Quality & Quantity
https://doi.org/10.1007/s11135-024-01956-0

A performance indicator and its decomposition according to the impacts of different aspects based on distributional data

Corrado Crocetta[1] · Antonio Irpino[2] · Laura Antonucci[3] · Claudia Marin[4]

Accepted: 5 August 2024

Abstract

In this paper, we present a novel approach to customer satisfaction analysis of airport services based on the analysis of distributional data for constructing a bivariate performance indicator. Distributional data was introduced for describing macro-data coming from the aggregation of micro-data observed at the individual level. We use them to represent the distribution of the ratings given by 165 classes (macro-units) of airport customers for twelve observed aspects. We describe the trend of passenger satisfaction over time by extracting 165 macro units from a survey conducted among 13,047 passengers at Bari and Brindisi airports during the peak and off-peak seasons of 2015, 2016 and 2017. To obtain a performance indicator, we performed a multiple factor analysis for distributional data. To our knowledge, no other methods exist for the factor analysis of multiple distributional variables. Further, we propose a new visualization tool called Green Eye Iris plot, which allows a joint visualization of our set of distributional values. The obtained results show that the distributional data analysis approach can provide valuable information at macro level that could be hidden when analyzing micro-data or when macro data are represented only by some features coming from summary statistics of groups.

Keywords Symbolic data analysis · Performance indicator · Distributional data

1 Introduction

In customer satisfaction analyzes, it is very important to compare data over time. To do this, you need a primary key that allows you to track the same person's responses over time. If you have anonymous questionnaires to simulate a cohort analysis, you might be interested in exploring typologies of customers rather than the individual customer to get some insights at the group level rather than the individual level. In this analytical framework, Symbolic Data Analysis (SDA) (Bock and Diday 2000) provides with statistical tools for studying groups of individuals (macro-units), defined according to a classification or a clustering process, through the use of the so-called *symbolic data*. Symbolic data

Antonio Irpino, Laura Antonucci and Claudia Marin have contributed equally to this work.

Extended author information available on the last page of the article

are multivalued descriptions (list or intervals of values, frequency distributions of qualitative or quantitative variables) that allow describing a class of individuals. Other concurrent methodologies to SDA (Brito and Dias 2022) have been proposed for analyzing distributional data from the compositional (Hron et al. 2016) and functional data analysis (Petersen and Müller 2016). Unfortunately, none of them provides tools for the analysis of multiple distributional ones. For this reason, the choice of tools proposed by the symbolic data analysis approach has been privileged in this context.

In the framework of symbolic data analysis, many statistical techniques first conceived for single-valued variables have been extended to the analysis of multivalued symbolic variables. Among them, factor analysis techniques for distributional data allowed us to reduce the number of variables to be considered and obtain a composite performance indicator. In a data exploration task, data visualization is a key factor for planning the analysis tasks, choosing the proper methodology, and discovering the presence of particular patterns in the data. Due to the complexity of the information carried by a symbolic description, there is a lack of visualization techniques for such data that allows capturing interesting patterns in a symbolic data set simply and intuitively. In this paper, we propose an innovative visualization tool for distributional data that was inspired to the shape and the colors of the iris of an eye, where situations of dissatisfaction or non-compliance with the norms can be easily identified for each macro-unit by using different color scales. The proposed visualization tool will be combined with the factor scores to improve the interpretation of the obtained results.

This paper aims to create a management dashboard consisting of a selected number of Key Performance Indicators (KPIs) able to measure the main evaluation dimensions and capture their changes over time. The analysis of the deviations makes it possible to immediately identify any risk situations and take corrective action by integrating the data obtained from the passenger satisfaction analyzes with those of the company information system. The proposed approach is particularly innovative compared to the existing literature (Sect. 2), since it uses distributional data analysis to obtain data simulating the existence of a cohort of respondents. The use of the Green Eye Iris plot, based on a polar system of coordinates, relates to similar approaches used by the Sant'Anna Institute in Pisa (Nuti et al. 2009) for analyzing healthcare performance and helps to see the results immediately and intuitively and to identify the risk areas where intervention is needed easily.

The paper is structured as follows: Sect. 2 provides a literature review on airport service analysis. Section 3 describes our research project. Section 4 presents the methodology, the distributional data analysis and the construction of a bivariate performance indicator through the factor analysis of distributional data. Section 5 shows the results of our analysis. Section 6 concludes the paper.

2 Literature review of the analysis of airport services

The development of global air traffic has increased the demand for airport services and the need for more efficient procedures to handle aircraft, passengers and baggage. Studies on airport operations and services are currently conducted from very different perspectives. Authors have used different methodologies to evaluate airport services. Fodness and Murray (2007) created a conceptual model of airport service quality by surveying nearly a thousand passengers who frequently use airport services. This allowed the authors to propose a set of recommendations for measuring airport service quality. The results of their research showed that business and leisure travelers have different opinions about the importance of the services provided and the level of airport operational efficiency. Lubbe et al. (2011) claimed that analyzing passengers' expectations regarding airport services is extremely important. Fernandes and Pacheco (2010) analyzed the quality of airport services using the methods of fuzzy multicriteria analysis and the alpha-cut concept. They used a complex set of quality variables and their indicators to obtain a comprehensive quality assessment, identify the cause-effect relationship and establish a quality standard.

In assessing the quality of airport services, some authors (Chou et al. 2011; Erdil and Yıldız 2011) developed criteria according to the classical dimensions of the Servqual method (touchability, responsiveness, reliability, safety, and empathy): Erdil and Yıldız (2011) assessed quality using 22 criteria, while Chou et al. (2011) added the flight pattern criteria group to the quality dimensions and used a set of 28 criteria. Sutia et al. (2013) analyzed the relationship between human capital, leadership, and strategic orientation with organizational performance, especially the impact of human capital investment on airport performance. Moreover, in contrast to the studies of other authors, the present study showed that airport ownership form and management strategy did not necessarily affect the growth of airport productivity, which is consistent with the results of Lin and Hong (2006). In 2016, da Rocha et al. (2016) proposed a multicriteria approach for the comparative analysis of the operational performance of Brazilian airport terminals. The relationship between an airport's service quality and passengers' behavioral intentions was also discussed by Prentice and Kadan (2019), who explored the synergy of these relationships and indicated whether airports should be considered elements of the tourism experience.

3 Research design and data

In Italy airport operators are required to draw up their own annual Service Charter, which sets out the overall levels of quality guaranteed at the airport in relation to the services offered directly or through the handling companies represented at the airport. In this way, the passenger can have helpful and understandable information about a particular type of service, even if the operator handles only part of it. The Service Charter is divided into 10 sections, 9 of which relate to quality aspects, each measured by one or more items whose responses are rated on a scale of 1–10 points. On the airport handler's website stakeholders can find the quality standard promised and observed, so that customers can compare the perceived quality with their expectations (ENAC 2014) (Table 1).

