



Visualizing the historical COVID-19 shock in the US airline industry: A Data Mining approach for dynamic market surveillance

Darío Pérez-Campuzano^{a,b,*}, Luis Rubio Andrada^a, Patricio Morcillo Ortega^a, Antonio López-Lázaro^{c,d}

^a Universidad Autónoma de Madrid (UAM), Facultad de Ciencias Económicas y Empresariales, Calle Francisco Tomás y Valiente N5, 29049, Madrid, Spain

^b LLM Aviation, Paseo de la Habana N26, 28036, Madrid, Spain

^c Universidad Politécnica de Madrid (UPM), Escuela Técnica Superior de Ingeniería Aeronáutica y del Espacio, Plaza del Cardenal Cisneros N3, 28040, Madrid, Spain

^d Euroairlines, Paseo de la Habana N26, 28036, Madrid, Spain

ARTICLE INFO

Keywords:

Airlines
COVID-19
Data mining (DM)
Unsupervised learning
Self-organizing map (SOM)
K-means

ABSTRACT

One of the purposes of Artificial Intelligence tools is to ease the analysis of large amounts of data. In order to support the strategic decision-making process of the airlines, this paper proposes a Data Mining approach (focused on visualization) with the objective of extracting market knowledge from any database of industry players or competitors. The method combines two clustering techniques (Self-Organizing Maps, SOMs, and K-means) via unsupervised learning with promising dynamic applications in different sectors. As a case study, 30-year data from 18 diverse US passenger airlines is used to showcase the capabilities of this tool including the identification and assessment of market trends, M&A events or the COVID-19 consequences.

1. Introduction

From the time of the first digital data compilations around 1960, the amount of collected data has never stopped growing. During the last decades, database management systems have evolved until reaching advanced levels in terms of database size management, cloud storage capabilities or integration of sources, to name but a few. Nevertheless, these burgeoning amounts of data have become unmanageable and frequently the information or knowledge hidden inside them is not discovered or even tried to be disclosed due to the size and complexity of those databases. This situation has given rise to the well-known proverb *data rich but information poor*.

This is the reason behind the born of Data Mining (DM) which has become the favorite denomination to refer to other similar concepts such as Knowledge Discovery in Databases (KDD), knowledge mining from data, knowledge extraction, data or pattern analysis, data archaeology or data dredging. According to (Han et al., 2012), this term comprises advanced or intelligent techniques to analyze large data repositories (such as databases, data warehouses or web storages) in order to recover the most valuable knowledge from them. In the previous definition,

knowledge may refer to implicit, previously unknown information; relationships between the original data or simply summaries of it. In any case, this process is expected to add understanding and wisdom to the data owner by easing its assessment and visualization or representation. It is a creative process whose implementation widely depends on aspects such as eventual objective (summarizing, clustering, visualization, etc.) and data traits (structure, size, type, amount, noise, etc.).

As explained in (Zhou, 2003), DM has a great relationship with several sciences and gathers technologies from different branches such as database management, Machine Learning (ML), Statistics, Data Visualization or Artificial Intelligence (AI). In addition, it can be used for a great variety of applications such as characterization, prediction, clustering or surveillance (Campos and Rubio, 2017).

At the beginning of the past decade, DM and Big Data had already become the top corporate priority for most of the aviation companies (GE and Accenture, 2014) and it still represents one of the most promising AI technologies for air transport. Furthermore, the growing amount of available data of different nature in the industry (finance, operations, flight, etc.) is allowing corporations to find new ways of improving their overall performance leveraging on that information. As it happens in

* Corresponding author. Universidad Autónoma de Madrid (UAM), Facultad de Ciencias Económicas y Empresariales, Calle Francisco Tomás y Valiente N5, 29049, Madrid, Spain.

E-mail addresses: dario.perez@estudiante.uam.es (D. Pérez-Campuzano), luis.rubio@uam.es (L. Rubio Andrada), patricio.morcillo@uam.es (P. Morcillo Ortega), alopezlazar@euroairlines.es (A. López-Lázaro).

<https://doi.org/10.1016/j.jairtraman.2022.102194>

Received 22 June 2021; Received in revised form 21 February 2022; Accepted 21 February 2022

Available online 28 February 2022

0969-6997/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

other sectors, this kind of technology is starting to deliver some added value but it still shows great room for improvement and its fast theoretical evolution has created a gap between academic research and business application. As (Akerkar, 2014) concludes, aviation companies (including airlines and airports to name but two) will be forced to develop strong capabilities regarding data acquisition and processing with the aim of improving its service offering and efficiency as well as maintaining its competitiveness.

On the other hand, AI shows great potential for application in management departments of transport organizations as the historical meta-analysis of (Pérez-Campuzano et al., 2021b) suggests. In particular, airlines may capitalize on this opportunity since the COVID-19 crisis has further stressed their sector where firms had historically struggled to sustain profitability. Their strategy and finance could be enhanced by the application of this kind of tools as a way to complement other potential initiatives for the recovery (Dube et al., 2021). Especially, algorithms based on unsupervised learning (strongly related to DM) are envisaged as key potential facilitators for strategic market analysis in a broad sense (trends, rivalry, anomalies, etc.) even considering the impact of the COVID-19 pandemic in the business (Maneenop and Kotcharin, 2020).

In order to support the strategic decision-making process of the airlines, this paper proposes a DM approach (focused on visualization) with the objective of extracting market knowledge from any database (e.g. financial or operational) of industry players or competitors. Particularly, after a review of the state-of-the-art (section 2), a method that combines two clustering techniques (Self-Organizing Maps and K-means) is suggested for potential use in different sectors and with several promising dynamic applications (section 3). As a case study (section 4), after a preprocessing phase, 30-year data from 18 United States (US) passenger airlines are used to showcase the capabilities of this tool, which allows to draw general conclusions regarding the evolution of market trends and players, as well as particular observations in relation to relevant events, especially the COVID-19 consequences.

2. Literature review

Firstly, some state-of-the-art reviews are listed and briefly described in the following paragraphs, starting with and focusing on those more closely related to DM and database management techniques as well as to the air transport sector.

A review of aviation-related works from a DM perspective is gathered in (Akpınar and Karabacak, 2017). There, eight different topics are addressed including *airlines* and their *marketing*. As the authors state, DM implementation in the sector is still at an early stage and there is big room for innovation and improvement regarding the use of Business Intelligence (BI) tools in order to obtain more managerially interpretable knowledge. Similar conclusions are also drawn from the analysis carried out in (Akerkar, 2014).

Another review of the state of the art of BI applied to airports and airlines is found in (Zhang et al., 2011). In this case, the focus is not only on the DM techniques but also on data warehousing, which relates to the collection of information and how it is organized in a suitable manner. Since data is becoming increasingly available in the industry, authors conclude that BI will help decision-makers in making their actions more effective, reasonable, scientific and quick.

A holistic approach to Big Data-based analytics in the aviation sector is carried out in (Larsen, 2013): (i) databases are cited and briefly described, (ii) a generalist methodology for implementation is presented and (iii) some examples of application are shown. One of the outcomes of the work is that working with large aviation databases requires some effort for a reliable integration. This is caused by the lack of standardization, uniformity or defect controls in aviation data sources and the low homogeneity between publishers.

More recently, two literature reviews have been published regarding the application of general AI tools within the air transport sector:

(Beecroft, 2019) addresses the current situation and future challenges for security in the public transport and (Chung et al., 2020) focuses on aspects such as air transport network management or forecasting. It is worth also mentioning two recent books on AI application cases: (Shmelova et al., 2019) and (Shmelova et al., 2020).

Secondly, some papers which have developed specific methods leveraging DM tools within the air transport industry are described hereunder.

One of the first studies where DM is applied to aviation data is (Nazeri and Zhang, 2002). There, a relevant role is given to the data preparation stage. Insights regarding the influence of weather on airspace system performance are obtained. This is achieved by means of using classification, regression and clustering techniques and generating classification rules.

Another work that focuses on the preprocessing stage, of airline data in this case, is (Gürbüz et al., 2011). Several tools (regression, anomaly detection or classification) are applied in order to reduce the size of the original database. The traditional airline business model classification between Low Cost Carrier (LCC) and Full Service Network Carrier (FSNC) is questioned in (Urban et al., 2018). There, 42 operating airlines are clustered through the analysis of their current business model canvas (customer, value proposition, etc.) in search for new or arising standardized model groups or trends.

The competitiveness of 5 airlines operating 2 Taiwanese routes is assessed in (Wen and Chen, 2011) via Multiple Correspondence Cluster Analysis (MCCA). In addition to this technique, perceptual maps and clustering are applied to 21 service attributes gathered through 647 passenger surveys. This allows for an easy visualization of the results which indicate that two clusters of airlines (with common underlying strategies) are found in the market.

A similar visual approach is carried out in (Wen et al., 2014), where the factor-analytic choice map is used to illustrate the passenger perception of 5 different Asian airlines. The analysis of the results provides behavioral and managerial suggestions not only for airline managers but also for policy-makers.

Clustering techniques are also used with grouping purposes in (Vogel and Graham, 2013). Particularly, the financial performance of 73 airports is evaluated via 9 Key Performance Indicators (KPIs) giving rise to 3 different clusters. Real flight data is used to feed a random forest model in (Rodríguez-Sanz et al., 2021), which allows to analyze the passenger behavior at airport queues and obtain valuable knowledge to optimize its management through reinforced learning.

