the proposed methodologies is demonstrated by identifying new airline flights in the U.S. domestic market.

This paper is organized as follows: in section 2, a review on unmet demand studies, a review on segmentation or clustering methods in the air pax transportation industry studies, and a review on the latest papers that study socioeconomic factors for estimating Pax, airport competition, and airline market share are presented; in section 3, the methodology to estimate potential pax per OD pair is developed, and one strategy to determine what OD pairs could have enough unmet demand to open new flights is designed; in section 4, the empirical application and results are presented for the domestic U.S. passenger airline industry to show an applicability of the proposed methodology in real practice; finally, in section 5, conclusions are drawn, and future research topics are provided.

2. Literature review

Many models have been developed to forecast Pax based on historical data with the goal of finding balance between supply and demand which is a very complex problem to tackled (Birolini et al., 2021). As explained in the introduction, in the airline industry, historical data is booking data which is censored or data constrained what represents a major problem for forecasting, since these data is bias because of managerial decisions and the result is the loss of demand information. Five approaches exist to unconstraint or detruncate censored data: "directly observed and record latent demand; leave data constrained, ignoring the fact of censorship; use unconstrained data only and discard censored ones; replace censored data using imputation methods; or statistically unconstraint the data" (Guo, Xiao & Li, 2012). Direct observation considers declined bookings as unmet demand (Pölt, 2000). This methodology presents disadvantages: one customer can make multiple inquiries, it does not consider the difference between a denial and regret, and it also does not consider the level of uncontrol channel (Orkin, 1998). Ignore the censored data as if the censoring never happened, this approach is also known as method Naïve #1 (N1) (Weatherford & Pölt, 2002). This method presents one disadvantage: its application reports a potential significant difference between the forecast and real demand. Discard the censored data is a method, known as Naïve #2 (N2), that uses unconstrained data only and neglects the censored data to deal with incomplete data (Zeni, 2001). This method does well when data are censored totally at random, there is only small quantity of lost data, and the variables under study are not associated to the instrument triggering off censorship (Guo, Xiao & Li, 2012). This method presents one disadvantage: variables under study cannot be correlated because discarding them will lead to a biased forecast (Zeni, 2001). Imputation methods replace missing data with probable values. The goal is to complete the data set and treat them as unconstrained demand. The mean or Naïve #3 (N3), median, and percentile are imputation methods (Little & Rubin, 1987). Saleh (1997) studies the performance of the N1, N2 and N3 methods. His conclusion is that N3 method performs better than the N1 and N2. Finally, statistical unconstraint data methods are more complex and reliable than imputation methods because they are built on a foundation of statistics theory (Zeni, 2001). Statistical unconstrained data methods depend only on observed bookings and cover an array of optimization and heuristic techniques (Queenan et al., 2007). In the airline industry, these methods are divided into univariate and disaggregate demand models also known as single-class methods, multi-class methods, and multi-flight methods. Single-class methods estimate the parameters of an assumed probability distribution function to estimate the conditional mean or the conditional median, and variance. Their main disadvantage is they assume Pax is uncorrelated between flight classes. The Expectation-Maximization (EM) algorithm is a single-class method and it is the most popular for detruncating demand in revenue management (Dempster et al., 1977) because it has shown success where censored data, missing observations, and truncated distributions are present (McLachlan & Krishnan, 1997). The EM algorithm has been mixed with time series methods to improve forecasts (He & Luo, 2006) (Guo et al., 2008). Multi-class methods mix single-class methods to consider Pax correlation or interactions between flight classes. For example, Karmarkar et al. (2011) use the EM algorithm assuming a bivariate normal distribution. Finally, multi-flights methods calculate the probability that passengers buy tickets fares in specific flight classes. These detruncating methods are the most difficult to estimate because of their complexity and the level of desegregation of data (Guo, Xiao & Li, 2012).

The unmet demand methodology proposed in this paper is based on clustering methods for segmentation using socioeconomic factors that determine Pax in the air pax transportation industry, and other factors that assess airport connectivity, airport competition, and airline market concentration. Therefore, according with literature review,

the proposed methodology is a new single-class statistical unconstrained data method because it estimates the parameters of an assumed probability distribution function to estimate the conditional mean or the conditional median, and variance.

In literature, different methodologies have been developed to estimate the unmet demand of an OD pair. Anderson & Kraus, (1981), Andereck & Caldwell, (1994), and Park & Pan, (2018) directly concerned with the identification of unmet demand. Anderson & Kraus, (1981), Andereck & Caldwell, (1994), use market segmentation and Park & Pan (2018) mix buying funnel theory with gravity model. Park & Pan (2018) is a relevant study because they propose a model that estimates unmet demand on OD pairs to identify new non-stop flights. There are other studies that are closely related to the unmet demand estimation problem, and their concern with the identification of potential tourism demand of different needs or targeting potential markets (Andereck & Caldwell, 1994; Heung, Kukusta & Song, 2011; Jang, Morrison & O´Leary, 2002; Müller & Hamm, 2014; Tkaczynski, Rundle-Thiele & Beaumont, 2009). These studies use market segmentation as a methodology to identify the potential tourism demand. Target marketing is a strategy adopted by companies to divide the market into segments and design products or services to each segment (Camilleri, 2018). In the airline industry, segmentation has relayed mainly on business and economic passengers and flight class (Teichert, Shehu & Wartburg, 2008) but since the liberalization of the air pax transportation industry in the U.S., this industry has increased in competition because airlines are fighting for market share, and airports too. Traditional passenger segmentation in business and leisure is obsolete because passengers' preferences for both classes are becoming wider (Alamdari & Mason, 2006). The competition between airlines has resulted in the development of new airlines business models. The convergence between full-service carriers (FSC) business model and low-cost carriers (LCC) business model has produced different airline strategies, all seeking to exploit profitable opportunities by segmenting and commoditizing passengers' markets per OD pair by using fares (yields) as factor (Guillen & Lall, 2018). This increment in airline competition obligates to consider also other factors for OD pair segmentation such as average fare, average yields, flexibility, airport connectivity, airline competition, airline market share, and socio-demographic variables. Toh & Hu (1990) use multiple discriminant analysis to categorize between frequent-flier members and non-members, and after segment into light, medium and heavy consumers. They use behaviour, attitudes, and demographic factors. Teichert, Shehu & Wartburg (2008) propose a latent class model for OD pair segmentation to analyse the choice made by passengers. Latent class models assume units of the same segment show similar joint probability distribution amongst the observed variables (Vermunt & Magidson, 2003). Teichert, Shehu & Wartburg (2008) use behavioural and socio-demographic variables to segment Pax and then identify customer preferences. They conclude that marketing strategies (price, flexibility, comfort, and airport selection) interact with socio-demographic factors because Pax may be geographically different. Chiang (2017) develop a model using data mining technologies to determine Pax for airlines and find high-value airlines markets. Marcos, García-Cantú & Herranz (2018) develop a machine learning approach for the estimation of airlines route choices between an O airport and a D airport. This study is highly relevant because their methodology cluster routes (considering navigation charges, route length, and congestion) and segment flights (considering airline business type and time of the day). Their approach forecasts Pax in two steps: first, flights are segment based on flight characteristics; then, airline choices are modelled per segment. Urban et al. (2018) identify airline clusters besides LCCs and FSCs. First, they use the Business Model Canvas to identify key characteristics between airline business models. Second, they apply a two-step cluster analysis to segment airlines into different groups.