As required by the ENAC guidelines, the survey was conducted in different periods of the year choosing a typical week in the high and low seasons of the two airports of Bari and Brindisi. In a post-stratification procedure, the data were processed using expansion coefficients obtained by comparing the total number of departing passengers on a given day with the number of questionnaires collected on that day. To ensure better representativeness, it was decided to control for some factors, in particular, the airline used by the travelers surveyed, considering the main airlines: Ryanair, Alitalia and all other airlines. In 2015, 2016 and 2017, 13,047 passengers were surveyed, both in high season (summer) and low season (winter), as shown in the following Table 2.

Table 1 The sections of the questionnaire

Section	Description
A	Personal data (gender, age, education level, reasons for travel)
B	Security services (security checks for people and hand luggage, security of people and property at the airport)
C	Accuracy and punctuality of services (regularity and timeliness of services at the airport)
D	Cleaning and hygiene (cleanliness and functionality of restrooms, cleaning of the airport)
E	Comforts (availability of baggage carts, efficiency of passenger transfer systems, efficiency of air conditioning, general comfort)
F	Additional services (accessibility and reliability of Wi-Fi connections, availability and availability of recharging stations for cell phones or laptops, availability of vending machines for beverages and snacks, rating of restaurants and other shops)
G	Information services (effectiveness of information points, clarity and effectiveness of internal signage, professionalism of staff, updating and ease of reading website, accessibility and overall effectiveness of information)
H	Counter/gate services (ticket sales, waiting times at check-in counters, waiting times at security checkpoints)
I	Transportation network (clarity and effectiveness of external signage, adequacy of connections between the city and the airport)
L	Overall satisfaction (expected quality of airport services, Perceived quality of services used)

Table 2 Number of interviewed passengers

	Airport		
Season	Bari	Brindisi	Interviewed passengers
Winter 2015	1339	903	2242
Summer 2015	2724	1436	4160
Winter 2016	1202	649	1851
Summer 2016	784	587	1371
Winter 2017	926	791	1717
Summer 2017	768	938	1706
Total	7743	5304	13,047

4 Methods

4.1 Distributional data analysis

In classical data analysis, an input data table is provided where the rows represent the statistical units, and the columns are numeric or categorical variables. A classical data table is made of cells where each cell contains a number or a category, which is the measurement of the variable indicated in the column of the statistical unit (namely, a *micro-unit*) on the row. However, in many studies, the interest is to study groups of units whose description for a variable cannot be a single value but a multiple set of values that best synthesizes the group information without losing the inherent variability of the group. The treatment of such information originated the Symbolic Data Analysis (Bock and Diday 2000), where each statistical unit represents a group of individuals (namely, a *macro-unit*) that is described by a so-called Symbolic variable, a new concept of a variable whose realizations can be: intervals of numbers, sets of numbers, or categories, empirical frequency distributions having numeric or categorical support (also known as modal data).

When a unit is described by an empirical frequency distribution with numerical or categorical support, it is a *distributional data*. The analysis of distributional data is a recent approach to statistics derived from SDA, in which variables are referred to as *distributional variables*. A wide review of some of the most recent developments in this area of statistics was given by Brito and Dias (2022). There, one can find how classical statistical procedures such us basic statistics, regression, factor models, clustering, and classification techniques have been extended to analyze distributional data.

In this paper, a statistical unit is no longer considered as a micro-unit that assigns a numerical score to the evaluation of a service but as macro-unit, namely, a group of individuals, sharing some common characteristics (i.e., the airport and the season they were interviewed, the destination, the company, and the travel motivation) and, then, by the distribution of the score assigned to a particular aspect of the service. In this case, each aspect represents a distributional variable. Since each aspect is evaluated on a ten-point scale, we are dealing with *discrete distributional variables*. We considered 6 moments of customer satisfaction surveys, but the questionnaires collected are anonymous and do not allow comparison of passenger responses over time. For this reason, we identified 165 macro-units in which passengers were grouped based on some characteristics mentioned in Sec. 5.1. For each macro-unit, we obtained 6 different results that allowed us to study the temporal evolution of the phenomenon.

Let's suppose a data table as a $N \times P$ matrix, where N represents the number of statistical units and P is the number of considered variables, where each cell x_{ij}, $i = 1, \dots, N$ and $j = 1, \dots P$, contains the measure/value of the $j-th$ variable for the $i-th$ unit. A distributional data table is defined as follows:

$$X = [x_{ij}]_{N \times P}$$

where x_{ij} is an empirical frequency distribution. In particular, if all the P variables are discrete distribution functions x_{ij} is described as:

$$x_{ij} := D_j \rightarrow [0, 1]$$

where $D_j \in \mathbb{R}$ is a set of $K_j \in \mathbb{N}$ discrete values $d_{j1}, \dots, d_{jK_j}$ and $x_{ij} = [f_{ij1}, \dots, f_{ijK_j}]$ such that $\sum_{\ell=1}^{K_j} f_{ij\ell} = 1$, namely a set of empirical relative frequencies. In this paper, each cell contains a discrete empirical frequency function as in Table 3.

Most statistical methods and models for multivariate data analysis assume the definition of an appropriate distance function to compare data, measure dispersion, or obtain loss functions. In our case, the data are frequency distributions, so most methods for the analysis of distributional data require the definition of a distance between distributions. A wide range of distances and dissimilarities exist for comparing frequency distribution functions, which are derived from the distances or dissimilarities proposed for comparing probability distributions (Gibbs and Su 2002). Among them, the Wasserstein (Rüshendorff 2001) family of distances between probability distributions has shown interesting properties for the analysis of distributional data and the corresponding interpretation of the obtained results (Verde and Irpino 2008). In particular, the 2-Wasserstein distance, also known as Earth's Mover Distance (EMD), formed the basis of several statistical models for distributional

Table 3 A cell of a distributional data table where the distributional variable X_j has support in $D_j = \{1, \dots, 10\}$

Units / variables	…	X_j	…
…	…	…	…
Macro-unit i	…	D_j f_{ij} 1 0.00 2 0.03 3 0.02 4 0.05 5 0.15 6 0.30 7 0.25 8 0.10 9 0.04 10 0.01 1.00	…
…	…	…	…

data (Irpino and Verde 2006; Irpino et al. 2006, 2014; Irpino and Verde 2015; Verde et al. 2016; Verde and Irpino 2020). The L_2-Wasserstein distance between two distributional data can be viewed as a Euclidean distance between the quantile functions associated with each frequency distribution function, i.e., the inverse of the cumulative distribution function. A simple formulation of the L_2-Wasserstein distance between two density functions $f_1(x)$ and $f_2(x)$, having as cumulative distribution functions $F_1(x)$ and $F_2(x)$, and quantile functions $Q_1(p) = F_1^{-1}(p)$ and $Q_2(p) = F_2^{-1}(p)$ ($p \in [0, 1]$) is given as follows:

$$d_{\mathcal{W}}(f_1,f_2) := \sqrt{\int_0^1 |Q_1(p) - Q_2(p)|^2 dp}.$$

Without loss of generality, it can also be computed between two discrete distributions (Nguyen 2013).