A model for estimating costs of 15 airlines was developed in (López-Lázaro et al., 2018) in order to assess the economic impact of environmental measures. Operational and financial data from 12 airlines was also used in (Pineda et al., 2017) for developing a method for the improvement of airline performance. It combines DM techniques along with Multiple Criteria Decision Making (MCDM) along with other tools for the extraction of critical features and an optimization method to find the most adequate alternative.

A comprehensive analysis of the air services available at certain cities is conducted in (O'Connor and Fuellhart, 2012) focusing on airlines and aircraft data. One of the authors' conclusions is that a dynamic analysis would be useful to explore the changes over time due to evolution of fleet or demographics among other aspects.

Several works have tackled the safety topic in aviation through DM. Some of the most popular are: (Pagels, 2015) where 4 methods are applied for the identification of safety issues before they happen; (Koteeswaran et al., 2019) which develops a method for predicting accidents and (Merzbacher, 2019) which focuses on detecting explosives.

Other examples of works that have tackled the application of DM techniques to aviation data are: (Ayhan et al., 2013) developing predictive models based on flight data; (Liau and Tan, 2014) and (Eti and Mizrak, 2020) which mine tweets or text in order to analyze customer's perception over carriers; or (Gorripaty et al., 2017) where weather and traffic are analyzed for easing Air Traffic Management (ATM) decisions.

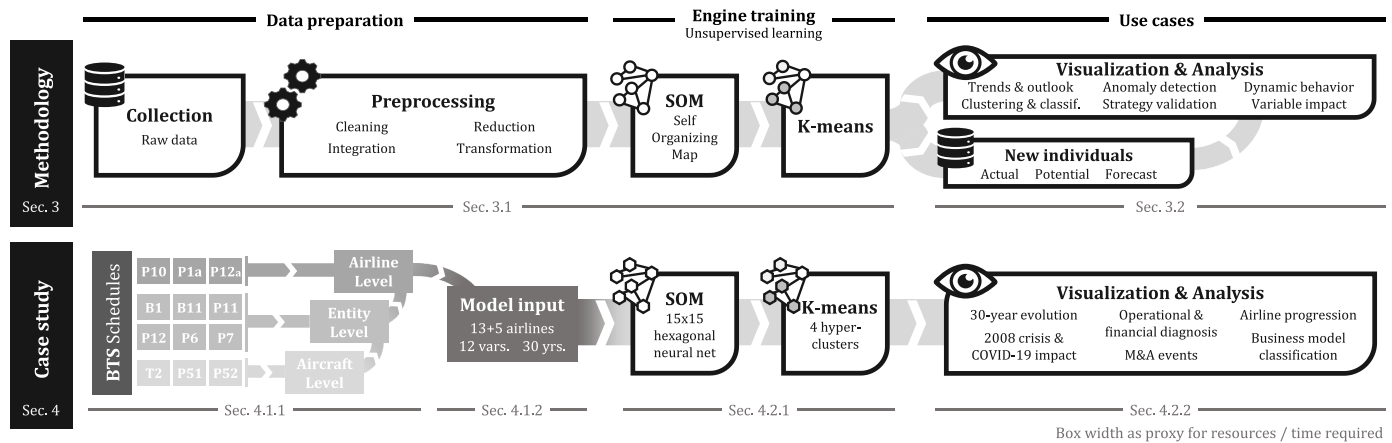


Fig. 1. Methodological workflow.

Thirdly, given the unprecedented impact the COVID-19 has had on air transport, abundant literature has addressed this topic. In fact, as analyzed in (Tanriverdi et al., 2020), much of the scientific production has focused on assessing the influence of the crisis on the business as well as on identifying potential recovery measures. In particular, both data and the opinion from a sample of aviation executives is examined in (Suau-Sanchez et al., 2020) in order to explore the forced transformation of the sector.

A couple of works that have tackled the practical application of ML algorithms during the pandemics are (Pérez-Campuzano et al., 2021a), which gathers and proposes a handful of strategic AI applications that could be implemented during these times of crisis, and (Tsa and Hung, 2021), where a model is modified to predict enterprise performance and compared against different engines such as Neural Networks.

All the aforementioned studies tackle the topic of using aviation-related data and apply DM techniques. However, none of them, to the knowledge of the authors, focuses on the analysis and visualization of the historical evolution of the market players (including the COVID-19 impact) neither in the use of Self Organizing Maps (SOMs), which is the objective of the proposed approach and main novelty of this paper.

3. Methodology: proposed Data Mining approach

This section defines the proposed DM approach including its main stages and characteristics as well as aspects to consider during the whole implementation process. A summary of this methodological workflow is featured in Fig. 1.

3.1. Model description

Firstly, it is worth mentioning that the collected raw database (which includes the information that will behave as the input for the model) will probably require a preprocessing phase including stages such as cleaning, integration, reduction, transformation or other statistical adjustments (Han et al., 2012). The objective is to prepare the data to be fed into the model, but the specific stages would be dependent on the nature, structure and quality of the raw data. In order to ease the process, it is advisable to build a database in which each individual or sample represents the performance or attributes (either financial or operational data) of each market player for a given timeframe (e.g. year or quarter).

The scope of the used variables or attributes is very flexible. For example, a comprehensive study could include high-level data of different areas of interest (e.g. financial, operational or even combined variables). On the other hand, and depending on the degree of detail of the raw database, the method could be applied at a lower level by only using variables from a specific area of expertise (e.g. direct and indirect expenses distributed in different allocations).

Table 1

Clustering algorithms initially considered and their generic features, non-comprehensive. Adapted from (Tan et al., 2013).

Algorithm	Nesting	Exclusiveness	Completeness	Visualization
SOM	Partitioning	Exclusive	Absolute	Maps
K-means	Partitioning	Exclusive	Complete	No
DBSCAN	Partitioning	Exclusive	Partial	No
Hierarchical	Hierarchical	Overlapping	Complete	Dendrogram
Fuzzy	Partitioning	Fuzzy	Complete	No

With the aim of finding and understanding the underlying relationships between the different market players, the proposed model engine is based on an unsupervised cluster analysis. This exploratory science develops and assesses different methods for grouping or structuring datasets according to intrinsic characteristics or attributes (Jain, 2010). Furthermore, it represents a taxonomic tool whose main goal is to maximize the intraclass similarity while minimizing the interclass similarity. Clustering is usually comprised by tools based on unsupervised learning since no labels are available a priori for classification.

Several algorithms for clustering have been developed in the last decades. A comprehensive review can be found in (Jain, 2010) along with a description of the challenges and issues while using them. As stated by (Tan et al., 2013), clustering algorithms can be categorized according to the following features: nesting, exclusiveness, completeness and visualization. With the objective of selecting the most suitable tools for the purpose of the analysis, five techniques are preliminarily considered in this study, those shown in Table 1.

In this work, *hierarchical, overlapping, fuzzy or partial* algorithms are discarded due to the requirements of the clustering requirements (see description below). Eventually two different engines, sequentially applied, have been chosen for the method:

A. Self-Organizing Maps (SOMs). The main idea behind SOMs, in addition to clustering, is visualizing on a 2D map the multidimensional relationships (in terms of links and distances) between those clusters (Kohonen, 2001). This visualization of the inner relationships may ease the analysis in real business cases. In this type of neural network, neurons (also referred to as nodes or units) act as centroids for each cluster and are set up as an interconnected net which maintains the topology of the original dataset. In fact, if this neighbor concept was removed, their behavior would be similar to traditional K-means. Some SOM outputs that can be used for the analysis include not only the clusters built for each neuron (and their correspondent individuals) but also the distances between neurons (a proxy for similarity) and the relative weights of the attributes for each of the neurons (a proxy for the classification relevance). As

shown in Table 1, SOM is not a pure *complete* algorithm (in the sense that all the samples are allocated to a given neuron or cluster, namely, no sample is left orphan). Instead, aiming at mapping the underlying structure of the information, there is also the possibility that some neurons may not be assigned any training sample (*absolute*). This is relevant when considering the potential subsequent use of the SOM to map or interpolate samples or individuals afterwards and independently from the training subset.

- B. **K-means.** This is the most common algorithm for *partitioning* clustering purposes and it is based on the grouping of the samples into an a priori k number of groups. It aims at minimizing the Sum of Squared Errors (SSE) or variance criterion, which is computed as the Euclidean distance between each individual and the cluster node to which it has been allocated (Bock, 2008). After initialization, the centroids of each cluster are moved iteratively following the points of the dataset which are assigned to it until the calculations converge. For the purposes of this work's method, and in order to hypercluster the previously SOM neurons, the desired algorithm must be *exclusive* (including each individual in only one cluster) and *complete* (grouping all the individuals but without creating empty clusters). This is the reason behind the selection of the K-means for this second step given the features in Table 1.

These two algorithms have been extensively compared in the literature. While SOMs show fair capabilities for data visualization and representation of data in a lower dimension (Mangiameli et al., 1996), K-means may achieve better accuracy (Mingoti and Lima, 2006). Hence, the characteristics of the target dataset as well as the purpose of the analysis can be factors to take into account when deciding on which algorithm to use.