Like airlines, airports have also been affected by the changes in the air transport industry. Airports have applied passengers' segmentation strategies to develop airport commercial strategies. Freathy & O´Connell (2012) use a framework of market segmentation to distinguish how passengers spend their time inside the airport terminal and classified them according with their shopping behaviour. Harrison, Popovic & Kraal (2015) propose a model for passenger segmentation based on time with the aim to allocate space for terminal planning and design. Leung, Yen & Lohmann (2017) perform a geo-demographic classification analysis using origin location data (home postcodes, trip characteristics and airport preferences) and socioeconomic variables.

Pax forecasting is relevant to this study because forecasting is needed to properly segment airports and OD pairs, because it is compulsory to understand what factors determine Pax. The number of papers addressing the problem of forecasting Pax is vast, but this paper focuses on recent publications addressing socioeconomic factors, airline and airport competition, and airline market share. Hofer, Kali & Mendez (2018) is one of the newest papers examining how socioeconomic factors determine Pax at U.S. airports. They develop an econometric model to estimate and forecast Pax. They use the socioeconomic mobility factors (SEM) developed in Chetty et al. (2014a) to design their model. Chetty et al. (2014a) develop two SEM factors (relative mobility and absolute mobility) with data from de-identified federal income tax record. As explained in Hofer, Kali & Mendez (2018) relative mobility (RelMob) measures the relation between parents with low-income rank and their progenies income, and

absolute mobility (AbsMob) measures the relation between the expected economic outcomes in terms of economic rank. Chetty et al. (2014a) calculate SEM measures for 741 "communiting zones" (CZs) in the U.S. Hofer, Kali & Mendez (2018) include these factors and develop an econometric model to estimate Pax. The contributions of Hofer, Kali & Mendez (2018) are highly relevant to this paper because in the previous paragraphs, we present different papers addressing the problem of targeting airline and airport markets with segmentation, all these papers use socioeconomic and socio-demographic factors, and Hofer, Kali & Mendez (2018) demonstrate that SEM factors are determinants of Pax. Then, their SEM factors (relative mobility and absolute mobility) can be used to segment U.S. airports at community zone level which is highly important in this study because the proposed methodology studies Pax at airport level. Moreover, Chetty et al. (2014a, 2014b) conclude that SEM captures important socioeconomic aspects of economic areas, because Chetty et al. (2014a, 2014b) demonstrate that: SEM are significantly different between U.S. CZs; SEM are correlated with socioeconomic factors such as race, public education, and social capital index; and SEM are uncorrelated with income level (Hofer, Kali & Mendez, 2018). These conclusions explain SEM as demographic factors too. Hofer, Kali & Mendez (2018) use other socioeconomic factors to measure airport and airline competition and market share. In this study, Hofer, Kali & Mendez (2018) factors are used to segment airports and to estimate Pax. These socioeconomic factors measure the competition between airports and between airlines. The competition between airports is measured with socioeconomic mobility factors and Pax determinants. The socioeconomic mobility factors are: AbsMob, RelMob, Income, and Population. Pax determinants are: AltAirport (number of commercial airports inside each CZ); DestCount (number of OD pairs from an origin airport); Yield (Average yields at an airport); LM_Hub (Airports are classified into medium and large hubs using the Federal Aviation Administration (FAA) classification); and AggRouteHHI (airline market concentration at an airport). So, the unmet demand methodology (Section 3) proposed in this paper considers socioeconomic mobility factors and Pax determinants to cluster airports, and then estimate the unmet demand for OD pairs in the U.S. air passenger industry.

The difficultness of forecasting Pax is finding the most accurate forecasting method. In literature, many quantitative forecasting methods have been used to calculate Pax in the air pax transportation industry. Quantitative forecasting methods are classified into causal models (multiple-regressions, data panel, and gravity models) and time-series methods (Solvoll, Mathisen & Welde, 2020). In this study, we decide to use a time-series method because these methods are commonly used to estimate Pax in the air transport industry. The most applied time-series methods are: Exponential Smoothing, Holt-Winters Additive and Multiplicative, Grey, and SARIMA models (Nieto & Carmona-Benitez, 2018). The reason why these methods are commonly used in this industry is the advantage of mixing them to consider trend, seasonality, data variability, and Pax distribution over time i.e. ARIMA+GARCH+Bootstrap developed by Nieto & Carmona-Benitez, (2018) and SARIMA Damp Trend Grey developed by Carmona-Benitez & Nieto (2020). So, new time-series methods can be developed as the combination of other time-series methods to improve forecast accuracy (Carmona-Benitez & Nieto, 2020). Knowing this, in this paper, the ARIMA+GARCH+Bootstrap forecasting method is applied to predict Pax because Nieto & Carmona-Benitez (2018) proves that this method is more appropriate for long-term Pax than the: ARIMA, Holt-Winters Additive, Holt-Winters Multiplicative, and the Damp Trend Grey Model (DTGM) which it has been proved to be better than the Grey Model (Carmona-Benitez et al., 2013).

3. Unmet demand methodology

This section develops a methodology that forecast the unmet demand of an OD pair. The main problem is, if actual Pax data is used, it is not possible to know what the real unmet demand is (Zeni, 2001). In fact, the unmet demand cannot be known, but an estimation of the potential Pax can be calculated by simulating the MS data behaviour. This can be done by describing the distribution of the MS historical data. A distribution function gives the possibility to calculate Pax without knowing the parameters that have an influence on it (Carmona-Benitez, 2012).

The proposed methodology estimates the unmet demand of an OD pair in five steps:

- Step 1: segment airports in airport clusters using a machine learning classification method,
- Step 2: segment OD pairs in O-D relations or sub clusters,
- Step 3: forecast the MS of the OD pairs,
- Step 4: forecast the Pax of the OD pair,
- Step 5: calculate the OD pair unmet demand (UPax).

Step 1: segment airports in airport clusters using a machine learning classification method.

A machine learning classification method is applied to segment airports in clusters in a way that each airport belongs to one and only one airport cluster. To do so, clustering partitioning approach is implemented. One of the most widely used partitioning approaches is the *k*-means algorithm. Its purpose is to divide a sample in a predetermined number of groups, *k*. The *k*-means algorithm is the following (Peña, 2002):

1. Select *k* points as centre of the starting groups,
2. For the *k* groups, calculate the Euclidean distance from each element to the centre.
3. Define an optimality criterion to test if reallocating elements of one group to another the criteria improves.