4.2 Exploratory analysis and construction of composite indicators through the factor analysis of distributional data

The input data matrix contains the aggregated information about the macro-units considered in the study. In particular, twelve distributional variables are considered.

First, we introduce a new visualization tool termed *Green Eye Iris* plot (GEI), allowing the joint visualization of a set of distributional values measured on a macro-unit (Fig. 1).

The GEI plot of a macro-unit described by P distributional variables is essentially a stacked percentage barchart represented in polar coordinates. The main steps for generating a GEI plot are as follows:

- For each distributional value, we generate a stacked percentage bar chart where each part of the bar is proportional to the relative frequency of each element of the support. The lowest (in our case, 1) to the highest (in our case, 10) value of support is associated with a filling color scale, ranging from a dark red to a dark green hue.

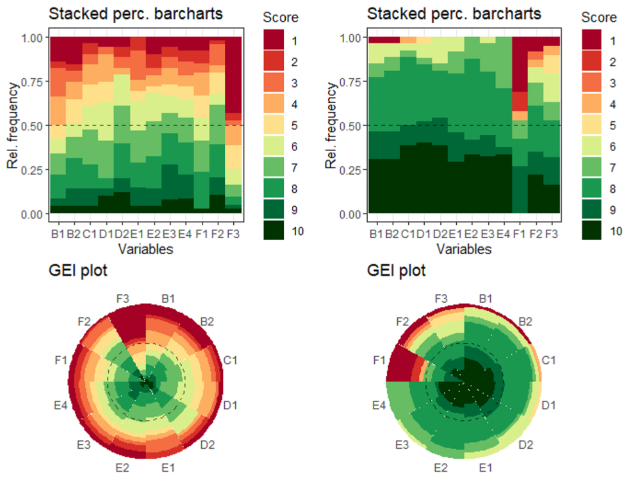


Fig. 1 Two macro units described by two stacked percentage barcharts and the corresponding GEI plots. On the left one can see a macro unit with low scores, while on the right, a macro unit with high scores. We considered a set of variables as selected in Sect. 5.1

- The P distributional values, one for each variable, allow us to obtain P stacked percentage barcharts whose tiles are arranged in reverse scale such that the highest values are positioned at the bottom of the plot and the lowest at the top of it.
- A polar coordinate system is used. Each stacked barplot referred to a distributional variable represents a sector such that the angle is equal to $\frac{360°}{P}$. Each sector, is divided from the center outward, accordingly to the relative frequency of each value, but, in this case the values are in reverse order (at the center values are the highest). In this way, each stacked barplot represents a sector of a circle. The order of the distributional values can be chosen in advance by the user accordingly to some apriori knowledge.

The GEI plot in Fig. 1 summarize the information related to our 12 variables at a time. The plot can be perceived as pleasant when it is completely green, while it is perceived as negative when it goes toward the red. When a person sees the iris of an eye, it is more attractive when it is completely green, while they feel bad when the iris is red. Note that, in this case, we reversed the order of the values for emphasizing the presence of low scores, which are more evident if they occupy an external position with respect to the center because the size of external subsectors appear greater. The distorsion in size introduced by the polar coordinate plot, even if in general is a disavantage, in this case is useful to the user that can be interested more on low scores frequencies than on high scores, which is typical in the exploration of items related to quality satisfaction. We have enriched the GEI plot using a dotted circle representing the 50% (namely, the level of the frequency distribution representing the median).[1] In Fig. 1, we show an example of two GEI plots representing, respectively, two macrounits with low and high ratings for the variables selected in Sect. 5.1.

[1] Other glyphs can be introduced in the plot for considering information about the skewness of each distribution, but we don't use it here for avoiding an overload of information represented in the plot.

4.3 Extracting composite performance indicators using factor methods for distributional data

Composite indicators are constructed by combining variables into a single score or index. Factor extraction methods, such as Principal Component Analysis (PCA), are often used to construct formative composite indicators because they allow us to reduce the dimensionality of the data and identify the underlying patterns or dimensions that explain most of the variation in the original variables.

Recently, factor analysis methods extended to distributional data have been proposed (Verde et al. 2016; Verde and Irpino 2020) in the framework of SDA. In the current paper, we considered 12 distributional variables and used an extension of the classical Multiple Factor Analysis (MFA) (Escofier and Pagès 1994) according to the MFA for distributional variables proposed by Verde and Irpino (2020). MFA builds upon PCA and produces a set of common factors that can be used to project data that is characterized by multiple sets of variables onto a common subspace, allowing for a compromise solution. The method proposed by Verde and Irpino (2020) assumes that each distributional variable represents a block of columns each of them described by a predefined set of quantiles (usually 25 are sufficient for describing each distribution).

Let $\mathbf{x}_i$ be the set of P distributions x_{ij} (for $j = 1, \ldots, p$) describing the i-th macro-unit with respect to the P variables. We fix in advance a number q of quantiles (usually $q \geq 25$ is sufficient) such that each distribution x_{ij} will be coded into a vector $\mathcal{Q}_{ij}$of $q + 1$ values corresponding to quantiles associated as follows:

$$\mathcal{Q}_{ij} = \left[Q_{ij}(0), Q_{ij}\left(\frac{1}{q}\right), Q_{ij}\left(\frac{2}{q}\right), \ldots, Q_{ij}\left(\frac{q-2}{q}\right), Q_{ij}\left(\frac{q-1}{q}\right), Q_{ij}(1) \right].$$

The MFA will have as input a column-wise block matrix as follows

$$\mathcal{Q} = [\mathcal{Q}_1 | \mathcal{Q}_2 | \ldots | \mathcal{Q}_P] \tag{1}$$

having N rows and $P \cdot (q + 1)$ columns. The classical MFA algorithm (Escofier and Pagès 1994) is performed on centered quantiles only in order to preserve the Wasserstein-based variance of each distributional variable (for further details, see Verde et al. 2016; Verde and Irpino 2020).