Attention must be paid in both cases (but specially for SOMs) to the training stage. During this phase, all the network parameters or cluster nodes and weights are defined by means of an unsupervised learning algorithm that aims to optimize a predefined objective function (usually minimizing the sum of errors estimated as the Euclidian distance between each individual and its designated node).

Another consideration is that SOMs, due to their 2D-interconnected neurons, may work better than K-means for multidimensional populations that are suspected to be warped in a lower-dimensional space (which can be the case when some variables are partially correlated). SOMs seem also to show a more robust behavior in the sense that K-means usually enforces clusters to be of a similar size, while some of the SOM neurons can be empty (with no individuals assigned, see description for *absolute* above).

One important parameter that requires special attention is the number of clusters (neurons in the case of SOMs) chosen during the analysis (Fraley and Raftery, 1998). Commonly, K-means is used with a low number of groups (2–10) while SOMs are usually comprised by nets in the order of 100 neurons (e.g. in a 10×10 grid). The eventual selection of this number of clusters should be the result of a trade-off between six main factors: (i) the desired outcome of the of the analysis, (ii) the easiness of visualization, (iii) the number of samples or individuals that are being clustered, (iv) the intra-cluster cohesion and intra-cluster separation, (v) the final sum of errors derived from the clustering, and (vi) computing resources required. Although the decision and unsupervised validation may be taken empirically (based on external expert decision for i or ii) or a priori (e.g. based on iii), there are qualitative techniques such as the *silhouette coefficient* (Kaufman and Rousseeuw, 1990) and the *elbow* or *scree test* (Pérez-Campuzano et al., 2018) that can help to find a preliminary estimate (based on iv and v respectively).

The approach proposed here consists in firstly set up the SOM with a relatively high number of neurons (based on i, ii and iii) and feeding it with the initial dataset of individuals and, after that, applying the K-means over the resulting neurons of the SOM, using this time a lower number of hyperclusters. This can eventually ease the analysis phase since the complex SOM can be broken down into fewer hyperclusters

with similar features (thus grouping the individuals into a high-level classification) while also maintaining the topology of the network as well as its lower-level detail and 2D visualization.

3.2. Potential use cases

This proposed method offers several opportunities as potential use cases or applications. In terms of their time scope, they can be divided into the two categories below:

- **Historical analysis.** The most straightforward use is the assessment of the past evolution of the market by the direct application to actual historical data. Market trend identification and competitor surveillance are some of the potential outcomes.
- **Forecasting.** A prognosis could be carried out based on the actual past behavior and its future evolution either by assuming prospective market constraints or the impact of potential actions a company is assessing within its strategic roadmap.

In both cases, the method can be used to extract insights in many different aspects. Some of them are the impact of intrinsic or extrinsic crises or disruptive events (such as the 9/11 attacks, the financial crisis in 2008 or the COVID-19), the analysis of the variables or attributes (especially by the assessment of the neurons weights and their relative impact in each of the clusters) or the analysis and comparison of competitors.

The business model definition (FSNC, LCC, etc.) has been an historical research topic in the air transport literature (Sengur and Sengur, 2017). In the past, several models have been developed in order to classify the different carriers through the business model spectrum: (Mason and Morrison, 2008), (Lohmann and Koo, 2013) and (Jean and Lohmann, 2016). According to the DM method proposed in this paper, an examination of the network architecture and its underlying relationships with the variables employed could also lead to specific conclusions in this regard.

Another potential area of study is the Mergers & Acquisitions (M&A) activity. This could be addressed by analyzing the consequences of past mergers in terms of the evolution of both the acquiree and, specially, the acquirer or by feeding the model with potential individuals build based on the consolidated figures of two existing carriers. The results could be also compared with other conclusions extracted from the literature such as the relation between the profitability of the airlines involved in the M&A operation and its impact on the cost structure of the resulting firm (Gudmundsson et al., 2020).

Regulation is also a relevant topic in the air transport industry since it has played a major role in the past by means of impacting the procedures, operations or finance of the carriers. The method could be used in order to estimate the impact of future regulations in the industry as well as to support the definition of potential mandatory requirements such as those enforced after the COVID-19 outbreak (Abate et al., 2020).

A final consideration for the use of this method is that, once the SOM or K-means model is built and trained (i.e. all the network parameters or cluster nodes and weights are defined), it can be fed with new data (additional competitors, years, hypotheses, etc.). This can be useful in order to analyze in which clusters these new individuals (actual or based on assumptions) should be allocated. However, if any of these new samples fall outside the domain of the initial training set or population the resulting extrapolation may result inaccurate or biased; hence retraining with the extended dataset might be recommended.

4. Case study: historical evolution of the US airline industry

As a way to showcase the possibilities of the method, in this section the proposed approach is applied to a historical database of financial and operational data from 18 US airlines gathered through the years 1991–2020. The arrangement of the process workflow has been featured

Table 2
Summary of aggregated databases developed from BTS data for the years 1991–2020.

Database	Samples		Variables/Attributes	
	Total	Per year (average)	Total	Included <i>schedules</i>
Airline	3492	116	366	T2, P51, P52, B1, B11, P11, P12, P6, P7, P10, P1a, P12a
Entity	5317	177	324	T2, P51, P52, B1, B11, P11, P12, P6, P7
Aircraft	17,936	598	93	T2, P51, P52

in the bottom half of Fig. 1.

4.1. Data preparation

As already mentioned, data preparation represents a very important (and resource consuming) process when dealing with DM or ML projects. In this section, the main actions taken in order to handle the data are described, from the raw sources up to the model input.

4.1.1. Raw data cleaning and integration

Data in this analysis is gathered from the Bureau of Transportation Statistics (BTS), one of the federal agencies in the US that collects data from transportation economics, multimodal freight and commercial aviation. In particular, data from the Form 41 and other air operational schedules is used in this study. Original sources can be found at (BTS, 2018) and a useful introduction to these resources is gathered in (Durso, 2007).

Aviation databases from BTS are available in a semi-structured format and each of the different *forms* and *schedules* shows a different attribute distribution and temporal scope, which difficulties its usage for data analysis. For example, while some databases are segregated by airline and with annual data, others are built at a monthly aircraft level.

In order to prepare this semi-structured data for the analysis, a pre-processing stage was required. The main preprocessing steps for DM according to (Han et al., 2012) are summarized in the subsequent sections. They are presented in the most common order although this may vary depending on the particular data and application.

The first step comprises cleaning actions such as noise and

inconsistent data removal or missing values handling (Li, S. & Shue, 2004). In this case, certain bugs or defects in the raw databases were found (such as missing identifying attributes or incorrect year labeling) and fixed during the process.

Secondly, integration was required considering that the original data was extracted from diverse databases, each of which presents a different layout, arrangement, composition, identifying attributes, size, etc. This step aims to merge and combine all that data into a common structure and format. With this purpose, the different *forms* and *schedules* were merged by (i) aggregating figures in annual intervals, (ii) using certain common attributes as identifiers (year, airline, aircraft, etc.) and (iii) adding each particular variable (revenue, expenses, Available Seat Miles, etc.) as a new attribute to each individual.

Following these premises, data was consolidated into three different

Table 4
Model variables: Finance, Operations and Composites.

	Variable	BTS Variable (<i>Schedule</i>)	Units
Finance	Operating Revenue	Op. Revenue (P12)	USD
	Operating Margin	Op. Margin (P12)	-
	Profit Margin	Op. Revenue (P12)	-
		Pr. Loss (P12)	-
Working Capital Ratio	Op. Revenue (P12)	-	
	Curr. Assets (B1)	-	
Operations	Load factor	RPM (T2)	-
		ASM (T2)	-
	Average Route	Miles Flown (T2)	mi/departure
	Distance	Departures (T2)	-
Fuel efficiency	RPM (T2)	(PAX*mi)/gallon	
	Fuel cons.(T2)	-	
Composite	Fuel Cost Ratio	Fuel Cost (T2)	-
	Employee yield	Op. Expenses (P12)	-
		Op. Revenue (P12)	USD/FTE
	Cost per ASM	Total FTE (P10)	-
		Op. Expense (P12)	USD/(seat*mi)
	Revenue per ASM	ASM (T2)	-
Op. Revenue (P12)		USD/(seat*mi)	
Yield	ASM (T2)	-	
	Op. Revenue (P12)	USD/(PAX*mi)	
		RPM (T2)	-

Table 3
Airlines (existing and extinct) included in the study.