A problem for the application of the *k*-means algorithm is to choose the number of clusters. To solve this problem, different methodologies have been proposed. One of them is based on an approximation of the *F* test of variability reduction with the aim to create homogenous groups that minimize the within-cluster sum of squares. This test calculates the proportional variability reduction obtained when go from *k* to *k+1* cluster (Peña, 2002). The *F* test value is compared with the critical value of a *F* distribution with *p*, *p(n-k-1)* degrees of freedom, where *p* is the number of variables used in the *k*-means methodology for segmenting data. Another procedure commonly used to calculate the number of clusters is the elbow method which is an empirical method to calculate the best number of clusters. The algorithm initializes with one cluster (k=1), then the number of clusters are incremented (k>1) and the sum squared error (SSE) is calculated per cluster until a predetermined maximum number of clusters (max *k*) are evaluated. The SSE of the *k* clusters form a graph where the angle of the elbow is the minimum SSE which is related to the best cluster number known as optimal *k* (Umargono, Suseno & Gunawan, 2020).

Step 2: segment OD pairs in sub clusters (O-D relation).

In this step, sub clusters are created based in the airport clusters generated in step 1 by linking them. The aim is to create connections between airport clusters (O-D relation) by interpolating the sub clusters median to airports that do not have certain connections.

Once all airports in sample are classified into an airport cluster, OD pairs can be classified into sub clusters according with their O cluster and their D cluster (O-D relation). Data gathered is the OD pairs median Pax per quarter of year because average bias must be avoided, and the median statistic allows that. In the following step, a model for forecasting the distributions of O-D relations is proposed.

Step 3: forecast the MS of the OD pairs.

The MS of an OD pair is calculated by forecasting the distributions of Pax of its O-D relation, as follows: Let assume that the MS of OD pairs follows the next model:

$$MS_t = \mu_t + \sigma_t \epsilon_t \quad (1)$$

Where:

μ_t = MS_t conditional mean at time t [passengers]
σ_t = MS_t conditional standard deviation at time t [passengers]
ϵ_t = MS_t distribution of the error term with zero mean and unit variance at time t [passengers]

where, $t = 1..., T$ and T is the length of Pax time series.

Pax within each O-D relation follows model of Eq. 1. This paper applies the ARIMA+GARCH+Bootstrap methodology developed by Nieto & Carmona-Benitez (2018) to estimate and forecast MS_t distribution per OD pair allowing to calculate any statistics metrics from MS_t distribution, specifically any quantile. In the ARIMA+GARCH+Bootstrap methodology, μ_t is estimated with ARIMA forecasting model, σ_t is estimated using a GARCH(1,1) forecasting model, and the distribution of ϵ_t is approximated using the bootstrap methodology. As it has been explained, the reason of using this forecasting method is because Nieto & Carmona-Benitez (2018) have previously validated its superior predictive ability by comparing it with the most frequently used forecasting models in the air pax transportation industry such as Holt-Winters additive and multiplicative models, ARIMA model and the DGTM models.

Step 4: forecasting the Pax of an OD pair.

Let assume the Pax of OD pairs follows the next model:

$$Pax_s = \mu_s + \sigma_s \epsilon_s \quad (2)$$

Where:

μ_s	= Pax_s conditional mean at time s	[passengers]
σ_s	= Pax_s conditional standard deviation at time s	[passengers]
ϵ_s	= Pax_s distribution of the error term with zero mean and unit variance at time s	[passengers]

where, $s=1,\ldots,S$ and S is the length of Pax time series.

Pax within each OD pair follows model of Eq. 2. This paper also applies the ARIMA+GARCH+Bootstrap method developed by Nieto & Carmona-Benitez (2018) to estimate and forecast Pax_s distribution per OD pair allowing to calculate any statistics metrics from Pax_s distribution, specifically any quantile. In the ARIMA+GARCH+Bootstrap methodology, μ_s is estimated with the ARIMA forecasting model, σ_s is estimated using the GARCH(1,1) forecasting model, and the distribution of ϵ_s is approximated using the bootstrap methodology.

Even though model of Eq. 1 and Eq. 2 can be used for forecasting the MS_t at every time t of an OD pair and/or Pax_s at every time s of an OD pair, the estimated values of the parameters for each model are different, i.e., a Bootstrap distribution of MS_t and/or Pax_s at every time t and s is obtained considering the MS_t of an OD pair and Pax characteristics of the OD pair. Measures of central tendency and dispersion, confidence intervals, quantiles, etc. are statistics that can be calculated when the distribution of MS_t and/or Pax_s are known at every time t and s.

The advantage of using Bootstrap methods for the distribution of MS_t and/or Pax_s every time t and s is that no known distribution is assumed which is what other researchers normally do, i.e. Normal, Gamma, Weibull, log-normal have been proposed in Zeni (2001), Swan (2002), and Balcan et al. (2009) respectively. The problem of assuming a distribution is that the evolution of MS_t and/or Pax_s overtime changes which is dynamic and not static. To solve the problem, bootstraps methods capture MS_t and/or Pax_s changes overtime in a dynamic way.

Step 5: calculate the unmet demand of an OD pair,

In this paper, the unmet demand of an OD pair ($UPax_t$) is equal to the difference between the (1-α) quantile of distribution of Pax of the O-D relation, which is the MS of the OD pair ($MS_{1-\alpha,t}$), and the (1-α) quantile of the distribution of Pax of the OD pair ($Pax_{1-\alpha,t}$).

$$UPax_t = MS_{1-\alpha,t} - Pax_{1-\alpha,t} \quad (3)$$

Where:

$UPax_t$	=	unmet demand per route at time t	[passengers]
$MS_{1-\alpha,t}$	=	(1-α) quantile of the distribution of MS of the OD pair at time t	[passengers]
$Pax_{1-\alpha,t}$	=	(1-α) quantile of the distribution of Pax of the OD pair at time t	[passengers]

With models of Eq. 1 and Eq. 2, $MS_{1-\alpha,t}$ and $Pax_{1-\alpha,t}$ are calculated as (1-α) quantile of the distribution of Pax of the O-D relation, which is assumed to be the MS of the OD pair, and for the distribution of Pax of the OD pair at time t. The corresponding quantile is considered as the amount that accumulates (1-α) of the probability, i.e., the probability that the O-D relation Pax would be higher than $MS_{1-\alpha,t}$ is α; and the probability that Pax of the OD pair would be higher than $Pax_{1-\alpha,t}$ is α.

3.1. Approach for identifying new airline flights

In this section with the purpose of demonstrating an applicability of the proposed methodology to calculate *UPax*, we design a qualitative approach to determine whether an OD pair could have enough unmet demand to open new flights. This approach is based on the calculation of the unmet demand. The qualitative approach is as follows: in an OD pair, whether the calculation of $UPax_t$ is positive, it is concluded that there is an opportunity for opening new services at period t. This is because according with the $Pax_{1-\alpha,t}$ forecast, the $MS_{1-\alpha,t}$ is not satisfied. Then, an opportunity for opening new services and fulfil such demand exists, and $UPax_t$ is the maximum number of extra seats to offer by opening new services. In this situation, airlines could be offering less seats, at period t. Here, $MS_{1-\alpha,t}$ indicates that airlines would tend to provide more services than needed, but they are not offering enough, therefore an opportunity to open new services might exist at period t. Contrary, whether $UPax_t$ is negative or zero, it can be concluded that the OD pair would not have opportunities for opening new services at period t, because $Pax_{1-\alpha,t}$ is greater than $MS_{1-\alpha,t}$. In other words, according to the MS model (Eq. 1), when $Pax_{1-\alpha,t}$ is greater or equal than $MS_{1-\alpha,t}$, airlines might be offering more seats than needed at period t. When this happens, the possibility to open new services might not exist.