In the first step of MFA, a PCA for each block is performed, then each block is standardized by the first corresponding eigenvalue. A second PCA is then performed on the standardized data and, after fixing the number of retained components $\alpha \leq P$, matrices of $P \cdot (q + 1) \times \alpha$ loadings, and $N \times \alpha$ scores are obtained.

5 Results

5.1 The distributional data

Based on the 13, 047 interviews (as reported in Table 2), we grouped the interviewed passengers according to the combination of the following characteristics:

Airport: Bari, Brindisi;

Season: from low season 2015 to high season 2017;
Destination: Rome, Milan, Other italian destinations, International;
Flight Company: Alitalia, Ryanair, Other airlines;
Motivation: Leisure, Business.

Groups with fewer than 20 passengers were not included, giving us 165 macrounits.

Each unit is described by the frequency distribution for each of the following items:

B. Security services.

B1 : luggage screening service and personal safety;
B2 : property protection.

C. Accuracy and punctuality of services;

C1 : overall perception of accuracy and punctuality.

D. Cleaning and hygiene.

D1 : toilets cleanliness;
D2 : overall cleanliness of the airport.

E. Comforts.

E1 : luggage trolley availability;
E2 : efficiency of the system transfer passengers.
E3 : efficiency of the air conditioning system;
E4 : perception of the overall comfort level of the airport facility.

F. Additional services.

F1 : wifi service;
F2 : vending machine availability;
F3 : retrievability of seats for charging phones/laptops.

The other items had a high proportion of missing values and were not considered. The analysis was performed on 12 distributional variables observed for 165 macrounits.

In classical data analysis, a numerical variable can be summarized by a single value, i.e., the average value. When data are distributions, we speak of a barycenter represented by a distribution that has the smallest distance between all other distributions. Figure 2 shows the bar diagrams constructed for each variable, representing the intermediate distributions of the 12 distribution variables that preserve the common features of all distributions. For example, B1, the item regarding hand luggage control, is represented by a discrete distribution that has a skewness of −0.86 and an average of 7.37 with a standard deviation of 1.77. $F1$, the item related to wifi service, does not appear to have significant patterns in the data. In fact, Table 4 shows that it is the one with the highest Wasserstein standard deviation (i.e., a high diversity between the distributions observed for the units) and a high standard deviation for the Wasserstein mean (i.e., high variability of its representative). Wasserstein means are obtained by using the

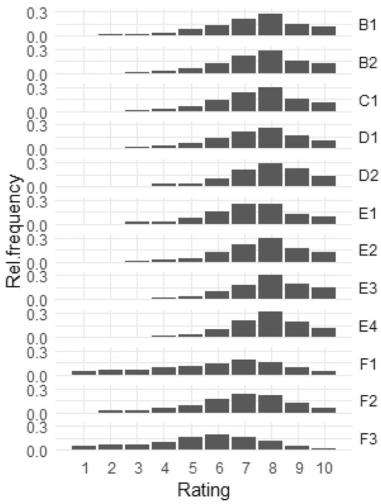


Fig. 2 Wasserstein means of distributional variables

Table 4 Wasserstein basic statistics for each variable and for the mean distributions

Variable	Wass. means statistics			Skewness
	Mean	Median	St. Dev	
B1	7.369	8	1.762	−0.862
B2	7.543	8	1.627	−0.814
C1	7.498	8	1.549	−0.810
D1	7.395	8	1.705	−0.879
D2	7.857	8	1.443	−0.805
E1	7.269	7	1.639	−0.805
E2	7.618	8	1.541	−0.817
E3	7.890	8	1.484	−0.868
E4	7.761	8	1.451	−0.863
F1	5.983	6	2.379	−0.749
F2	6.869	7	1.861	−0.836
F3	5.603	6	2.183	−0.645

approach proposed in Irpino and Verde (2015) and in Brito and Dias (2022)(Chap. 3), which is based on L_2-Wasserstein distance.

The Wasserstein correlation and covariance matrix As can be seen from Table 5, there is a positive correlation/association between all the considered variables, even if they do not have particularly high values. Covariances and correlations between distributional variables have been obtained following the approach proposed in Irpino and Verde (2015) and in Brito and Dias (2022)(Chap. 3), which is based on L_2-Wasserstein distance, too.

Table 5 Variance–covariance–correlation matrix

	B1	B2	C1	D1	D2	E1	E2	E3	E4	F1	F2	F3
B1	**0.596**	0.718	0.547	0.482	0.448	0.455	0.475	0.461	0.485	0.069	0.211	0.409
B2	*0.409*	**0.544**	0.574	0.507	0.516	0.449	0.541	0.532	0.577	0.065	0.253	0.396
C1	*0.275*	*0.275*	**0.423**	0.507	0.510	0.455	0.497	0.500	0.523	0.064	0.290	0.317
D1	*0.256*	*0.257*	*0.227*	**0.472**	0.590	0.448	0.469	0.513	0.537	0.111	0.262	0.316
D2	*0.221*	*0.243*	*0.212*	*0.259*	**0.408**	0.400	0.509	0.577	0.582	0.037	0.312	0.284
E1	*0.226*	*0.213*	*0.191*	*0.198*	*0.165*	**0.415**	0.500	0.461	0.466	0.094	0.259	0.311
E2	*0.248*	*0.269*	*0.218*	*0.217*	*0.219*	*0.217*	**0.455**	0.578	0.605	0.033	0.278	0.385
E3	*0.237*	*0.261*	*0.216*	*0.234*	*0.245*	*0.198*	*0.259*	**0.442**	0.663	0.013	0.269	0.320
E4	*0.250*	*0.283*	*0.227*	*0.246*	*0.247*	*0.200*	*0.272*	*0.294*	**0.444**	0.051	0.268	0.350
F1	*0.054*	*0.049*	*0.042*	*0.078*	*0.024*	*0.062*	*0.023*	*0.009*	*0.035*	**1.045**	0.216	0.205
F2	*0.122*	*0.140*	*0.141*	*0.135*	*0.149*	*0.125*	*0.141*	*0.134*	*0.134*	*0.166*	**0.563**	0.289
F3	*0.304*	*0.280*	*0.198*	*0.209*	*0.174*	*0.193*	*0.250*	*0.205*	*0.224*	*0.201*	*0.208*	**0.925**

Wasserstein variances (in bold) for each distributional variable are reported on the main diagonal. The elements outside the diagonal represent the covariances (in italics) on the lower triangular part and the correlations on the upper triangular part