	Carrier	IATA Code	Model	Recorded years	Revenue (€bn)			
					2019	2020	Δ	
Existing	1	Delta Airlines	DL	FSNC	30 91-20	47.1	17.1	-64%
	2	American Airlines	AA	FSNC	30 91-20	45.8	17.3	-62%
	3	United Airlines	UA	FSNC	30 91-20	43.3	15.4	-65%
	4	Southwest Airlines	WN	LCC	30 91-20	22.4	9.0	-60%
	5	Alaska Air Group	AS	FSNC	30 91-20	8.8	3.6	-59%
	6	JetBlue Airways	B6	LCC	21 00-20	8.1	3.0	-63%
	7	Spirit Airlines	NK	LCC	29 92-20	3.8	1.8	-53%
	8	SkyWest Airlines	OO	Reg.	18 03-20	2.9	2.1	-28%
	9	Hawaiian Airlines	HA	FSNC	30 91-20	2.8	0.8	-70%
	10	Frontier	F9	LCC	27 94-20	2.5	1.3	-50%
	11	Allegiant Air	G4	LCC	23 98-20	1.7	0.9	-47%
	12	Envoy Air	MQ	Reg.	30 91-20	1.4	1.0	-29%
	13	Republic Airlines	YX	Reg.	16 05-20	1.3	1.0	-27%
Extinct	Carrier	IATA Code	Model	Recorded years	Last 2 years revenue (€bn)		Merged into	
	1	US Airways	US	FSNC	25 91-15	15.8 ¹⁴	7.5 ¹⁵	AA
	2	Northwest	NW	FSNC	19 91-09	14.1 ⁰⁸	10.9 ⁰⁹	DL
	3	Continental	CO	FSNC	21 91-11	14.0 ¹⁰	16.2 ¹¹	UA
	4	AirTran	FL	LCC	19 94-12	2.9 ¹¹	0.7 ¹²	WN
5	Virgin America	VX	LCC	11 07-17	1.7 ¹⁶	1.6 ¹⁷	AS	

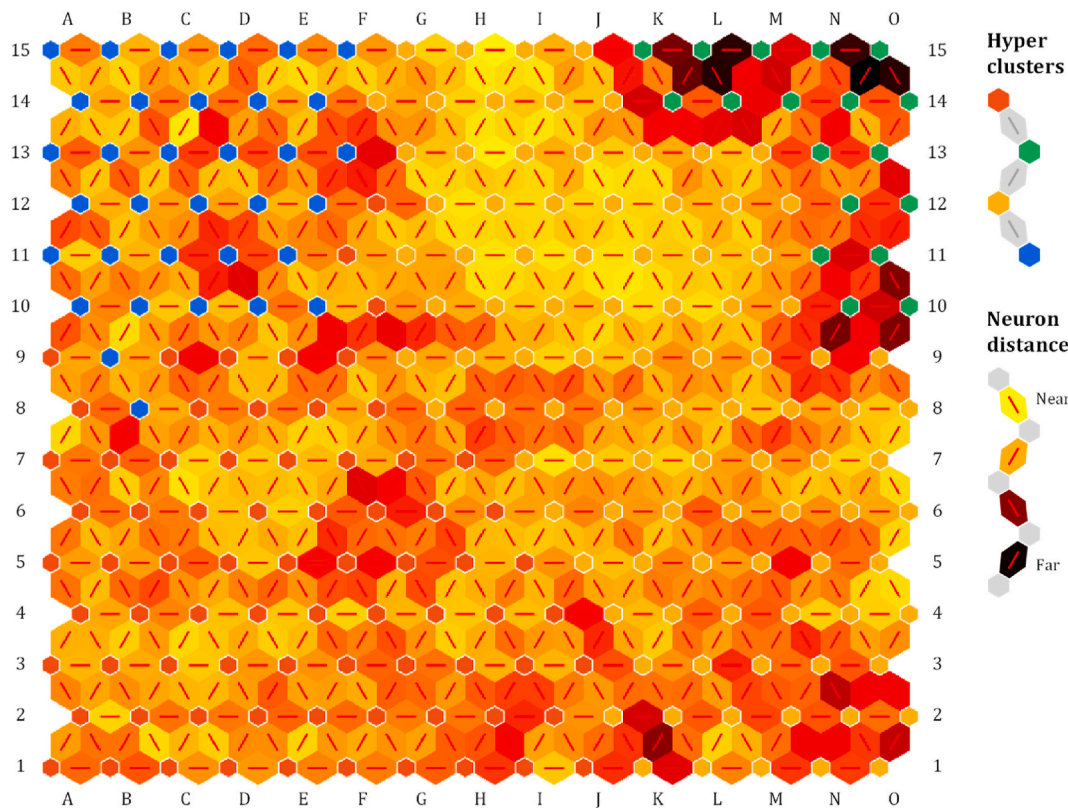


Fig. 2. Non-dimensional distances between the SOM neurons.

levels (from higher to lower): airline, entity (airline's market unit which operates one region) and aircraft. Not all the data is available at entity or aircraft level in the BTS repository so these databases comprise fewer variables or attributes than the one at airline level. The resulting aggregated databases include data from 1991 to 2020 and a summary of their main features can be read in Table 2.

4.1.2. Model input reduction and transformation

For this case study, one of the aforementioned aggregated databases was used, the one which gathers annual data at airline level. Particularly, the 18 airlines shown in Table 3 have been included in the analysis as a reduced population. This selection is originally based on the top 13 US passenger carriers in terms of revenues in 2019. In addition, 5 extinct airlines that were merged into some of the former in the recent past have been also included in the analysis (last five rows in Table 3). This aims to properly visualize how the M&A consolidation modified the industry landscape. In total, the population included in the simulation amounts to 439 individuals (total sum of the recorded years).

Regarding the variables used to feed the model, Table 4 shows the 12 different attributes analyzed for each carrier. The selection of these variables has been based on external a priori considerations such as (i) their capacity to represent the global business performance (thus combining some of the original variables), (ii) their availability in the BTS database (low ratio of missing values on the records) or (iii) the possibility of being accessible in other markets or organizations in order to reproduce and compare this analysis, or even to include airlines from other countries directly into this model. It must be also mentioned that the input dataset was transformed through standardization in order to avoid statistical bias in the model due to different orders of magnitude in the variables' units.

4.2. Results

In this section, the results of the application of the methodology to

the historical US case study are shown along with some observations arisen from its analysis and visualization. Firstly, the resulting network structure and parameters are discussed and later the analysis focuses on the specific individuals (carriers throughout the years) and their allocation in the network.

4.2.1. Network architecture

Once the model input has been prepared and preprocessed it is ready to be fed into the engines. In this case the SOM is preconfigured as a 15×15 neural network with hexagonal neurons (each one being connected to other 6). After training the SOM, the K-means is applied to the resulting neurons using 4 hyperclusters (Red, Green, Yellow and Blue). Both processes are carried out by means of ML engines and specifically with unsupervised learning algorithms, since no labelling or categorization was attached to the data a priori. As a way to better understand the relationships between neurons and between hyperclusters, Fig. 2 shows the non-dimensional multivariate distances among them.

It can be observed how the distances relatively resemble the results of the K-means grouping. The higher distances usually correspond to the borders between the hyperclusters built by the K-means engine. This is clearly visible in the Blue-Red-Yellow border around F13 or F10, in the Red-Yellow border in J1-H7 or in the Yellow-Green border in J15-K13-M13.

However, due to the complexity of the network, the distances represented in Fig. 2 also allow to extract a deeper detail in the relationship between neurons within the same hypercluster. For example, the upper area of the Yellow hypercluster seems quite homogeneous (small distances) while the Green hypercluster seems quite fragmented given that some specific neurons have been allocated far away from their neighbors (L15, O15 and O10).

Another type of analysis can be conducted by assessing the weights of the neurons assigned to each variable, represented in the maps of Fig. 3. This may be useful in order to assess the relevance of the attributes in the SOM architecture definition or to easily identify outliers among the