4. Empirical Application

In this paper, the proposed unmet demand methodology is applied to forecast MS of OD pairs with Eq. 1, forecast Pax of OD pairs with Eq. 2, and forecast the unmet demand of OD pairs with Eq. 3, are set up by analysing The United States domestic air pax transport market from the 1st quarter of 1993 to the 4th quarter of 2015 (Bureau of Transportation Statistics, BTS). Data from the 1st quarter of 1993 to the 4th quarter of 2012 is used for estimation and to calibrate the models of Eq. 1 and Eq. 2. Data from the 1st quarter of 2013 to the 4th quarter of 2015 is used to validate the models.

In step 1, the unmet demand methodology classifies airports into different set of airports using the *k*-means algorithm. For the study case, The United States has 5,080 public airports (Bureau of Transportation Statistics, BTS). The elbow method is applied to find the optimum number of clusters to restrict the *k*-means algorithm. In this application, the elbow method considers two confidence levels 0.95 and 0.99. One hand, a confidence level of 0.95 finds three optimum values k=12, k=13, and k=14. The problem with k=13 and k=14 is the existence of clusters with one airport which restricts the proposed methodology specially when the cluster includes one airport with few connections. Contrary, there are any cluster that include one airport when k=12. Therefore, k=12 is better for the applicability of the proposed methodology. On the other hand, a confidence level of 0.99 always finds the upper limit value of k which must be set for the calculation of the elbow method. The problem of this confidence value is the existence of clusters with one airport. A sensitivity analysis is performed to assure that the results are not dependent on the cluster size. So, we analyse the optimum values obtained with the elbow methods (k=11, k=12 and k=13).

Once the optimum value of k is calculated with the elbow method, the *k*-means algorithm classified them into 11, 12, and 13 airport clusters (A, B, C, D, E, F, G, H, I J, K, L, and M). This algorithm classifies airports based on different socioeconomic factors that determine Pax in a certain market: Distance, AbsMob, RelMob, Income, and Population. Pax determinants are: AltAirport (number of commercial airports inside each CZ); DestCount (number of OD pairs from an origin airport); Yield (Average yields at an airport); LM_Hub (Airports are classified into medium and large hubs using the Federal Aviation Administration (FAA) classification); and AggRouteHHI (airline market concentration at an airport).

As an example, Table 1 shows the cluster centre for each socioeconomic factor of k=12. The factor mean of each cluster indicates the characteristics of the airports that belong to each cluster. For example, the airports of cluster 1 have an average number of 2.71 million passengers per quarter, 69.36 HHI, 0.33 RelMob, 40.65 AbsMob, 6.56 million of population, 181 Destcount, 1.77 Yield, and 1 AltAirp.

Table 1. Cluster centres

Factor	Cluster											
	1	2	3	4	5	6	7	8	9	10	11	12
PAX (000)	2,709.59	1,361.46	1,989.00	212.84	767.23	970.27	202.15	14.48	253.45	760.27	380.97	99.74
HHI	69.36	1,068.79	960.26	58.85	1,208.89	483.75	850.00	833.27	645.43	622.45	465.57	517.99
RelMob	0.33	0.19	-	0.09	0.13	0.09	0.17	0.22	0.20	0.16	0.11	0.26
AbsMob	40.65	27.36	-	17.04	15.37	12.79	26.79	31.07	27.64	19.66	13.63	33.72
Pob (000)	6,558.26	3,206.78	18,788.80	18,788.80	4,985.27	9,678.10	8,393.37	243.44	2,749.45	7,113.74	23,894.16	1,105.32
Destcount	181.00	138.00	134.00	58.00	88.00	100.00	38.00	10.00	67.00	94.00	47.00	50.00
Yield	1.77	3.11	2.53	0.94	2.37	4.15	6.57	5.79	1.20	4.25	4.63	3.14
AltAirp	1.00	2.00	4.00	3.00	1.00	2.00	1.00	1.00	1.00	3.00	4.00	1.00

In step 2, the unmet demand methodology classifies OD pairs into O-D relations or sub clusters. As an example, for the case study k=12, The United States air pax domestic market has 4,384 OD pairs (Bureau of Transportation Statistics, BTS) that are classified into 132 O-D relations or sub clusters (Table 2).

Table 2. O-D relations or sub clusters, k=12

ORI/DES	B	C	D	E	F	G	H	I	J	K	L
A	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
B	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL
C	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL
D	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL
E	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL
F	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL
G	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL

H	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL
I	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL
J	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK	JL
K	KB	KC	KD	KE	KF	KG	KH	KI	KJ	KK	KL
L	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LL

In step 3, the MS is forecasted for two O-D relations (Table 3) using the ARIMA+GARCH+Bootstrap methodology (Eq. 1) for k=11, k=12, and k=13. In step 4, Pax is forecasted for these two OD pairs also with the ARIMA+GARCH+Bootstrap methodology (Eq. 2) for k=11, k=12, and k=13. In step 5, the unmet demand is estimated for these two OD pairs with a probability of 95% of confidence level (Eq. 3) for k=11, k=12, and k=13. Finally, a prediction interval of 95% is estimated per each O-D relation to make sure the decisions of the approach for identifying new airline flights are statistically significant. In this paper, quantile 4 is set up as the decision maker that measures the risk to miss calculate the unmet demand, at this quantile the risk (α) is small, if this happens, a negative unmet demand means it is highly probable that Pax is already attended, and a positive unmet demand means it is highly probable the calculated unmet demand represent an opportunity to open new services. Contrary, it is not recommended to set up quantile 1, 2 and/or 3 as decision maker because the probability to miss calculated the unmet demand is higher than quantile 4, and when this happens, a negative unmet demand could mean no opportunity to open new services might exist when actually a high probability of opening new airline services might exist, and a positive unmet demand could mean an opportunity of open new services might exist when actually it is highly probable an opportunity to open new services might not exist.

All the calculations of the unmet demand for all OD pairs in the US market cannot be shown because, for example, they are 4,384 OD pairs for k=12, reason why, in this section, two OD pairs and two O-D relations that have been randomly selected are shown as example.

Table 3. Selected OD pairs for analysis

No.	k	Sub cluster	OD pair	
			Airport Origin	Airport Destination
1	11	JF	Los Angeles (LAX)	John F. Kenedy (JFK)
2	11	EK	Portland (PDX)	Boston (BOS)
3	12	CK	Los Angeles (LAX)	John F. Kenedy (JFK)
4	12	IG	Portland (PDX)	Boston (BOS)
5	13	FF	Los Angeles (LAX)	John F. Kenedy (JFK)
6	13	IF	Portland (PDX)	Boston (BOS)

Table 4, Table 5 and Table 6 show the results for the two OD pairs under analysis (LAX – JFK and PDX – BOS) for k=11, k =12, and k =13 respectively. The *MS_lb* is the *MS* prediction interval lower bound, the *MS_up* is the *MS* prediction interval upper bound, the *Upax_lb* is the *Upax* prediction interval lower bound, the *Upax_up* is the *Upax* prediction interval upper bound.