Table 6 MFA first 15 eigenvalues and explained variance

Comp.	Eigenvalue	% of variance	Cumulative % of variance
1	7.75	34.38	34.38
2	2.13	9.45	43.84
3	1.24	5.51	49.35
4	0.91	4.06	53.40
5	0.66	2.94	56.35
6	0.61	2.71	59.05
7	0.52	2.30	61.35
8	0.44	1.95	63.30
9	0.40	1.75	65.05
10	0.33	1.44	66.50
11	0.30	1.32	67.82
12	0.29	1.29	69.11
13	0.27	1.19	70.30
14	0.26	1.14	71.44
15	0.23	1.04	72.48

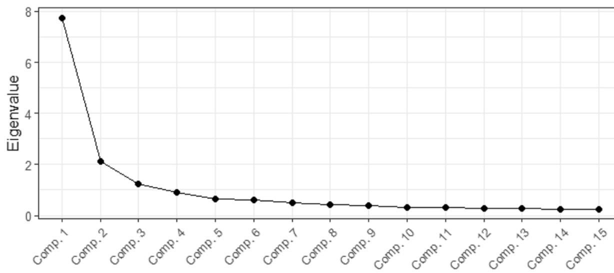


Fig. 3 The scree plot of the MFA. Only the first 15 eigenvalues of 164 are considered

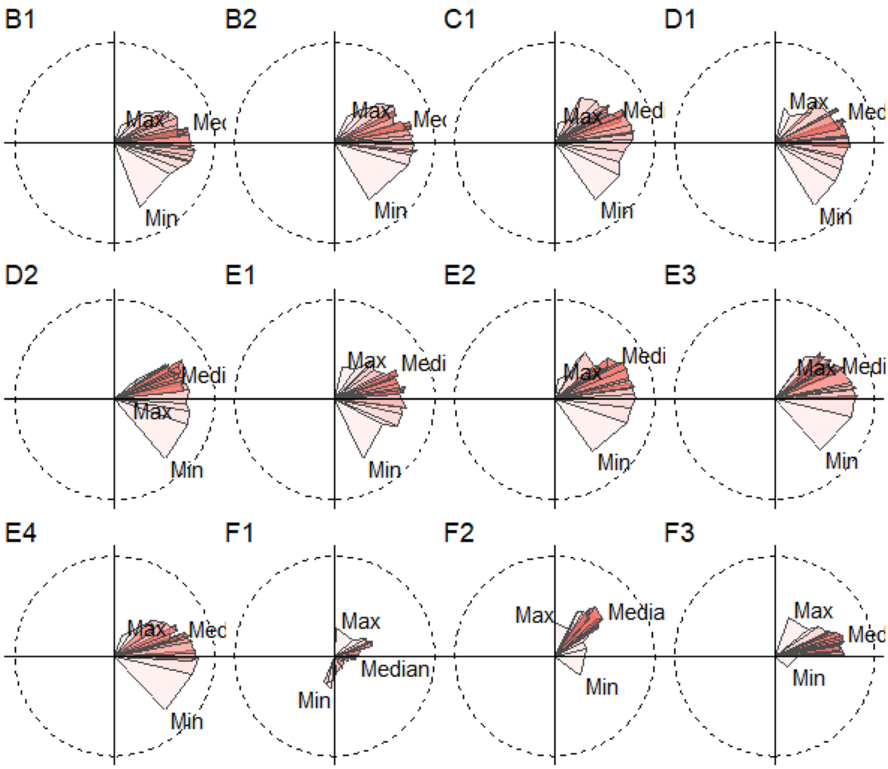
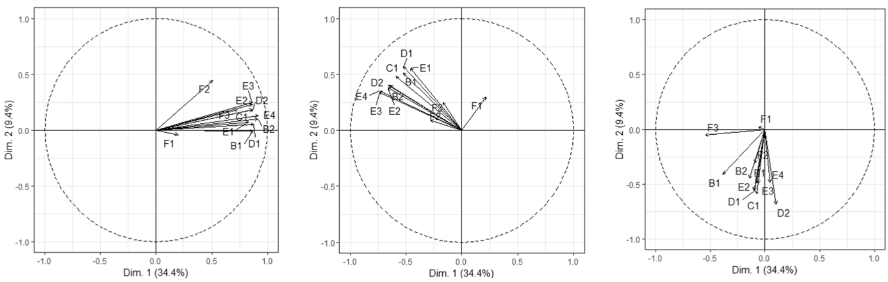


Fig. 4 The Spanish-fan plots for each distributional variable on the first factorial plane. The dashed circle represents the classical unit circle correlation bound of PCA-like methods

5.2 The MFA output

We performed an MFA on a $165 \times (12 \cdot 26)$ matrix $\mathcal{Q}$ (as in Eq. 1), where each distributional variable is represented by a block of $q = 25 + 1$ columns $\mathcal{Q}_j$ $j = 1, \ldots, 12$ containing the 25 quantiles (plus the 0-th quantile which represents the minimum) of each



(a) Means (b) Standard deviations (c) Skewness

Fig. 5 Correlation plots of means, standard deviations, and skewness indices with respect to the first two dimensions extracted from the MFA

distribution.[2] Each set of 26 columns is referred to each item and represent a block in the analysis.

Table 6 reports the first 15 eigenvalues of the MFA. We reported the percentage of explained variance and the respective cumulative percentage, too.

The screeplot From the analysis of the scree plot associated with the eigenvalues extracted from the MFA, in Fig. 3, and according to the elbow method of selection, we retain only the first two components, which synthesize 43.8% of the total variance.

Plot of variables: the Spanish-fan plots The variables are represented by the Spanish-fan plots, proposed by Verde and Irpino (2020). Figure 4 shows the correlation between the quantiles of each distributional variable and the first two dimensions extracted from the MFA. We recall that the Spanish-fan plot of a single distributional variable projected on the first factorial plane is constructed by connecting each quantile-column vector and coloring it to look like the familiar Spanish fans.

The shape of the fans suggests some peculiar patterns for the interpretation of the ditributions.

The first axis shows units on the left with generally low ratings, while on the right are positioned units with a high rate for all the quantiles. The second axis is mainly related to a left (on the top) versus a right (on the bottom) skewness of distributions and for high scores for the variable $F2$.

Such patterns are also corroborated by the analysis of the contributions of the quantiles to the axes (see supplementary information provided in Sect. 7).

The interpretation of the dimensions is then:

- Dimension 1, which explains the 34.4% of variability: from left to right, the units are ranked by their average score from not very satisfied to very satisfied for almost all the variables (except $F1$).
- Dimension 2, which explains the 9.4% of variability: from top to bottom, units with lower skewness than the mean skewness (see Table 4) are ranked against units with higher skewness. It means that points on the top of the plane correspond to units

[2] We chose 25 quantiles because increasing the number of quantiles the results of the MFA are substantially the same.

