Networks and Spatial Economics
https://doi.org/10.1007/s11067-024-09624-4

RESEARCH

Spatiotemporal Analysis of Traffic Data: Correspondence Analysis with Fuzzified Variables vs. Principal Component Analysis Using Weather and Gas Price as Extra Data

Pierre Loslever[1,2]

Accepted: 21 March 2024

Abstract

Study of large rail traffic databases presents formidable challenges for transport system specialists, more particularly while keeping both space and time factors together with the possibility of showing influencing factors related to the users and the transport network environment. To perform such a study, a bibliographic analysis in both statistics and transport revealed that geometrical methods for feature extraction and dimension reduction can be seen as suitable. Since there are several methods/options with, in principle, required input data, this article aims at comparing Principal Component Analysis (PCA) and Correspondence Analysis (CA) for traffic frequency data, both methods being actually used with such data. The procedure stands as follows. First a grand matrix is built where the rows correspond to time windows and the columns to all the possible origin-destination links. Then this large frequency matrix is studied using PCA and CA. The next part of the procedure consists in studying the effects of influencing factors with the possibility of keeping the quantitative scales with PCA or using fuzzy segmentation with CA, the corresponding data being considered as supplementary column points. The procedure is applied on a rail transport network including 10 stations (one corresponding to the airport) and one-hour time windows for 4 months, the available influencing factors being the temperature, rain level and gas price. The comparative analysis shows that CA graphical outputs are more complicated than PCA ones, but reveal more specific results, e.g. the network user behavior related to the airport, while PCA mainly opposes link clusters with low vs. high frequencies. Fuzzy windowing performed using actual and simulated data reduces the loss of information when averaging, e.g. over time, and can show non-linear relational phenomena. The possibility of displaying new traffic data in real time is also considered.

Keywords Traffic analysis · Origin-destination matrix · Correspondence analysis · Principal component analysis · Economics · Weather · Fuzzy sets

Extended author information available on the last page of the article

Published online: 03 May 2024

1 Introduction

Transport network studies yield many kinds of variables making it difficult to analyse them all simultaneously once the corresponding data sets have been collected. First, let us focus on the diversity of data that can be collected.

Firstly, this diversity comes from the data origin i.e. either the human being or not. In the first origin or modality, we can mention a data piece given by anyone concerned in the transport network, including the *user* (the pedestrian, cyclist, car driver or anyone using public transit) and *the person working on the network* (the aircraft pilot, the bus driver, the car maintenance technician, train safety engineer, subway traffic manager), keeping in mind that the behaviour of these persons can be more or less impacted by factors as the weather (Nosal and Miranda-Moreno 2014), strike (Exel and Rietveld 2009), traffic congestion or stress (Khoo and Asitha 2016; Cabral and Kim 2022), dysfunction (Hong et al. 2019) or accident (Ryder et al. 2019). With network users, data is often recorded through a survey questionnaire (Exel and Rietveld 2009; Schmöcker et al. 2010; Truong and Somenahalli 2015; McCarthy et al. 2016; Khoo and Asitha 2016; Ahmed et al. 2021; Yin and Leurent 2022). Individual data can be related to absolute or relative assessment (e.g. perception of the security, preference for a mode of transport, off-hour delivery types, location privacy, level of cycling comfort) (Cottrill and Thakuriah 2015; Marcucci and Gatta 2017; Cabral and Kim 2022), or to more general personal characteristics (gender, age, job, dwelling place, household size, income and household assets) (Exel and Rietveld 2009; Truong and Somenahalli 2015; Farber et al. 2016; Khoo and Asitha 2016; Mattioli and Anable 2017; Jain and Tiwari 2019; Cabral and Kim 2022).

Secondly, there are mainly possible signals coming from an automatic measurement device. The data piece can be a position (or its derivatives) (Khoo and Asitha 2016, Rostami-Nasab and Shafahi 2020), temperature (Nosal and Miranda-Moreno 2014), CO2 quantity (Chen and Lei 2017), a digital image from an embedded or exterior camera (Khoo and Asitha 2016, Rostami-Nasab and Shafahi 2020, Cabral and Kim 2022) or the cellular network (Montero et al. 2019; Fekih et al. 2021). If most of the data is based on a quantitative measurement scale, a qualitative scale is sometimes encountered as with using/not using the ticket machine or presence/absence of ticket reimbursement (Exel and Rietveld 2009). Costs (of a ticket, transport vehicle, accident, strike) or duration values (delays, trip, bridge construction,…) (Alonso et al. 2019; Hong et al. 2019) also belong to this second modality.