In Table 4, Table 5 and Table 6 the unmet demand prediction intervals are all negative in the LAX – JFK OD pair. These results mean the Pax forecast of the OD pairs (*Pax*) are greater than the prediction interval of the O-D relation (*MS*) with a probability of 95% confidence level. It means the risk that *MS* is greater than *Pax* is 5%. The statistical results allow to conclude that these OD pairs have 95% probability of being highly congested according with the OD pairs (*Pax*) and O-D relations (*MS*) distribution functions, meaning that an opportunity to open new services might not exist.

In Table 4, Table 5 and Table 6 the unmet demand (*UPax*) prediction intervals are all negative in the LAX – JFK OD pair. These results mean the *Pax* forecast of the OD pairs are greater than the prediction interval of the *MS* with a probability of 95% confidence level. It means the risk that *MS* is greater than *Pax* is 5%. The statistical results allow to conclude that these OD pairs have 95% probability of being highly congested according with the *Pax* and *MS* distribution functions, meaning that an opportunity to open new services might not exist.

Table 4. O-D relations, OD pairs and unmet demand forecasts for k=11 [thousands].

Route	LAX-JFK	PDX-BOS

Year	Q	*MS_lb*	*MS*	*MS_ub*	*Pax*	*Upax_lb*	*Upax*	*Upax_ub*	*MS_lb*	*MS*	*MS_ub*	*Pax*	*Upax_lb*	*Upax*	*Upax_ub*
2013	1	4.61	4.62	4.63	35.95	- 31.33	- 31.33	- 31.31	2.86	2.89	2.93	30.55	- 27.70	- 27.67	- 27.62
2013	2	6.90	6.92	6.95	31.26	- 24.36	- 24.34	- 24.31	4.22	4.23	4.25	1.21	3.01	3.03	3.04
2013	3	4.89	4.90	4.92	36.65	- 31.76	- 31.75	- 31.73	3.93	3.95	3.97	6.68	- 2.75	- 2.73	- 2.71
2013	4	- 9.34	- 7.90	- 6.40	29.35	- 38.69	- 37.26	- 35.75	3.70	3.71	3.74	8.22	- 4.53	- 4.52	- 4.49
2014	1	17.29	17.65	17.98	36.42	- 19.14	- 18.78	- 18.45	3.10	3.12	3.15	1.49	1.62	1.63	1.66
2014	2	5.46	5.48	5.49	34.80	- 29.33	- 29.32	- 29.31	4.42	4.43	4.43	24.34	- 19.91	- 19.91	- 19.91
2014	3	26.40	26.76	27.26	32.53	- 6.13	- 5.77	- 5.27	4.93	4.93	4.93	56.10	- 51.18	- 51.18	- 51.18
2014	4	5.80	5.81	5.83	34.05	- 28.25	- 28.24	- 28.22	6.56	6.59	6.74	0.97	5.59	5.62	5.78
2015	1	5.23	5.25	5.27	34.67	- 29.43	- 29.42	- 29.39	4.28	4.28	4.28	4.58	- 0.30	- 0.30	- 0.30
2015	2	6.39	6.40	6.40	37.77	- 31.38	- 31.38	- 31.37	4.82	4.83	4.85	0.75	4.08	4.09	4.10
2015	3	4.51	4.52	4.54	42.41	- 37.90	- 37.89	- 37.88	6.25	6.26	6.27	3.24	3.01	3.02	3.03
2015	4	5.32	5.32	5.32	40.82	- 35.50	- 35.50	- 35.50	6.62	6.63	6.63	28.73	- 22.11	- 22.10	- 22.10

In Table 4, Table 5 and Table 6 the *Upax* prediction intervals are positive for some quarters of year in the PDX-BOS OD pair. These results mean that *Pax* forecast of the OD pairs are greater than the prediction interval of the *MS* with a probability of 95% confidence level. It is possible to conclude that this OD pair is not congested in some quarters of year according with its *MS*, this condition could be explained by the air pax demand seasonality.

Table 5. O-D relations, OD pairs and unmet demand forecasts for k=12 [thousands].

Route		LAX-JFK							PDX-BOS						
Year	Q	*MS_lb*	*MS*	*MS_ub*	*Pax*	*Upax_lb*	*Upax*	*Upax_ub*	*MS_lb*	*MS*	*MS_ub*	*Pax*	*Upax_lb*	*Upax*	*Upax_ub*
2013	1	7.43	7.43	7.43	35.95	- 28.52	-28.52	- 28.52	1.03	1.03	1.04	30.55	- 29.53	-29.52	- 29.52
2013	2	9.99	9.99	9.99	31.26	- 21.27	-21.27	- 21.27	1.73	1.74	1.74	1.21	0.53	0.53	0.54
2013	3	15.57	15.57	15.57	36.65	- 21.09	-21.09	- 21.09	2.20	2.21	2.22	6.68	- 4.48	- 4.47	- 4.46
2013	4	9.17	9.17	9.17	29.35	- 20.18	-20.18	- 20.18	1.10	1.11	1.12	8.22	- 7.13	- 7.12	- 7.10
2014	1	15.71	15.71	15.71	36.42	- 20.71	-20.71	- 20.71	1.08	1.08	1.08	1.49	- 0.41	- 0.41	- 0.40
2014	2	24.40	24.40	24.40	34.80	- 10.40	-10.40	- 10.40	1.93	1.93	1.93	24.34	- 22.41	-22.40	- 22.40
2014	3	15.00	15.00	15.00	32.53	- 17.53	-17.53	- 17.53	2.18	2.19	2.21	56.10	- 53.92	-53.91	- 53.90
2014	4	26.46	26.46	26.46	34.05	- 7.59	- 7.59	- 7.59	1.57	1.57	1.57	0.97	0.60	0.60	0.60
2015	1	9.88	9.88	9.88	34.67	- 24.78	-24.78	- 24.78	0.97	0.97	0.97	4.58	- 3.61	- 3.61	- 3.61
2015	2	14.58	14.58	14.58	37.77	- 23.19	-23.19	- 23.19	1.57	1.58	1.59	0.75	0.83	0.83	0.84
2015	3	15.35	15.35	15.35	42.41	- 27.07	-27.07	- 27.07	1.90	1.90	1.90	3.24	- 1.34	- 1.34	- 1.34
2015	4	13.55	13.55	13.55	40.82	- 27.27	-27.27	- 27.27	1.65	1.65	1.65	28.73	- 27.08	-27.08	- 27.08

Table 6. O-D relations, OD pairs and unmet demand forecasts for k=13 [thousands].