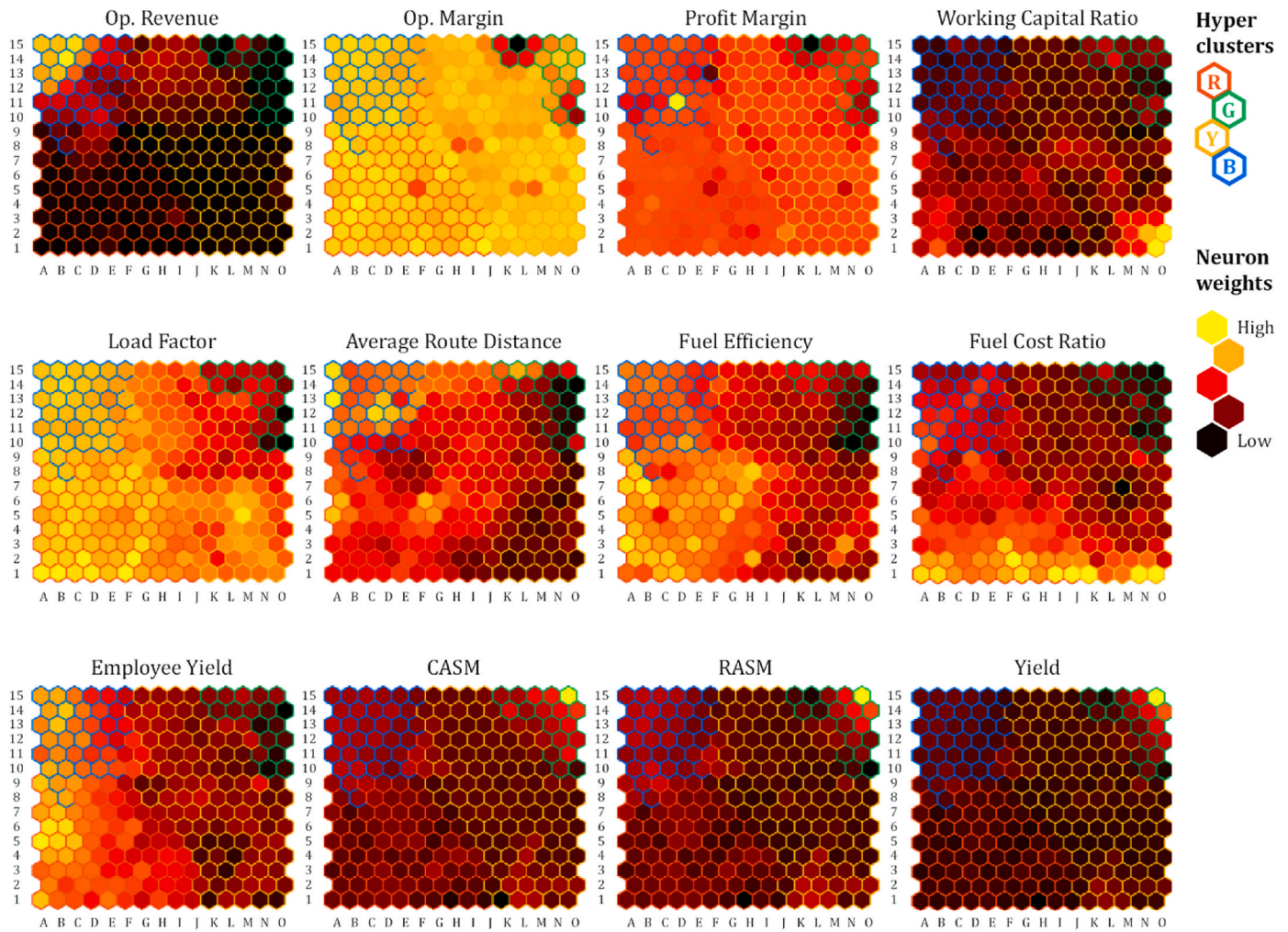


Fig. 3. Non-dimensional weights of the SOM neurons for each variable.

Table 5
Non-dimensional weights of the SOM neurons for each hypercluster and variable.

	Op. Revenue	Op. Margin	Profit Margin	Working Capital Ratio	Load Factor	Average Route Distance	Fuel Efficien.	Fuel Cost Ratio	Employ. Yield	CASM	RASM	Yield
Red												
Green												
Yellow												
Blue												

Neuron weights

High Low

population.

In some variables, the SOM has been automatically structured in such a way that the individuals with the highest values are clearly grouped towards specific areas, which may have eventually also influenced the definition of the hyperclusters. This happens for example for the *Op. Revenues* (top-left, corresponding with the Blue hypercluster) or the *Working Capital Ratio* (bottom-right).

Other variables seem quite biased due to outliers such as the *Op. Margin*, *Profit Margin* or the *CASM*, *RASM* and *Yield*, these last 3 also

showing a certain level of correlation among them. It is worth mentioning that most of these outliers (L15, O15 and O10) have been allocated to the fragmented Green hypercluster.

For the rest of the variables, the distribution is more heterogeneous although some patterns can also be identified. For example, for the *Load Factor*, *Average Route Distance* and *Employee Yield* the values grow from right (Yellow-Green) to left (Red-Blue), while for the *Fuel Cost Ratio* this happens from top (Blue-Yellow-Green) to bottom (Red-Yellow).

This information regarding the distribution of the weights of neurons

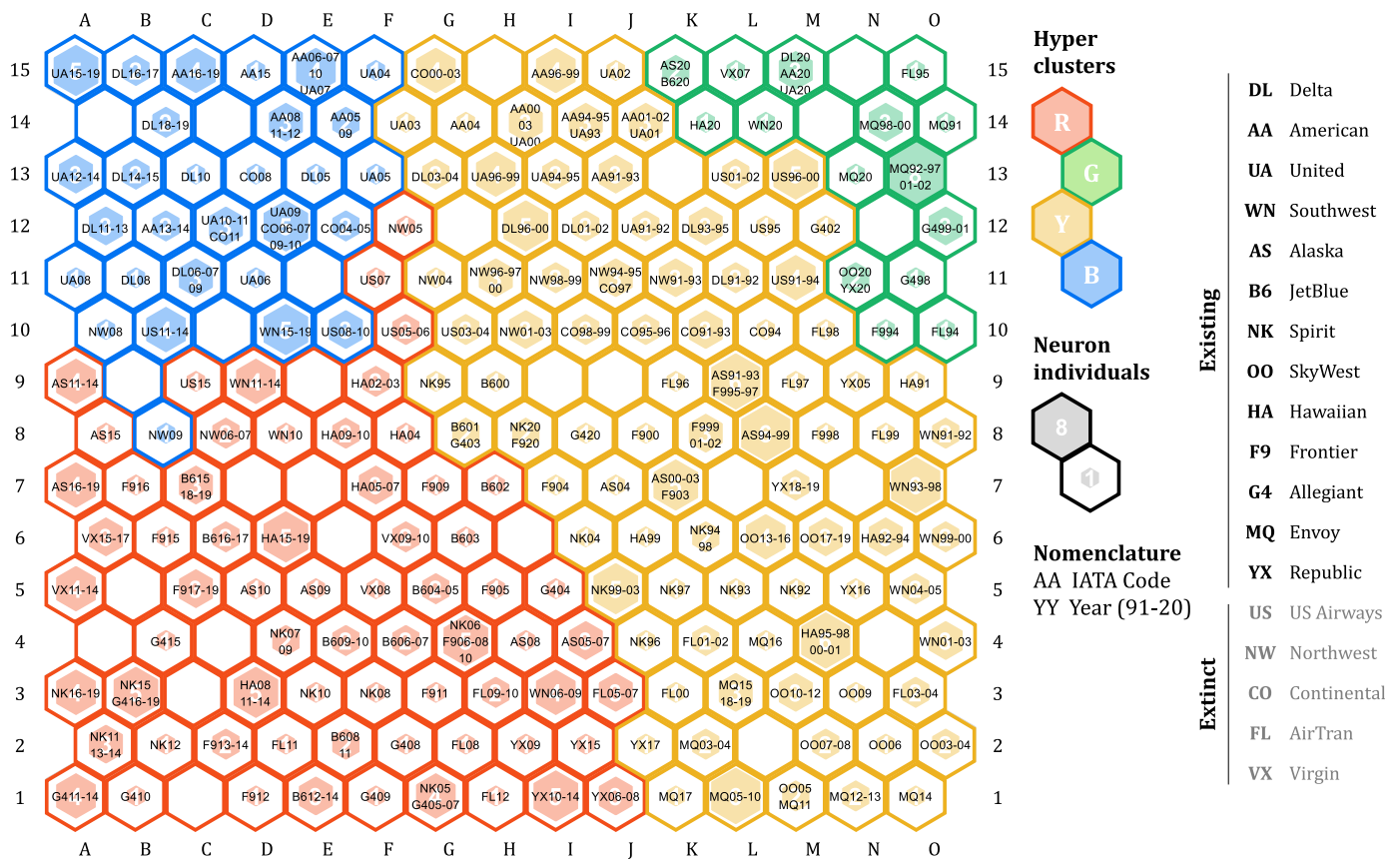


Fig. 4. Allocation of individuals (airline & year) within the SOM neurons.

for each of the hyperclusters is qualitatively summarized in Table 5. Whenever the neurons within a certain hypercluster featured very different weights values a striped hexagon is shown with the range of different values.

4.2.2. Airline visualization and analysis

This section analyzes the allocation of the specific individuals through the network. As a general picture, the resulting individual allocation is represented in Fig. 4. Each hexagon represents one neuron and includes the individuals (airline and year) allocated to it. Hyperclusters resulting from the application of the K-means algorithm (Red, Green, Yellow and Blue) are also represented using colors.

In order to ease the visualization of each of the carriers' evolution, the neurons where a certain airline is present have been colored and connected with lines following its annual evolution in Fig. 5. This leads to a representation that resembles star constellations and which illustrates how the individual has evolved throughout the SOM and along the years.

One of the main common behaviors shown in Fig. 5 is that most of the airlines move chronologically from right to left according to the SOM architecture. Additionally, the different business models seem to be allocated into certain regions of the SOM and the reasons may be partially explained by the value of the weights of the variables in Fig. 3 and Table 5. In particular, the big-scale FSNCs (DL, AA, UA, US, NW and CO) feature horizontal constellations that tend to be placed in the upper-left part of the SOM (Yellow and Blue hyperclusters) where the *Op. Revenue* and *Average Route Distance* are higher. Mid-scale LCCs occupy the mid-Yellow and Red hyperclusters, mainly characterized by high *Load Factor*, *Fuel Efficiency* and *Employee Yield*. Finally, Regionals (OO, MQ and YX) tend to populate the bottom-right area (Yellow) probably due to low *Average Route Distance* and *Fuel Efficiency* as well as high *Fuel*

Cost Ratio.

It is worth noting that, while the general evolution seems quite smooth for most of the individuals (with relatively stable direction and small jumps from neuron to neuron or year over year), there are some events that seem to cause more erratic or sharp movements in the map.

The 3 major mergers included in the analysis (US into AA, NW into DL and CO into UA) represent one example of these sudden changes in the constellations (AA in B12-D15, DL in C11-C13 or UA in C12-A13). Probably due to the smaller scale of the acquirees, the rest of M&A operations (FL into WN and VX into AS) do not seem to have caused a severe impact in the chronological trajectory of the acquirer.