Thirdly, the diversity comes from the presence of different measurement scale models (Stevens 1974; Chrisman 1998). Reporting this heterogeneity is important because, if one wants to carry out an analysis which considers all the data collected, this scale diversity will condition the choice of the statistical methods. Let us now focus on the diversity of statistical methods.

Firstly, given the presence of different data sets, more particularly with time variables, the data mining starts with *data characterizing*. We mean here the methods for a "mere summarizing" of the data using the three kinds of usual models, i.e. *mathematical* through global indicators as average traffic flow and

average speed (Khoo and Asitha 2016) or the overall probability of traffic breakdown (Tu et al. 2012); *graphical* using an overall view as the traffic accident histogram (Ryder et al. 2019), percentiles of car driver distance for food shopping (Mattioli and Anable 2017) or traffic boxplot (Guardiola et al. 2014); *verbal* as "the density is low" (Khoo and Asitha 2016). Obviously, the term "mere summarizing" is not pejorative but seeks to distinguish a deeper analysis of an expert from a large set of either raw values of multidimensional signals (the components may be related to the driver and car behavior in the overtaking task) (Younsi et al. 2011) or indicators (within an origin–destination matrix of a traffic analysis) (Parry and Hazelton 2012; Fekih et al. 2021).

Secondly, comes methods in a multifactor and/or a multivariate context where the main aims are to find cause-to-effect relationships e.g. the impact of the traffic image information system on the driver travel choices (Khoo and Asitha 2016; Schmöcker et al. 2010) and/or relation between measurement variables, such as density and flow (Alonso et al. 2019). Given these main aims, the statistical analysis can be used with either a descriptive or inferential context and either a monovariate or multivariate approach, each approach with many options. For instance, with the inferential context, let us mention Bayesian inference (Parry and Hazelton 2013) and ANOVA for the monovariate approach (Crawford et al. 2017; Mattioli and Anable 2017; Jain and Tiwari 2019), MANOVA or logistic regression for the multivariate approach (Schmöcker et al. 2010; Cottrill and Thakuriah 2015; Truong and Somenahalli 2015; Cabral and Kim 2022); with the descriptive context, let us mention Correspondence Analysis (CA), Multiple Correspondence Analysis (MCA) (Cottrill and Thakuriah 2015; Truong and Somenahalli 2015; McCarthy et al. 2016; Ahmed et al. 2021; Yin and Leurent 2022), Cluster Correspondence Analysis (Mattioli and Anable 2017; Cabral and Kim 2022) and Principal Component Analysis (PCA) (Jain and Tiwari 2019; Krishnakumari et al. 2020), PCA being much more widely known/used than CA and MCA (Diana and Pronello 2010; Jalayer and Zhou 2016). Classification and finding clusters, rather than summarizing using a single representative of a set (often the arithmetic mean) are also part of second modality (McCarthy et al. 2016; Cheng et al. 2018; Egu and Bonnel 2020). All these methods come in the field that could be named *data analysis*.

A particularly interesting aspect in transportation comes with traffic data (mainly obtained using automatic measurement devices). If such data has often been addressed, the *data analysis* of both space and time aspects plus explanatory variables related to the users and the transport network environment (weather, pollution, interruption, socio-economic) is less frequent.

The structure of this paper is as follows. Section 2 is related to the analysis when these two kinds of data are present. Section 3 includes the description of the data and the statistical procedure (for reasons of confidentiality, the place of the study will not be provided and the units will often be put in AU, for Arbitrary Units). The literature suggests several main stages for the statistical procedure, e.g. 1) *Data transform* (from raw data to tabulated data), 2) *Visual mapping* (yielding a visual form), 3) *View transform* (yielding a view) and *Visual perception* (of the user) (Card et al. 1999). Here, the stages will be 1) *data characterizing*, 2) *data coding*, 3) *data table shaping*, 4) *table analysis* and 5) *result presentation*. The first two stages play a main role because

(a) due to the presence of factors linked to space, time and individuals, the raw data cannot be considered as is and
(b) both quantitative and qualitative variables may be present.