Route		LAX-JFK							PDX-BOS						
Year	Q	*MS_lb*	*MS*	*MS_ub*	*Pax*	*Upax_lb*	*Upax*	*Upax_ub*	*MS_lb*	*MS*	*MS_ub*	*Pax*	*Upax_lb*	*Upax*	*Upax_ub*
2013	1	1.06	1.06	1.06	35.95	- 34.88	-34.88	- 34.88	1.44	1.46	1.47	30.55	- 29.11	-29.10	- 29.08
2013	2	1.06	1.06	1.06	31.26	- 30.20	-30.20	- 30.20	2.12	2.13	2.15	1.21	0.92	0.92	0.94
2013	3	0.81	0.81	0.81	36.65	- 35.85	-35.85	- 35.85	2.82	2.87	2.92	6.68	- 3.86	- 3.81	- 3.76
2013	4	1.19	1.19	1.19	29.35	- 28.16	-28.16	- 28.16	2.14	2.15	2.17	8.22	- 6.09	- 6.07	- 6.06
2014	1	0.82	0.82	0.82	36.42	- 35.60	-35.60	- 35.60	1.25	1.26	1.27	1.49	- 0.24	- 0.23	- 0.22

2014	2	1.06	1.06	1.06	34.80	-	33.74	-33.74	-	33.74	3.12	3.18	3.23	24.34	-	21.21	-21.16	-	21.11
2014	3	0.96	0.96	0.96	32.53	-	31.57	-31.57	-	31.57	0.91	0.92	0.94	56.10	-	55.20	-55.18	-	55.16
2014	4	1.13	1.13	1.13	34.05	-	32.91	-32.91	-	32.91	2.17	2.19	2.21	0.97		1.21	1.22		1.24
2015	1	0.81	0.81	0.81	34.67	-	33.86	-33.86	-	33.86	2.32	2.33	2.36	4.58	-	2.26	- 2.25	-	2.22
2015	2	0.81	0.81	0.81	37.77	-	36.96	-36.96	-	36.96	1.58	1.58	1.58	0.75		0.83	0.84		0.84
2015	3	1.01	1.01	1.01	42.41	-	41.40	-41.40	-	41.40	1.67	1.68	1.69	3.24	-	1.57	- 1.56	-	1.55
2015	4	0.87	0.87	0.87	40.82	-	39.95	-39.95	-	39.95	1.63	1.64	1.65	28.73	-	27.10	-27.09	-	27.07

It is important to mention that in this analysis, we are considering 95% confidence level for both forecasts, however, it is possible to consider any percentage of confidence level, and therefore, the quantile difference, which is the *Upax* could be different, but this decision must be taken by an airline or airport manager with enough expertise in calculating *Pax* at OD pair level.

The sensitivity analysis shown in Table 4, Table 5 and Table 6 that the number of clusters (*k*) do not change the sign (positive or negative) of *Upax*. It means, the Pax at OD pair level presents equal opportunities to open new service regardless the number of clusters (*k*). Thus, it is possible to confirm that the opportunity to open new routes is not sensible to the number of clusters obtained applying the elbow method in step 1 of the proposed methodology.

In this paper, the root mean square error (RMSE) is calculated to measure the goodness of fit for the forecasts of *Pax* and *MS* per number of cluster *k*. The RMSE is a statistical metric that explains the forecast error by comparing real *Pax* and *MS* data against the proposed methodology calculations performed with the ARIMA+GARCH+Bootstrap forecasting method. The RMSE analyses indicate that *k*=13 shows the most accurate performance for the case of LAX-JFK, and that *k*=12 shows the most accurate performance for the case of PDX-BOS.

Table 7. The root mean square errors (RMSE) of *MS* and *Pax* per number of clusters *k*

	LAX-JFK	PDX-BOS
MS k=11	8,161.08	681.68
MS k=12	5,220.26	282.67
MS k=13	145.55	706.74
Pax	4,575.26	20,456.92

6. Conclusion

The forecast of the *UPax* is important for strategic decisions of new planning such as opening new OD pairs or airline routes, increasing/decreasing the number of services in an OD pair, aircraft route choice, etc. A challenge to airlines and airports when forecasting *Pax* is the fact that they mainly gather the actual number of bookings and not the *MS* willing to fly between an origin airport and a destination airport. This happens because customers could not book might be due to capacity constraints or booking limits and/or simple because the OD pair has not been operated by any airline yet.

The main problem of forecasting the *UPax* is the fact that *Pax* must be unconstrained to calculate *UPax*. It is because the *MS* of OD pairs must be extrapolated from censored booking data since it is known that unconstrained *Pax* data calculates better forecasts and improve revenues gains. Many models have been developed to forecast *Pax* based on historical data with the goal of finding balance between supply and demand which is a very complex problem to tackled. The first contribution of this paper is the development of a statical method to unconstraint or detruncate *Pax* at OD pair level using socioeconomic factors that have been proved to be significant variables to segment *Pax*. This methodology mixes time-series methods with the bootstrap distribution function and machine learning algorithms.

The *UPax* cannot be known but, in this paper, we develop a five-step methodology to calculate it as a second contribution. The first step uses a machine learning clustering method to create sets of airports according with socioeconomic mobility variables and *Pax* determinants. The segmentation using the SEM variables is another

contribution because other studies in literature use other variables mostly at country or state level, and SEM factors are reported at community zone level, which is the lowest desegregation of data, very important in our study, because SEM variables allow us to study *Pax* at airport level, otherwise data would not be consistent, and our study could not be done. For example, if the FAA airport classification would be used, the resulting number of clusters are four and the O-D relations or sub clusters would be sixteen. With the FAA airport classification, we would not efficiently discriminate among airports. So, the distribution functions of *Pax* and *MS* would include aggregated information about airports that are located in community zones that are different in socioeconomic and air transportation factors. However, it is possible to cluster airports according with different variables if required, but at community zone level and/or OD pair level. For example, it would be interesting to include fares per OD pair, unfortunately, we do not have access to these data, but the forecast distribution considers it because time-series models consider such info.

In the second step, we propose to create subsets of airports according with the cluster's classification in Step 1. This step recognizes that a level of relationship exists due to the airports socioeconomical relations that are considered in their cluster classification. The results of this paper show that the O-D relations or sub clusters have an important influence in the necessity of enplaned passengers to fly between airports. Step 2 is a small contribution of our methodology because we are proposing to create clusters and sub clusters in this way to be able to interpolate the median to airports that do not have certain connections.

In the third and fourth steps, we propose to use the ARIMA+GARCH+Bootstrap time series method to estimate the *MS* and *Pax* because, so far, this forecasting method has been proved to be the most accurate method in the airline pax industry. However, the methodology can be adapted to use any forecasting method. But the ARIMA+GARCH+Bootstrap has been validated to model the behaviour of *Pax*. Moreover, this time series method simulates the real distribution function giving the possibility to estimate *Pax* without the need of knowing the parameters that have an influence on it, because the distribution function considers all variables that have an influence and are determinants of *Pax* such as: gross domestic product, population, average ticket fares, distance, number of seats supplied, aircraft average seat, etc. Hence, the main advantage of the ARIMA+ GARCH+Bootstrap is that it does not need to know or to forecast the determinants of *Pax* to have the capacity to estimate it. A second advantage is that this model do not change estimation when other variables change. Therefore, it is an easy model to use when historical data are available.