Another event that seems to cause entanglements (broad jumps as well as back-and-forth movements) in the constellations is the crisis in 2008 and subsequent years (e.g. DL in B11-C11, AA in D14-E14-E15, B6 in E2-E4, NK in F3-D4, HA in D3-E8, F9 in G4-G7 or CO in D13-D12). On the contrary, other airlines show a smoother evolution through those difficult years (e.g. WN in I3, AS in H4, OO in M2 or G4 in F2) as a potential sign of higher resilience.

The COVID-19 has caused an unprecedented shock in the airlines, which is clearly noticeable in the different constellations. Most of the airlines abruptly jumped in 2020 (dashed lines) towards the top-right area of the SOM (Green hypercluster) probably as a consequence of the sharp decline in operations, revenues and profits.

Another simplified way to represent the results and evaluate the underlying dynamic behavior of the population is to observe the evolution of the airlines over the years through their hypercluster allocation. This is shown in Fig. 6, where the historical evolution of each carrier is classified according to the results of the K-means engine.

The first observation arising from Fig. 6 is that most of the carriers were allocated to the Yellow hypercluster until the early 2000's. Particularly, during the years 2004-06 there was a transition where

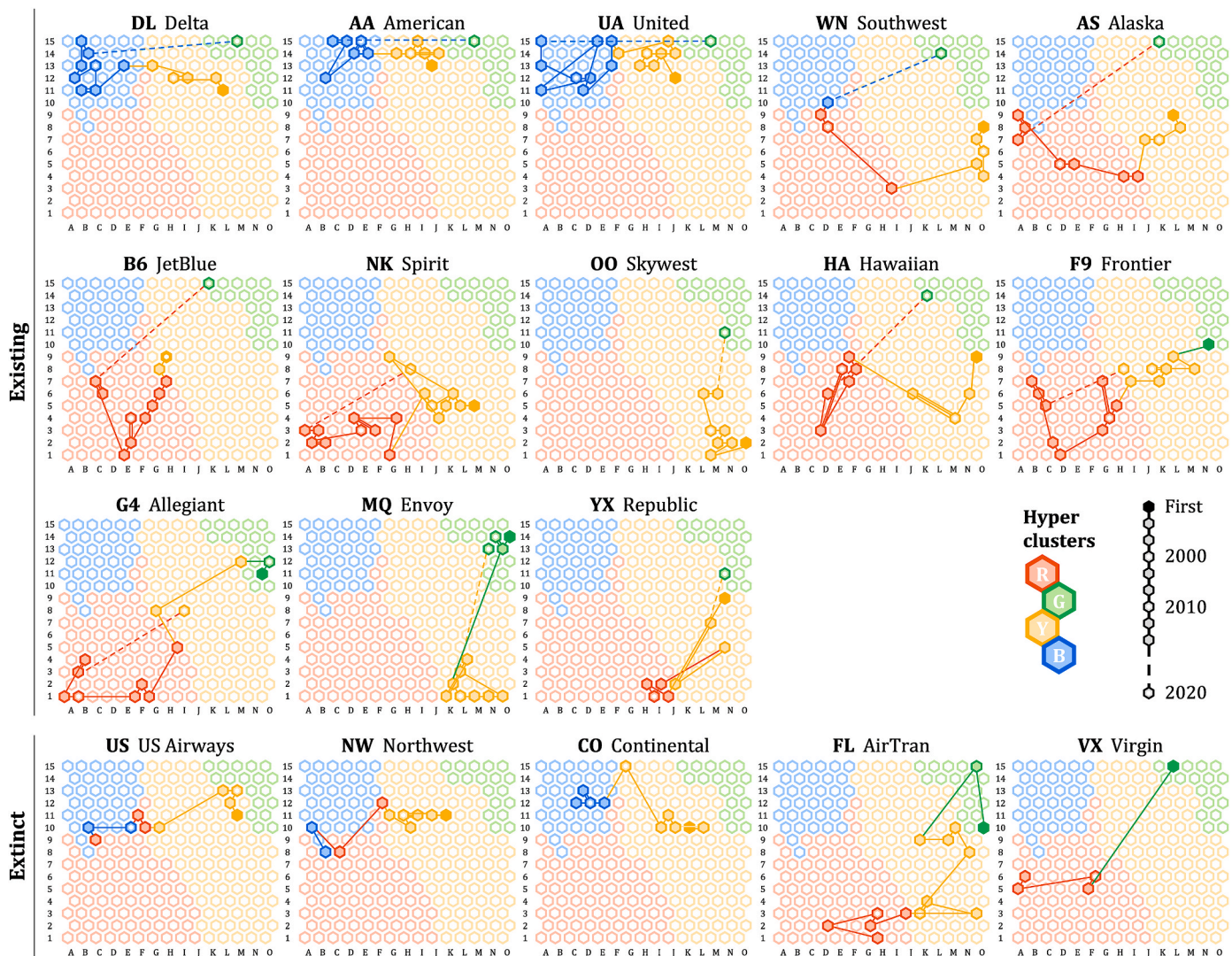


Fig. 5. Constellations: historical evolution and trajectory throughout the SOM for each airline.

most of the airlines moved towards Blue (in the case of FSNCs) or Red (in the case of LCCs). Regional carriers (OO, MQ and YX) show a particularly different behavior, allocated mostly to Yellow and Green. This demonstrates how the model, given the defined input variables, seems to show certain capability to discern between these 3 business models or even to help on the identification of business model shifts over time from a particular carrier.

As some exceptions or anomalies, it is worth mentioning the Green case, which seems to accommodate those individuals with low profitability such as the first years of operation (F9, G4, FL or VX) or most of the 2020 individuals affected by the pandemics. In fact, only 3 out of the 13 existing carriers have managed to move towards the Yellow hyper-cluster rather than to the Green one (NK, F9 and G4). This may be a manifestation of higher resiliency of those organization in face of market disruptions such as COVID-19.

5. Conclusions

In order to support the strategic decision-making process of the airlines, this paper proposes a Data Mining (DM) approach (focused on visualization) with the objective of extracting market knowledge from any database (e.g. financial or operational) of industry players or competitors. Particularly, after a review of the state-of-the-art, a method that combines two clustering techniques (Self-Organizing Maps -SOMs- and

K-means) is suggested for potential use in different sectors and with several promising dynamic applications. As a case study, after a pre-processing phase, 30-year data from 18 US passenger airlines is used to showcase the capabilities of this tool, which allows to draw general conclusions regarding the evolution of the market trends and players as well as particular observations in relation to relevant events, especially the COVID-19 consequences.

The model is configured as a SOM with a size of 15x15 hexagonal neurons which, after its training, are segregated into 4 different hyper-clusters using the K-means algorithm. A preliminary analysis of the resulting architecture helps to understand the underlying structure of the industry by examining the distances between neurons as well as the neurons weights that influence the allocation of the individuals according to the values of their variables.

A further analysis focusing on the specific carriers and their historical evolution throughout the network and hyperclusters allows to draw observations according to both their localization in the SOM as well as the degree of smoothness of its chronological trajectory. The former seems to show certain correlation with the corresponding business model and the hypercluster classification may be able to identify differences between FSNCs, LCCs and Regionals. Regarding the latter, the 3 biggest acquisitions included in the analysis cause sharp changes in the trajectory of the acquirers; this is not the case of the other 2 smaller mergers. Other events such as the 2008 crisis and, of course, the COVID-

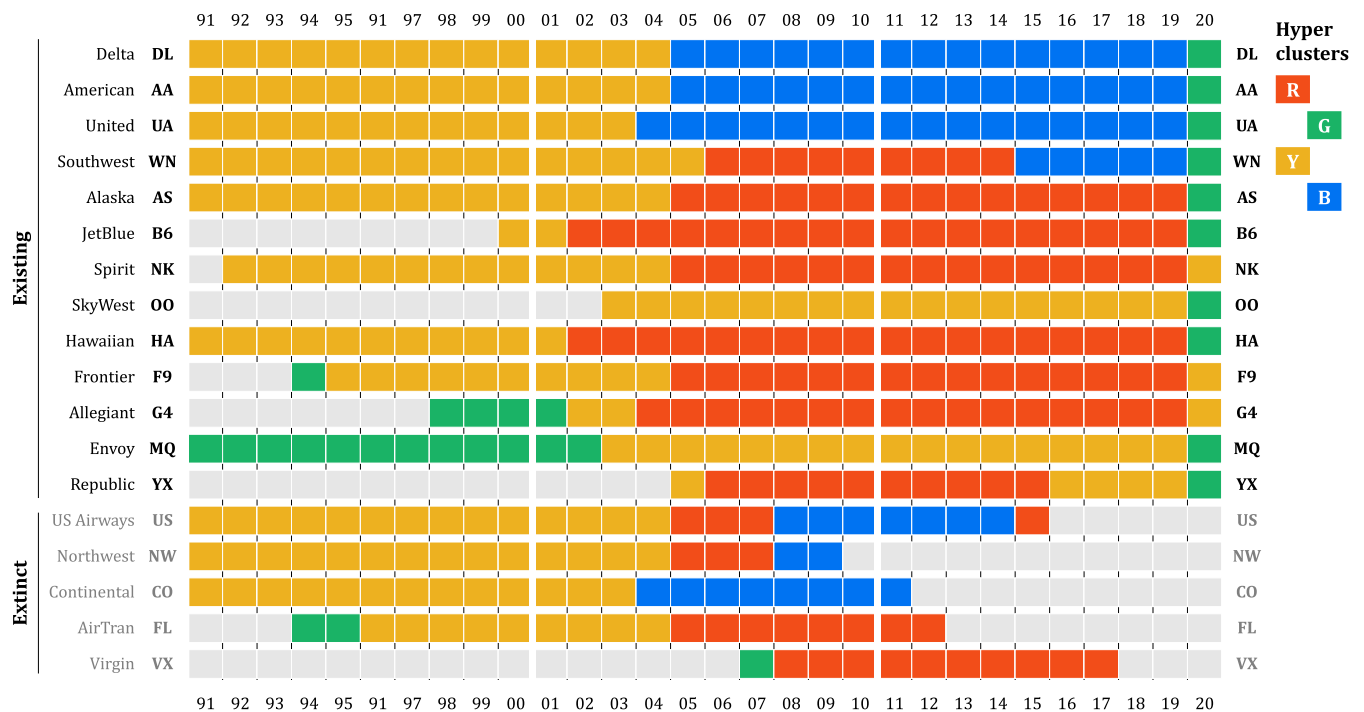


Fig. 6. Historical evolution throughout the hyperclusters for each airline.