Section 4 provides the results and Section 5 discusses the results of the methods and suggests some future works.

2 Statistical Analysis Principle of Data Related to Both the Traffic and The Network Environment

Initial data is available in two main ways. In the first one, raw data is present; here are four main examples for a metropolitan network, the first one concerning traffic data and the following three the environment (maintenance, weather and socioeconomic, respectively):

Example 1: a series of checks in/out for a subway station set (time and space aspects are present, plus possible user data when a season ticket is used);
Example 2: a series of breakdowns, each one with data describing time aspects (chronology and duration), space aspects (place, vehicle label, vehicle subpart label) and dysfunction aspects (failure types and cost);
Example 3: a series of rainfall levels, temperatures wind speeds and pollution indicators for a given space-time window (a 1 km by 1 km geographic area of the network during a given hour of a specific day);
Example 4: a series of fuel litre prices for a given space-time window.

The time and space windows are often different for the four-time series. In the second way, more globalized data sets are present e.g. statistics for a given time window (a day) and a given space window (a subway line or a city geographic part). Whatever the way the data sets are available, a data characterizing stage is required.

Data characterizing methods contain a lot of techniques for summarizing traffic data and then showing these summaries (thus statistical indicators), see (Chen et al. 2015; Kolaczyk and Csardi 2014) for an overview. According to (Chen et al. 2015), the basic summarizing statistical operations for traffic data are spatial, temporal, directional and attribute-related aggregation. Kolaczyk and Csardi (2014) focus more on the concept of network graph, thus on the space aspect, including vertices (nodes), edges (links) and attributes. If attributes are related to frequencies, these two ways of looking at things allow building specific statistical views (line charts, parallel coordinate plots, circular plots, maps) (Gao et al. 2019) and decorating the graph (nodes and/or links) (Tu et al. 2012), respectively. Frequencies being shown using a specific graph, it is worth noting the high difficulty to add more than one supplementary variable such as rain, temperature, pollution, travel cost, accident, delay, individual characteristics (job, age…) or user rating of the network (Iseki et al. 2018; Gao et al. 2019; Klingen 2019).

Data analysis methods allow to go further in data operating. One of the most studied data sets concerns count data. A main category concerns the "trip" which

may contain either two extremities only, as with network origin destination (OD) flow rates, or more than two points (maybe with different kinds of transport devices). In this latter case, several names are used, as the activity sequence pattern (Ahmed et al. 2021), trip chaining (Yin and Leurent 2022) or multimodal pattern of individual mobility (Schmöcker et al. 2010). Another kind of frequency data concerns visitor counts at sites e.g. transit stations, tourist attraction points, stocks and transactions of bicycles at docking stations or journey-to-work (Chen 2018; Klingen 2019; Wan et al. 2021).

It is worth noting that there are few statistical methods capable of considering, in the same analysis, the two following generic data sets,

set A: network traffic data (count data related to both space and time windows),
set B: data linked to the network environment (weather, socioeconomic, breakdown…) with perhaps also data relating to users (sex, age, job, satisfaction level…),

Therefore in the presence of variables with different scale mathematical models (e.g. quantitative vs. qualitative) and with different window sizes (e.g. the hour vs. the week for a time window).

For a very first approach in the exploration of such large and different data sets, the main idea consists in building a series of plots from data matrices as follows (Jobson 2012):

1. data processing performed by the computer where a) each row of a matrix is associated with a point in a high dimension space (the space size being the number of columns, each column corresponding to an axis) and b) few new axes that are interesting combinations of the initial axes are computed (the dimension reduction coming from a least square principle). The same principle is used for the column points with the possibility to show relationships between row points and column points.
2. row and column point plot analysis performed by the human with several loops including either both stages 1 and 2 (more particularly if outliers are displayed) or stage 2 only (e.g. to increase the understanding using extra points or the quality of the graphics).