In Step 5, we propose to calculate the *UPax* as the difference between the *MS* and *Pax* forecasts quantiles, and it provides a confidence level of (1-α)% which can be interpreted as a measure of risk as it is commonly used in risk management. The value of α is a decision of airlines and airports managers as it is done in any business when a forecasting model is applied. In this paper, the case study is performed with 95% confidence level. The *UPax* allows to determine whether the possibility of opening new services or offering more seats in an OD pair might exist according with the behaviour of the historical airline pax data. Although, the *UPax* is unknown because such data do not exist at all, the proposed methodology uses the quantile of the forecast distribution function of the data per O-D relation or sub clusters and per OD pair, so the methodology considers the characteristics of *Pax* at OD pair level. Hence, the calculation of the *UPax* is supported by the statistical theory and analysis. It can be concluded that the proposed methodology allows to simulate the *UPax* data which is not known. Besides, one way to assess the predictions provided by the proposed methodology is an estimation of the *UPax* calculated by simulating the *MS* data behaviour. This is done by describing the distribution of the *MS* historical data, because distribution functions give the possibility to calculate any statistics of *Pax* without knowing the parameters that have an influence on it.

The *UPax* methodology can calculate negative values because of the α percentile that is set up. In this paper, quantile 4 is set up as the decision maker that measures the risk to miss calculate the *UPax*, at this quantile the risk (α) is small, if this happens, a negative *UPax* means it is highly probable that *Pax* is already attended, and a positive *UPax* means is highly probable the calculated *UPax* represent an opportunity to open new services. Contrary, if quantile 1 is set up as the decision maker the probability to miss calculated the *UPax* is high, if this happens, a negative *UPax* means no opportunity to open new services exist when actually a high probability of opening new airline services do exist, and a positive *UPax* means an opportunity of open new services exist when actually it is highly probable an opportunity to open new services do not exist. In this paper, prediction intervals are performed for *MS* to avoid false positives and false negatives.

The calculations of the *UPax* for LAX-JFK indicate that this OD pair connection is under extreme competition, because the *UPax* forecasts are negative for all quarters of year. Contrary, the *UPax* for PDX-BOS indicate that this OD pair connection is not under extreme competition for some quarters of year. It is because thus connection

is between a small airport and a primary airport, both located in community zones in big cities. Thus, The *UPax* is positive in some quarters of year in this connection, meaning that socioeconomic factors between airport community zone locations change according with the seasonality of year.

The third contribution of this paper is an approach for identifying new airline flights to demonstrate the applicability of the proposed methodology that calculates *UPax*. This paper presents the application of the methodology to only two routes chosen aleatory. But it is important to mention that the US air pax industry has been studied as a whole. The results demonstrate that airports and airlines could use the proposed methodology to calculate *Pax* that is not satisfied to decide what routes to open and encourage. Airlines could decide what aircrafts, number of seats and frequencies to operate and cover the *UPax*. Governments and private companies can use the proposed methodology to study the *MS*, i.e. the results of the study case indicate that for the United States air pax industry, the hubs are under extreme competition, we did not find OD pairs where airlines could open new services or increase seats with low risk, it means, they can try, but as the calculated *UPax* are negative, high risks with high competition airlines will find if they tried to get in hubs OD pairs. Contrary, small airports, and primary airports located in big cities can open new services between them and some of them with hub airports. The results show that new frequencies could be open depending on the seasonality of *Pax*, meaning that socioeconomic factors between airport locations change according with the quarter of years.

Even though, the SEM variables used in this paper are recently proved to have an influence in the air pax industry by Hofer, Kali & Mendez (2018), one future work is to update socioeconomic variables and *Pax* data because the results of the application of the proposed methodologies depend on the variables used to cluster. A second future work is to modify the elbow method to find the optimum value of *k*, because if the *k*-means algorithm creates many clusters, the methodology cannot interpolate the median in many airports, and if the *k*-means algorithm creates few clusters, the proposed model cannot discriminate between airports, so *MS* would be similar in all OD pairs. In other words, the goal of this future work is to determine the minimum number of airports to set in a cluster to be statistically significant. Finally, a third future work is to study whether it is possible and convenient to apply an oversampling technique such as the Synthetic Minority Over-sampling Technique (SMOTE) to generate extra data to overcome the shortage of data and develop a one-step clustering approach rather than a two-steps clustering approach. This would allow to add relevant OD pair specific factors which are more disaggregated information at OD pairs level such as: airline business models, airport business models, number of airlines, number of frequencies, and airlines market share to assess levels of competition. According with literature review, these variables have been used to estimate latent demand in air transport management.

References

Alamdari, F., Mason, K. (2006). The future of airline distribution. Journal of Air Transport Management 12 (3), 122–134. DOI: 10.1016/j.jairtraman.2005.11.005

Andereck, K. L., Caldwell, L. L. (1994). Variable selection in tourism market segmentation models. Journal of Travel Research, 33(2), 40-46. DOI: 10.1177/004728759403300207

Anderson, J. E., Kraus, M. (1981). Quality of service and the demand for air travel. The Review of Economics and Statistics, 63 (4), 533-540. DOI: 10.2307/1935849

Balcan, D., Colizza, V., Goncalves, B., Hu, H., Ramasco, J. J., Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences Dec 2009, 106 (51) 21484-21489; DOI: 10.1073/pnas.0906910106.

Birolini, S., Antunes, A. P., Cattaneo, M., Malighetti, P., Paleari, S. (2021). Integrated flight scheduling and fleet assignment with improved supply-demand interactions. Transportation Research Part B: Methodological, 149, 162-180. https://doi.org/10.1016/j.trb.2021.05.001

Bureau of Transportation Statistics (2017). Retrieve from https://www.transtats.bts.gov/airports.asp accessed on 23/October/2017

Camilleri M.A. (2018). Market Segmentation, Targeting and Positioning. In: Travel Marketing, Tourism Economics and the Airline Product. Tourism, Hospitality & Event Management. Springer, Cham. DOI: 10.1007/978-3-319-49849-2_4

Carmona-Benítez, R.B., 2012. The Design of a Large Scale Airline Network. PhD Dissertation. Delft University of Technology, The Netherlands. TRAIL Research School, (Chapter 5). ISBN: 9789055841547

Carmona-Benítez, R.B., Carmona-Paredes, R.B., Lodewijks, G., Nabais, J.L. (2013). Damp trend Grey Model forecasting method for airline industry. International Journal of Expert Systems with Applications 40 (12), 4915–4921. DOI: 10.1016/j.eswa.2013.02.014.

Carmona Benitez, R.B., Nieto, M.R. (2020). SARIMA damp trend grey forecasting model for airline industry. Journal of Air Transport Management, 82, 101736. DOI: 10.1016/j.jairtraman.2019.101736

Chandukala, S. R., Edwards, Y. D., Allenby, G. M. (2011). Identifying Unmet Demand, Marketing Science, 30 (1), 61-73. DOI: 10.1287/mksc.1100.0589

Chetty, R., Hendren, N., Kline, P., Saez, E. (2014a). Where is the land of opportunity? The geography of intergenerational mobility in the United States. The Quarterly Journal of Economics, 129 (4), 1553–1623. DOI: 10.1093/qje/qju022

Chetty, R., Hendren, N., Kline, P., Saez, E., Turner, N. (2014b). Is the United States still a land of opportunity? Recent trends in intergenerational mobility. American Economic Review, 104 (5), 141–147. DOI: 10.1257/aer.104.5.141

Chiang, W.Y. (2017). Discovering customer value for marketing systems: an empirical case study. International Journal of Production Research, 55 (17), 5157-5167. DOI: 10.1080/00207543.2016.1231429

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39 (1), 1–38.