19 shock are observable in the airlines' behavior, being characterized by back-and-forth movements in the former case and by an abrupt jump towards the low-profit area of the map in the latter.

A final study regarding the hypercluster allocation per carrier through the 3 decades seems to show a certain capability by the model to discern between the 3 business models considered (FSNC, LCC and Regionals). Outliers and the evolution of the airlines can be also identified and assessed using this visualization.

As potential pathways for future developments, the nature of the analysis could be modified using as input databases segregated by data nature such as operational data, revenues, expenses, etc. This could give more insight regarding airline grouping by categories or business models, not only the traditional ones (FSNC, LCC, etc.) but also other alternative types such as the Airlines-within-Airlines (Pearson and Merkert, 2014). Segregation could be also carried out in terms of different aircraft, entities (markets served) or air service classes. Furthermore, players from other markets (EU, APAC, etc.), segments (passenger, cargo), modes (road, rail or water), value chain stages (manufacturers, airports, handling firms, etc.) or industries (renewable energy, banking, health, etc.) could be also studied using the proposed DM approach.

The model engine or clustering techniques used here could be also further explored. In particular, other unsupervised clustering tools could be applied to the databases such as t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), fuzzy c-means, or traditional hierarchical clustering algorithms as mentioned in (Li, S. & Shue, 2004).

In addition, anomaly detection techniques could be also used for finding anomalous cases in the industry, which may be useful to understand past events and foresee coming disruptions or failures. A similar approach has been applied in (Li et al., 2016) but for flight operation and safety monitoring in that case. Besides, supervised learning methods may be used for developing predictive or estimation models.

Regarding the variables used, a dimensionality reduction and feature selection stage could be also done using filters, wrappers or embedded methods with techniques such as autoencoders, correlation coefficients

(both input-input or input-output), or Principal Component Analysis (PCA) as in (Pérez-Campuzano et al., 2016).

To conclude, this and many other similar papers show how new digital technologies may enable the extraction of additional wisdom from databases and complement the traditional decision-making process. Hopefully the approach showcased throughout this paper may encourage practitioners to start applying these Artificial Intelligence tools within the industry.

Funding

This work was supported by LLM Aviation and Euroairlines.

Declaration of competing interest

The authors declare that they have no conflicts of interest with the contents of this paper.

CRedit authorship contribution statement

Darío Pérez-Campuzano: Conceptualization, Methodology, Investigation, Software, Visualization, Writing - original draft. **Luis Rubio Andrada:** Conceptualization, Investigation, Validation, Writing - review & editing, Project administration. **Patricio Morcillo Ortega:** Conceptualization, Supervision, Writing - review & editing. **Antonio López-Lázaro:** Data curation, Supervision, Writing - review & editing, Funding acquisition.

Acknowledgements

This study was carried out with the support of researchers and practitioners from LLM Aviation, Euroairlines and the Universidad Autónoma de Madrid (UAM); hence their cooperation is hereby gratefully acknowledged. This paper is comprised within the Aviation Research Framework led by those and other public and private entities from the air transport sector. The authors are also indebted to the reviewers for their valuable suggestions and comments.