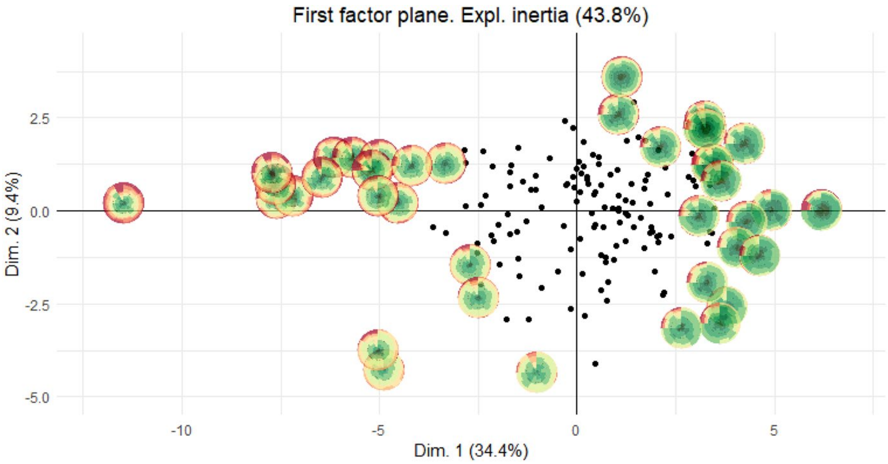


Fig. 6 MFA first factor plane. GEI plots are shown for those individuals with a quality of representation above 0.5 (squared cosines) on the plane

associated with more left-skewed distributions than the average. Also, the units with an increasing average value for variable $F2$ are shown from bottom to top.

The above patterns become clearer when looking at Fig. 5, which shows correlation plots of the means, standard deviations, and Fisher's skewness indices of distributions with respect to the first two dimensions extracted from MFA.

Since the measurement scales of each original variable are 10-point scales, we may observe some natural patterns in the data. For example, the more the average score provided for each variable increases, the more the corresponding standard deviation should decrease. Actually, the maximum standard deviation observable for a 10-point scale variable[3] is equal to $\sqrt{\left(1^2 + 10^2\right)0.5 - 5.5^2} = 4.5$. As shown in Fig. 5, this relationship holds for the data. In fact, the vectors related to the standard deviations have a slightly opposite direction with respect to the mean vectors. The second dimension also reveals an interesting pattern. As we saw in Table 4, the Wasserstein mean distribution of each variable has a negative (left) skewness, an aspect that is very common in customer satisfaction surveys (Peterson and Wilson 1992). Following the vertical direction (from the bottom upwards) of the first factorial plane, almost all distributions become less left-skewed and show a tendency towards symmetry. To catch this pattern, one must consider that when a set of distributions is heavily left-skewed the GEI plots appear with a higher presence of yellow and red color. In our application, it appears that the red on green ratio decreases from the bottom upwards and this suggests that the sets of distributions on the top of the plane have a lower proportion of low scores (under the condition that the compared GEI plots are horizontally close).

[3] It is easy to prove that the maximum standard deviation observable for a random variable with support bounded by $[a, b]$ is equal to $\sqrt{\left(a^2 + b^2\right)0.5 - \left(\frac{a+b}{2}\right)^2}$.

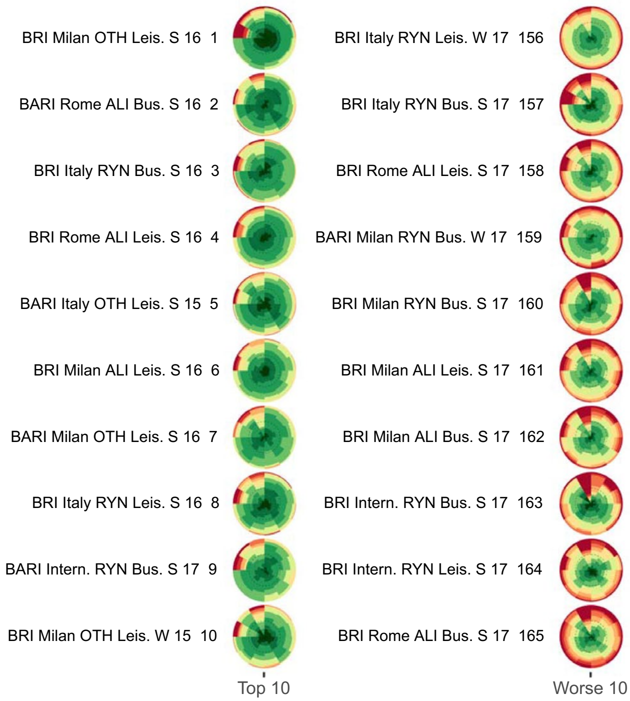


Fig. 7 GEI plots of the top and worse 10 units, ranked accordingly to the "Average score" Ap_i indicator. The label of each plot indicates the airport, the destination, the flight company, the motivation, the season and the rank position

Plot of individuals The above conclusions about the first two dimensions seem clearer when we look at Fig. 6, in which the GEI plots of those units whose square cosines is greater than 0.5 are overlapped to the points. Following a north-east direction in reading the plot reveals units with a higher average score and a more symmetric distribution score for each variable (dark green zones are opposite the medium yellow ones about in roughly equal proportions).

Scores According to the results of MFA, we will use the first two dimensions to measure the performance of the 165 units. The first indicator is associated with the first factor and represents an "Average score" of performance, while a second indicator, associated with the second dimension, is considered a "propensity toward symmetry" indicator. As for the first indicator, it is straightforward that it provides information about the average level of service. Another interesting aspect is that the first dimension is also associated with the decrease of the standard deviation moving from left to right: the higher the mean level of the service the more the users return concordant scores (namely, the distributions with a higher mean have a lower standard deviation). More interesting is the second indicator, which allows one to identify if, independently from the average score, the users scored the service in a more symmetric way, namely, letting the mean score be representative of a central tendency. We recall that, in our case, the origin of the axis in Fig. 6 is represented by the mean (in the sense of Wasserstein) the distributions shown in Fig. 2, where all the distributions are left-skewed (as reported in Table 4).