Freathy, P., O'Connell, F. (2000). Market segmentation in the European airport sector. Marketing Intelligence & Planning, 18 (3), 102–111. DOI: 10.1108/02634500010327890

Guillen, D., Lall, A. (2018). Commoditization and segmentation of aviation markets. Transportation Policy and Economic Regulation. Essays in Honor of Theodore Keeler, 53-75. DOI: 10.1016/B978-0-12-812620-2.00003-1

Guo, P., Xiao, B., Li, J. (2012). Unconstraining Methods in Revenue Management Systems: Research Overview and Prospects. Advances in Operations Research, 1-23 https://doi.org/10.1155/2012/270910

Harrison, A., Popovic, V., Kraal, B. (2015). A new model for airport passenger segmentation. Journal of Vacation Marketing, 21 (3), 237-250. DOI: 10.1177/1356766715571390

Haensel, A., Koole, G. (2011). Estimating unconstrained demand rate functions using customer choice sets. Journal of Choice Modelling, 4 (3), 75-87. https://doi.org/10.1016/S1755-5345(13)70043-5

He, D.Y., Luo, L. (2006). An improved winters model for airline demand forecast. Journal of Transportation Systems Engineering and Information Technology, 6 (6), 103–107.

Heung, V. C., Kucukusta, D., Song, H. (2011). Medical tourism development in Hong Kong: An assessment of the barriers. Tourism Management, 32 (5), 995-1005. DOI: 10.1016/j.tourman.2010.08.012

Hofer, C., Kali, R., Mendez, F. (2018). Socio-economic mobility and air passenger demand in the U.S. Transport Research part A: Policy and Practice, 112, 85-94. DOI: 10.1016/j.tra.2018.01.009

Jang, S.C., Morrison, A.M., O'Leary, J.T. (2002). Benefit segmentation of Japanese pleasure travelers to the USA and Canada: Selecting target markets based on the profitability and risk of individual market segments. Tourism Management, 23 (4), 367-378. DOI: 10.1016/S0261-5177(01)00096-6

Karmarkar, S., Goutam, D., Tathagata, B. (2011). Revenue impacts of demand unconstraining and accounting for dependency. Journal of Revenue and Pricing Management, 10, 367–381. https://doi.org/10.1057/rpm.2009.54

Leung, A., Yen, B.T.H., Lohmann, G. (2017). Why passengers' geo-demographic characteristics matter to airport marketing. Journal of Travel & Tourism Marketing, 34 (6), 833-850. DOI: 10.1080/10548408.2016.1250698

Little, R.J.A., Rubin, D.B. (1987). Statistical Analysis with Missing Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA.

Marcos, R., García-Cantú, O., Herranz, R. (2018). A Machine Learning Approach to Air Traffic Route Choice Modelling. DOI:10.48550/arXiv.1802.06588

McLachlan, G.J., Krishnan, T. (1997). The EM Algorithm and Extensions. Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA

Müller, H., Hamm, U. (2014). Stability of market segmentation with cluster analysise - A methodological approach. Food Quality and Preference, 34, 70-78. DOI: 10.1016/j.foodqual.2013.12.004

Neufville, R. (2017). Airport systems planning and design. In Budd, L. & Ison, S. (Eds.), Air Transport Management - An International Perspective (pp. 64). New York: Routledge Taylor and Francis Group.

Nieto, M.R., Carmona-Benítez, R.B. (2018). ARIMA + GARCH + Bootstrap forecasting method applied to the airline industry. Journal of Air Transport Management. 71, 1–8. Doi: 10.1016/j.jairtraman.2018.05.007.

Orkin, E.B. (1998). Wishful thinking and rocket science: the essential matter of calculating unconstrained demand for revenue management. The Cornell Hotel and Restaurant Administration Quarterly, 39 (4), 15–19. https://doi.org/10.1016/S0010-8804(98)80027-X

Park, S.Y., Pan, B. (2018). Identifying the next non-stop flying market with a big data approach. Tourism Management, 66, 411-421. DOI: 10.1016/j.tourman.2017.12.008

Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill Interamericana, Spain. ISBN: 8448136101.

Polt, S. (2000). From bookings to demand: the process of unconstraining: in Proceedings of the AGIFORS Reservations and Yield Management Study Group, New York, NY, USA.

Queenan, C.C., Ferguson, M., Higbie, J., Kapoor, R. (2007). A comparison of unconstraining methods to improve revenue management systems. Production and Operations Management, 16 (6), 729–746. https://doi.org/10.1111/j.1937-5956.2007.tb00292.x

Saleh, R. (1997). Estimating lost demand with imperfect availability indicators: in Proceedings of the AGIFORS Reservations and Yield Management Study Group, Montreal, Canada.

Swan, W. M. (2002). Airline route developments: a review of history. Journal of Transport Management, 8, 349-353. DOI: 10.1016/S0969-6997(02)00015-7.

Tkaczynski, A., Rundle-Thiele, S., Beaumont, N. (2009). Segmentation: A tourism stakeholder view. Journal of Travel Research, 30 (2), 169-175. DOI: 10.1016/j.tourman.2008.05.010

Teichert, T., Shehu, E., Wartburg, I.V. (2008). Customer Segmentation Revisited: The Case of the Airline Industry. Transportation Research Part A: Policy and Practice, 42 (1), 227–242. DOI: 10.1016/j.tra.2007.08.003

Toh, R.S., Hu, M.Y. (1990). A Multiple Discriminant Approach to Identifying Frequent Fliers in Airline Travel: Some Implications for Market Segmentation, Target Marketing, and Product Differentiation. Logistics and Transportation Review, 26 (2), 179-197.

Solvoll, G., Mathisen, T.A., Welde, M. (2020). Forecasting air traffic demand for major infrastructure changes. Reach in Transportation Economics, 82, 1-8. DOI: 10.1016/j.retrec.2020.100873

Umargono, E., Suseno, J.E., Gunawan, S.K.V. (2020). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019). Atlantis Press. ISBN: 9789462391680.

Urban, M., Klemm, M., Ploetner, K.O., Hornung, M. (2018). Airline categorisation by applying the business model canvas and clustering algorithms. Journal of Air Transport Management, 71, 175-192. DOI: 10.1016/j.jairtraman.2018.04.005

Vermunt, J.K., Magidson, J. (2003). Latent class models for classification. Computational Statistics & Data Analysis, 41 (3-4), 531–537. DOI: 10.1016/S0167-9473(02)00179-2

Weatherford, L., Pölt, S. Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. Journal of Revenue and Pricing Management, 1 (3), 234–254 (2002). https://doi.org/10.1057/palgrave.rpm.5170027

Wickham R.R. (1995). Evaluation of forecasting techniques for short-term demand of air transportation [M.S. thesis], Massachusetts Institute of Technology, Cambridge, Mass, USA.

Zeni, R. H. (2001). Improved forecast accuracy in revenue management by unconstraining demand estimates from censored data. Published by Dissertation.com, PhD in management program, United States of America. ISBN: 9781581121414