References

- Abate, M., Christidis, P., Purwanto, A.J., 2020. Government support to airlines in the aftermath of the COVID-19 pandemic. *J. Air Transport. Manag.* 89, 101931. <https://doi.org/10.1016/j.jairtraman.2020.101931>.
- Akerkar, R., 2014. Analytics on big aviation data: turning data into insights. *Int. J. Comput. Sci. Appl.* 11 (3), 116–127. Retrieved from <http://www.tmrfindia.org/jicsa/v11i35.pdf>.
- Akpınar, M.T., Karabacak, M.E., 2017. Data Mining Applications in Civil Aviation Sector: State-Of-Art Review. Turkey: Retrieved from <https://www.semanticscholar.org/paper/Data-mining-applications-in-civil-aviation-sector-%3A-Akpınar/3a9a112832e72ad6decf2e981c28865deb0f2601>.
- Ayhan, S., Pesce, J., Comitz, P., Sweet, D., Bliesner, S., Gerberick, G., 2013. Predictive analytics with aviation big data, 2013. In: Paper Presented at the Integrated Communications, Navigation and Surveillance Conference (ICNS), 2013, pp. 1–13.
- Beecroft, M., 2019. The future security of travel by public transport: a review of evidence. *Res. Transport. Bus. Manag.* 32 <https://doi.org/10.1016/j.rtbm.2019.100388>. UNSP 100388.
- Bock, H., 2008. Origins and extensions of the k-means algorithm in cluster analysis. *J. Electronique d'Histoire Des Probabilités Et De La Stat.* 4 (2).
- BTS, 2018. Aviation Data Library. Bureau of transportation statistics (BTS). Retrieved from https://www.transtats.bts.gov/databases.asp?Mode_ID=1&Mode_Desc=Aviation&Subject_ID2=0.
- Campos, J.P., Rubio, L., 2017. Vigilancia tecnológica e inteligencia competitiva: elementos de apoyo al desarrollo de una cultura de innovación en las organizaciones. el caso ALSA. *Econ. Ind.* 406, 81–90.
- Chung, S., Ma, H., Hansen, M., Choi, T., 2020. Data science and analytics in aviation. *Transport. Res. E Logist. Transport. Rev.* 134, 101837 <https://doi.org/10.1016/j.trre.2020.101837>.
- Dube, K., Nhamo, G., Chikodzi, D., 2021. COVID-19 pandemic and prospects for recovery of the global aviation industry. *J. Air Transport. Manag.* 92, 102022 <https://doi.org/10.1016/j.jairtraman.2021.102022>.
- Durso, J.C., 2007. An Introduction to DOT Form 41 Web Resources for Airline Financial Analysis. Louisville, KY. Rubel School of Business, Bellarmine University.
- Eti, S., Mızrak, F., 2020. Analysing customer satisfaction of civil aviation companies of Turkey with text mining. In: Dincer, H., Yüksel, S. (Eds.), *Strategic Outlook for Innovative Work Behaviours: Interdisciplinary and Multidimensional Perspectives*. Springer International Publishing, Cham, pp. 21–41. https://doi.org/10.1007/978-3-030-50131-0_2. Retrieved from.
- Fraley, C., Raftery, A.E., 1998. How many clusters? which clustering method? answers via model-based cluster analysis. *Comput. J.* 41 (8), 578–588. <https://doi.org/10.1093/comjnl/41.8.578>.
- GE Accenture, 2014. *Industrial Internet Insights Report for 2015*. Accenture and General Electric.
- Gorripaty, S., Liu, Y., Hansen, M., Pozdnukhov, A., 2017. Identifying similar days for air traffic management. *J. Air Transport. Manag.* <https://doi.org/10.1016/j.jairtraman.2017.06.005>.
- Gudmundsson, S.V., Merkert, R., Redondi, R., 2020. Cost structure effects of horizontal airline mergers and acquisitions. *Transport Pol.* 99, 136–144. <https://doi.org/10.1016/j.tranpol.2020.08.017>.
- Gürbüz, F., Özbakir, L., Yapici, H., 2011. Data mining and preprocessing application on component reports of an airline company in Turkey. *Expert Syst. Appl.* 38 (6), 6618. <https://doi.org/10.1016/j.eswa.2010.11.076>.
- Han, J., Kamber, M., Pei, J., 2012. In Elsevier. In: *Data Mining: Concepts and Techniques, third ed.* Morgan Kaufmann.
- Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.* 31 (8), 651. <https://doi.org/10.1016/j.patrec.2009.09.011>.
- Jean, D.A., Lohmann, G., 2016. Revisiting the airline business model spectrum: the influence of post global financial crisis and airline mergers in the US (2011–2013). *Res. Transport. Bus. Manag.* 21, 76–83. <https://doi.org/10.1016/j.rtbm.2016.06.002>.
- Kaufman, L., Rousseeuw, P.J., 1990. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, Inc.
- Kohonen, T., 2001. *Self-organizing Maps, third ed.* Springer.
- Koteeswaran, S., Malarvizhi, N., Kannan, E., Sasikala, S., Geetha, S., 2019. Data mining application on aviation accident data for predicting topmost causes for accidents. *Cluster Comput.* 22 (5), 11379–11399. <https://doi.org/10.1007/s10586-017-1394-2>.
- Larsen, T., 2013. Cross-platform aviation analytics using big-data methods, 2013. In: Paper Presented at the Integrated Communications, Navigation and Surveillance Conference (ICNS), 2013, pp. 1–9.
- Li, L., Hansman, R.J., Palacios, R., Welsch, R., 2016. Anomaly detection via a Gaussian mixture model for flight operation and safety monitoring. *Transport. Res. C Emerg. Technol.* 64, 45–57. <https://doi.org/10.1016/j.trc.2016.01.007>.
- Li, S., Shue, L., 2004. Data mining to aid policy making in air pollution management. *Expert Syst. Appl.* 27 (3), 331. <https://doi.org/10.1016/j.eswa.2004.05.015>.
- Liau, B.Y., Tan, P.P., 2014. Gaining customer knowledge in low cost airlines through text mining. *Ind. Manag. Data Syst.* 114 (9), 1344–1359. <https://doi.org/10.1108/IMDS-07-2014-0225>.
- Lohmann, G., Koo, T.T.R., 2013. The airline business model spectrum. *J. Air Transport. Manag.* 31, 7–9. <https://doi.org/10.1016/j.jairtraman.2012.10.005>.
- López-Lázaro, A., Pérez-Campuzano, D., Benito, A., Alonso, G., 2018. Analyzing carbon neutral growth and biofuel economic impact for 2017–2025: a case study based on Spanish carriers. *Proc. IME G J. Aero. Eng.* <https://doi.org/10.1177/0954410018768610>.
- Maneep, S., Kotcharin, S., 2020. The impacts of COVID-19 on the global airline industry: an event study approach. *J. Air Transport. Manag.* 89, 101920. <https://doi.org/10.1016/j.jairtraman.2020.101920>.
- Mangiameli, P., Chen, S.K., West, D., 1996. A comparison of SOM neural network and hierarchical clustering methods. *Eur. J. Oper. Res.* 93 (2), 402. [https://doi.org/10.1016/0377-2217\(96\)00038-0](https://doi.org/10.1016/0377-2217(96)00038-0).
- Mason, K.J., Morrison, W.G., 2008. Towards a means of consistently comparing airline business models with an application to the 'low cost' airline sector. *Res. Transport. Econ.* 24 (1), 75–84. <https://doi.org/10.1016/j.retrec.2009.01.006>.
- Merzbacher, M., 2019. Lessons Learned: Data Mining and Aviation Explosives Detection Systems, 2019. <https://doi.org/10.1117/12.2518776>. Paper Presented at the , 10999 doi:10.1117/12.2518776 Retrieved from.
- Mingoti, S.A., Lima, J.O., 2006. Comparing SOM neural network with fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *Eur. J. Oper. Res.* 174 (3), 1742. <https://doi.org/10.1016/j.ejor.2005.03.039>.
- Nazeri, Z., Zhang, J., 2002. Mining aviation data to understand impacts of severe weather on airspace system performance, 2002. In: Paper Presented at the Proceedings. International Conference on Information Technology: Coding and Computing, pp. 518–523. <https://doi.org/10.1109/ITCC.2002.1000441>.
- O'Connor, K., Fuellhart, K., 2012. Cities and air services: the influence of the airline industry. *J. Transport Geogr.* 22, 46. <https://doi.org/10.1016/j.jtrangeo.2011.10.007>.
- Pagels, D.A., 2015. *Aviation data mining*. Scholarly Horizons: Univ. Minnesota Morris Undergrad. J. 2 (1), 3.
- Pearson, J., Merkert, R., 2014. Airlines-within-airlines: a business model moving east. *J. Air Transport. Manag.* 38, 21–26. <https://doi.org/10.1016/j.jairtraman.2013.12.014>.
- Pérez-Campuzano, D., Morcillo Ortega, P., Rubio Andrada, L., López-Lázaro, A., 2021a. Artificial intelligence potential within airlines: a review on how AI can enhance strategic decision-making in times of COVID-19. *J. Airl. Airt. Manag.* 11 (2) <https://doi.org/10.3926/jairm.189>.
- Pérez-Campuzano, D., Gómez, E., Gallego, C., 2016. Wind Turbine Fatigue Loads Statistical Estimation from Standard Signals. Retrieved from <http://oa.upm.es/45684/>.
- Pérez-Campuzano, D., Gómez-de-las-Heras, E., Gallego-Castillo, C., Cuerva, A., 2018. Modelling damage equivalent loads in wind turbines from general operational signals: exploration of relevant input selection methods using aeroelastic simulations. *Wind Energy* 21 (6), 441–459. <https://doi.org/10.1002/we.2171>.
- Pérez-Campuzano, D., Rubio Andrada, L., Morcillo Ortega, P., López-Lázaro, A., 2021b. A 32-year meta-analysis on artificial intelligence research in aviation. *ESIC Digit. Econ. Innov. J.* (1), 138–157. Retrieved from <https://revistasinvestigacion.esic.edu/edej/index.php/edej/article/view/29/33>.
- Pineda, P.J.G., Liou, J.J.H., Hsu, C., Chuang, Y., 2017. An integrated MCDM model for improving airline operational and financial performance. *J. Air Transport. Manag.* <https://doi.org/10.1016/j.jairtraman.2017.06.003>.
- Rodríguez-Sanz, Á., Fernández de Marcos, A., Pérez-Castán, J.A., Comendador, F.G., Arnaldo Valdés, R., París Loreiro, Á., 2021. Queuing behavioural patterns for passengers at airport terminals: a machine learning approach. *J. Air Transport. Manag.* 90, 101940 <https://doi.org/10.1016/j.jairtraman.2020.101940>.
- Sengur, Y., Sengur, F.K., 2017. Airlines define their business models: a content analysis. *World Rev. Intermodal Transp. Res.* 6 (2), 141–154. <https://doi.org/10.1504/WRIIR.2017.082732>.
- Shmelova, T., Sikirda, Y., Rizun, N., Kucherov, D., 2019. Cases on Modern Computer Systems in Aviation. IGI Global, Hershey, PA, USA. <https://doi.org/10.4018/978-1-5225-7588-7>.
- Shmelova, T., Sikirda, Y., Sterenharz, A., 2020. *Handbook of Research on Artificial Intelligence Applications in the Aviation and Aerospace Industries*. IGI Global, Hershey, PA, USA. <https://doi.org/10.4018/978-1-7998-1415-3>.
- Suau-Sanchez, P., Voltes-Dorta, A., Cugueró-Escofet, N., 2020. An early assessment of the impact of COVID-19 on air transport: just another crisis or the end of aviation as we know it? *J. Transport Geogr.* 86, 102749 <https://doi.org/10.1016/j.jtrangeo.2020.102749>.
- Tan, P., Steinbach, M., Kumar, V., 2013. *Introduction to Data Mining*, first ed. Pearson. Retrieved from <https://www.pearson.com/us/higher-education/product/Tan-Introduction-to-Data-Mining/9780321321367.html?tab=overview>.
- Tanrıverdi, G., Bakır, M., Merkert, R., 2020. What can we learn from the JATM literature for the future of aviation post covid-19? - a bibliometric and visualization analysis. *J. Air Transport. Manag.* 89, 101916 <https://doi.org/10.1016/j.jairtraman.2020.101916>.
- Tsa, J., Hung, C., 2021. Improving AdaBoost classifier to predict enterprise performance after COVID-19. *Mathematics* 9 (18). <https://doi.org/10.3390/math9182215>. Retrieved from.
- Urban, M., Klemm, M., Ploetner, K.O., Hornung, M., 2018. Airline categorisation by applying the business model canvas and clustering algorithms. *J. Air Transport. Manag.* 71, 175–192. <https://doi.org/10.1016/j.jairtraman.2018.04.005>.
- Vogel, H., Graham, A., 2013. Devising airport groupings for financial benchmarking. *J. Air Transport. Manag.* 30, 32. <https://doi.org/10.1016/j.jairtraman.2013.04.003>.

- Wen, C., Chen, T., Fu, C., 2014. A factor-analytic generalized nested logit model for determining market position of airlines. *Transport. Res. Pol. Pract.* 62, 71. <https://doi.org/10.1016/j.tra.2014.02.001>.
- Wen, C., Chen, W., 2011. Using multiple correspondence cluster analysis to map the competitive position of airlines. *J. Air Transport. Manag.* 17 (5), 302. <https://doi.org/10.1016/j.jairtraman.2011.03.006>.
- Zhang, P., Fan, C., Xu, Q., Ran, X., Yu, L., Fang, D., Zhang, Z., 2011. Applications of business intelligence technology in the airports and airlines companies. *Int. J. Appl.* 1 (5).
- Zhou, Z., 2003. Three perspectives of data mining. *Artif. Intell.* 143 (1), 139. [https://doi.org/10.1016/S0004-3702\(02\)00357-0](https://doi.org/10.1016/S0004-3702(02)00357-0).