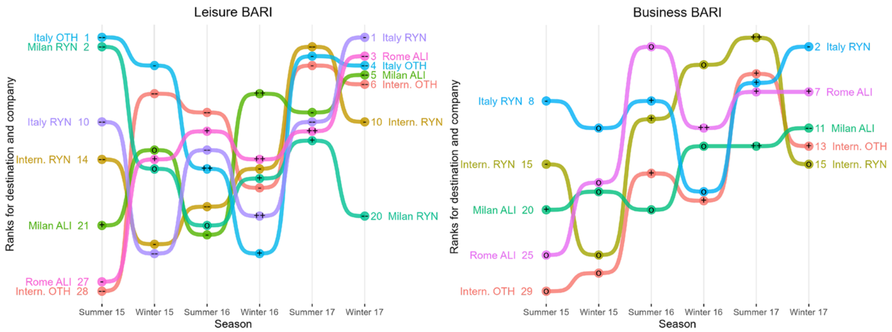


Fig. 8 Bumping plots of the ranks per season of those units observed at the *Bari* airport for *leisure* and *business* motivation along the six seasons for the first dimension Ap_i enriched with symbols representing the quantile of the second dimension Sy_i: "–" for strong left, "-" moderate left, "o" average, "+" moderate right and "++" rightest mean skewness

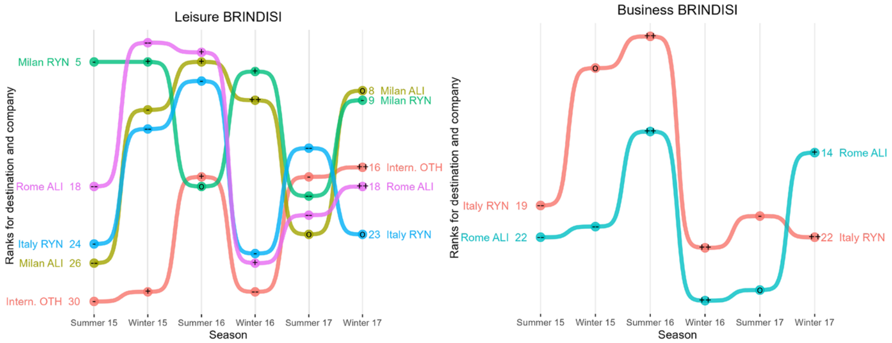


Fig. 9 Bumping plots of ranks per season of those units observed at the *Brindisi* airport for *leisure* and *business* motivations along the six seasons for the first dimension Ap_i enriched with symbols representing the quantile of the second dimension Sy_i: "–" for strong left, "-" moderate left, "o" average, "+" moderate right and "++" rightest mean skewness

We propose a bivariate performance indicator

$$P_i = (Ap_i, Sy_i) \tag{2}$$

using the standardized scores of the MFA units for the first dimension (Ap_i, Average Performance of the $i-th$ unit) and the second one (Sy_i, Symmetry score of the $i-th$ unit).

In Fig. 7, using the GEI plots, we see the best and the worst ranking units for the Ap_i standardized score related to the "Average score". We remark that the first top units are generally associated with the 2016 summer season, while the worse units are associated with both the 2017 winter and summer seasons.

The longitudinal analysis To provide a straightforward interpretation of both dimensions over time for the considered units, we propose to use *bump charts* for each airport and motivation-related unit. A bump chart is a visualization chart typically used in business analytics tools that looks like bumps in the road. It represents an alternative for time series analytics over rank for charts like a line chart. The proposed bump chart is related to the Ap_i and are enriched with some glyphs providing information about the Sy_i indicator: "–" for strong left, "-" moderate left, "o" average, "+" moderate right and "++" rightest mean skewness. We show the main results in Figs. 8 and 9. From the plots, we may observe that the rankings of the units related to Bari airport have generally improved in the last two seasons and that this improvement is also accompanied by a moderate improvement in the mean skewness of the distributions, especially for the business-related ones. The last consideration shows that the business-related units seem to have more symmetrical patterns in the observed distributions. As for Brindisi airport, in the first two seasons there is an improvement in the ranking of units and in the mean skewness of the distibution for both leisure and business travelers.

The distributional approach would lead to further detailed analysis. For example, by exploiting the properties of the Wasserstein-based analysis of distributional data coming from the optimal transportation theory (Villani 2009), it would be possible to reveal how distributions are changed over time or to explain differences between distributions using, for example, transportation maps, but, for the sake of brevity, we will not consider carrying such further analysis.

6 Conclusion

In this paper, we proposed a methodology of analysis of macrodata, i.e., data derived from aggregating microdata (at the individual level) into distributions describing groups of individuals. Such groups of individuals can be viewed as segments of a market.

Because we had anonymous questionnaires collected over 6 different time periods and there was no primary key that would have allowed us to track cohorts over time, we identified subgroups of respondents that we considered complex statistical units for which we observed trends across the 6 surveys we conducted. This innovative approach allowed us to transform our study into a cohort analysis, compared to the usual methods of measuring customer satisfaction described in Sect. 2, and to measure the satisfaction of specific groups of travelers.

We introduced a novel visualization for units described by a set of numerical distributions: the GEI plot. This representation allows to intuitively identify areas of improvement for the different aspects considered, and provides a dashboard of KPIs that can analyze a complex phenomenon such as that of passenger satisfaction, highlighting the latent variables that most influence the overall satisfaction of travelers. The comparison over time between the 6 available collection points allows to follow the evolution of the phenomenon over time and to immediately identify the most vulnerable situations.

We showed how a factor analysis technique extended to distributional data may provide useful information which is difficult to observe in the analysis of classical single-valued data.

We presented an application in the framework of customer satisfaction for services offered at two Italian (Apulian) airports and derived a bivariate performance indicator able to account for the different sources and types of variability carried by the distributions.

Bumping plots of ranks allowed us to track the satisfaction of different categories of travelers over time. For reasons of synthesis, we analyzed only some of the 165 groups considered by distinguishing the overall satisfaction of the indicators according to the asymmetry of the distributions. The analysis showed a slight improvement in quality standards at Bari airport, while the situation at Brindisi airport is more stationary. The results summarized in graphical form allowed us to capture the changes that occurred over time in the KPIs analyzed. The comparison between the 2 airports considered did not reveal any significant differences.The business intelligence visualization tools used allowed the synthesis of very complex phenomena and their monitoring over time, responding to the need to measure continuous improvement as required by Total Quality Management.

7 Supplementary information

The MFA input data and the analysis with all the detailed and intermediate results are freely available in a Github at the following URL: https://github.com/Airpino/Air_customer. The analysis was conducted using the R software and all the code is available for replicability issues at the aforementioned URL.

Funding Open access funding provided by Università degli Studi della Campania Luigi Vanvitelli within the CRUI-CARE Agreement. This study was carried out within GRINS - Growing Resilient, Inclusive and Sustainable and received funding from the European Union Next-GenerationEU (NATIONAL RECOVERY AND RESILIENCE PLAN NRRP). MISSION 4, COMPONENT 2, INVESTMENT 1.3 - D.D. 1558 11/10/2022, PE00000018 Spoke 8).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This manuscript has not been published anywhere and is not being considered for publication elsewhere. This manuscript reflects only the authors'view and opinions, neither the European Union nor the European Commission can be considered Responsible for them.

References

Bock, H.H., Diday, E.: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer, Cham (2000)

Brito, P., Dias, S.: Analysis of Distributional Data. Chapman and Hall/CRC, Boca Raton (2022)

Chou, C.-C., Liu, L.-J., Huang, S.F., Yih, J.-M., Han, T.-C.: An evaluation of airline service quality using the fuzzy weighted servqual method. Appl. Soft Comput. **11**(2), 2117–2128 (2011). https://doi.org/10.1016/j.asoc.2010.07.010

da Rocha, P.M., de Barros, A.P., da Silva, G.B., Costa, H.G.: Analysis of the operational performance of Brazilian airport terminals: a multicriteria approach with de borda-ahp integration. J. Air Transp. Manag. **51**, 19–26 (2016). https://doi.org/10.1016/j.jairtraman.2015.11.003

ENAC: Qualitá dei servizi nel trasporto aereo: le carte dei servizi standard per gestori aeroportuali e vettori aerei. Retrieved from https://www.enac.gov.it/sites/default/files/allegati/2018-Lug/GE 06.pdf (In Italian only) (2014)

Erdil, S.T., Yıldız, O.: Measuring service quality and a comparative analysis in the passenger carriage of airline industry. Proc. Social Behav. Sci. **24**, 1232–1242 (2011). https://doi.org/10.1016/j.sbspro.2011.09.117

Escofier, B., Pagès, J.: Multiple factor analysis (afmult package). Comput. Stat. Data Anal. **18**(1), 121–140 (1994). https://doi.org/10.1016/0167-9473(94)90135-X

Fernandes, E., Pacheco, R.R.: A quality approach to airport management. Qual. Quant. **44**(3), 551–564 (2010). https://doi.org/10.1007/s11135-008-9212-9

Fodness, D., Murray, B.: Passengers' expectations of airport service quality. J. Serv. Mark. **21**(7), 492–506 (2007). https://doi.org/10.1108/08876040710824852

Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. Int. Stat. Rev. **70**(3), 419–435 (2002). https://doi.org/10.1111/j.1751-5823.2002.tb00178.x

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P.: Simplicial principal component analysis for density functions in bayes spaces. Comput. Stat. Data Anal. **94**, 330–350 (2016). https://doi.org/10.1016/j.csda.2015.07.007

Irpino, A., Verde, R.: A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Batagelj, V., Bock, H., Ferligoj, A., Žiberna, A. (eds.) Data Science and Classification, pp. 185–192. Springer, Berlin (2006)

Irpino, A., Verde, R.: Basic statistics for distributional symbolic variables: a new metric-based approach. Adv. Data Anal. Classif. **9**(2), 143–175 (2015). https://doi.org/10.1007/s11634-014-0176-4

Irpino, A., Verde, R., de Carvalho, F.A.T.: Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Syst. Appl. **41**(7), 3351–3366 (2014). https://doi.org/10.1016/j.eswa.2013.12.001

Irpino, A., Verde, R., Lechevallier, Y.: Dynamic clustering of histograms using Wasserstein metric. Compstat 2006, pp. 869–876, Physica-Verlag (2006)

Lin, L.C., Hong, C.H.: Operational performance evaluation of international major airports: an application of data envelopment analysis. J. Air Transp. Manag. **12**(6), 342–351 (2006). https://doi.org/10.1016/j.jairtraman.2006.08.002

Lubbe, B., Douglas, A., Zambellis, J.: An application of the airport service quality model in South Africa. J. Air Transp. Manag. **17**(4), 224–227 (2011). https://doi.org/10.1016/j.jairtraman.2010.08.001

Nguyen, X.: Convergence of latent mixing measures in finite and infinite mixture models. Anna. Stat. **41**(1), 370–400 (2013)

Nuti, S., Bonini, A., Murante, A.M., Vainieri, M.: Performance assessment in the maternity pathway in Tuscany region. Health Serv. Manag. Res. **22**(3), 115–121 (2009). https://doi.org/10.1258/hsmr.2008.008017

Petersen, A., Müller, H.-G.: Functional data analysis for density functions by transformation to a Hilbert space. Anna. Stat. **44**(1), 183–218 (2016). https://doi.org/10.1214/15-AOS1363

Peterson, R.A., Wilson, W.R.: Measuring customer satisfaction: fact and artifact. J. Acad. Mark. Sci. **20**(1), 61–71 (1992). https://doi.org/10.1007/BF02723476

Prentice, C., Kadan, M.: The role of airport service quality in airport and destination choice. J. Retail. Consum. Serv. **47**, 40–48 (2019). https://doi.org/10.1016/j.jretconser.2018.10.006

Rüshendorff, L.: Wasserstein Metric, Encyclopedia of Mathematics, Springer, p. 21 (2001)

Sutia, S., Sudarma, M., Djumahir, R.: The influence of human capital investment, leadership and strategic orientation on airport performance. Int. J. Bus. Manag. Invent. **2**(6), 26–32 (2013)

Verde, R., Irpino, A.: Dynamic clustering of histogram data: using the right metric. In: Brito, P.E.A. (ed.) Selected Contributions in Data Analysis and Classification, pp. 123–134. Springer, Berlin (2008)

Verde, R., Irpino, A.: Multiple factor analysis of distributional data. Stat. Appl. Italian J. Appl. Stat. **29**(23), 305–330 (2020). https://doi.org/10.26398/IJAS.0029-017

Verde, R., Irpino, A., Balzanella, A.: Dimension reduction techniques for distributional symbolic data. IEEE Trans. Cybern. **46**(2), 344–355 (2016). https://doi.org/10.1109/TCYB.2015.2389653

Villani, C.: Optimal Transport: Old and New. Springer, Berlin (2009)

Authors and Affiliations

Corrado Crocetta[1] · Antonio Irpino[2] · Laura Antonucci[3] · Claudia Marin[4]

✉ Antonio Irpino
antonio.irpino@unicampania.it

Corrado Crocetta
corrado.crocetta@uniba.it

Laura Antonucci
laura.antonucci@unifg.it

Claudia Marin
claudia.marin@uniba.it

1 Department of Research and Humanistic Innovation, University of Bari, Piazza Umberto I, 70121 Bari, BA, Italy

2 Department of Mathematics and Physics, University of Campania “L. Vanvitelli”, Viale A. Lincoln, 5, 81100 Caserta, CE, Italy

3 Department of Clinical and Experimental Medicine, University of Foggia, Viale Luigi Pinto, 71122 Foggia, FG, Italy

4 Department of of Education, Psychology, Communication, University of Bari, Via Crisanzio, 42, 70121 Bari, BA, Italy