Several generic expressions are used for these methods (with controversy points of view for the expressions) such as *factor analysis* (Awad et al. 2023), *ordination (or inertia) methods* (Chahouki 2012), *geometric methods for feature extraction and dimension reduction* (Burges 2009) or *data reduction* (Jobson 2012). The latter expression being not precise enough to designate the methods (for instance using the arithmetic mean is also *data reduction*), we will use the expression of Burges (2009) with the acronym GM. Keeping in mind the presence of sets A and B, two specific GM are worth being mentioned, i.e. Principal Component Analysis (PCA) and Correspondence Analysis (CA) (Jobson 2012). The possible generic inputs for these methods and the reasons for considering here PCA and CA are as follows.

PCA and CA are based on the same principle: singular value decomposition obtained from an optimal solution to a least square criterion related to a two-entries table $\boldsymbol{Y}_{RxC}$ with R rows and C columns. When PCA and CA were designed, the main difference came from the type of $\boldsymbol{Y}_{RxC}$ content (Jobson 2012):

- for PCA, $\boldsymbol{Y}_{RxC}$ was related to V QuanTitative (QT) variables obtained for I individuals (humans, countries, network stations…) with the organisation as follows: $C=V$ and $R=I$ (Volle 1997; Pagès 2013; Bellanger and Tomassone 2014; Husson et al. 2016; Cornillon 2018). It is worth noting that I can be seen as the number of levels for a given QuaLitative (QL) variable such as with I space (or time) windows and that rows and columns don't play symmetrical roles. An example that can be cited concerns I regions (prefectures) of a country and J variables related to the transport network topology (number of nodes, node eccentricity…) (Tsiotas and Tselios 2023);
- for CA, $\boldsymbol{Y}_{RxC}$ was obtained from the crossing of $V=2$ QuaLitative (QL) variables with J and K levels respectively with the organisation as follows: $C=J$ and $R=K$, or the contrary, rows and columns playing symmetrical roles. $\boldsymbol{Y}_{RxC}$ is often named (two ways) contingency, frequency, correspondence or incidence table (Benzecri 1992; Volle 1997; Jobson 2012; Blasius and Greenacre 2006; Lebart et al. 2006; Taylor 2006; Nishisto 2007; Saporta 2011; Beh and Lombardo 2014; Truong and Somenahalli 2015; Friendly and Meyer 2016; Husson et al. 2016). An example stands with I customer profiles and J travel modes, data coming from an attitudinal travel survey (Diana and Pronello 2010).

Then, other contents have been considered,

- for PCA, a first generic case comes when $\boldsymbol{Y}_{RxC}$ is related to V ordinal QL scales, each with L levels, the number of levels being large enough (let say $L \geq 5$), e.g. a 5-levels Likert-scale for questions related to transport attitudes and the use of mobile applications for travel (van Lierop and Bahamonde-Birke 2023). A second generic case is with $V=1$ QT variable obtained for 2 factors with J and K levels, e.g. two time factors with I years and J months of the year when the variable corresponds to the average value of a climatic indicator in a town for a given month of a year, with $C=J$ and $R=I$ (Pruscha 2013). For other generic cases, let us focus on traffic data and more particularly on frequencies related to a network. A first example (labelled Ex.a) is when $R=O$ origins and $C=D$ destinations, the variable indicating the number of travellers (or cars, trains…) for a given pair (o, d) of a network with N points (Djukic et al. 2012; Antolini and Giusti 2021). A second example (Ex.b) comes when the time aspect is considered through J time windows and the space aspect is present through N^2 levels. The table $\boldsymbol{Y}_{RxC}$ can be built using $C=N^2$ and $R=J$ or the contrary, i.e. the rows are the links and the columns are the time windows (Chawla et al. 2012; Montero et al. 2019). In fact, as stated by Tsiotas and Tselios (2023), PCA enjoys a variety of applications in transportation research;
- for CA, a first generic case comes when $\boldsymbol{Y}_{RxC}$ is related to V homogeneous non-negative QT variables obtained for I individuals, i.e. when the distance used in CA (the Chi-square distance) (Jobson 2012) as an actual meaning for comparing either

the rows or the columns of $\boldsymbol{Y}_{RxC}$ (Beh and Lombardo 2014; Husson et al. 2016). A good example comes with V anthropometric dimensions and I humans where $C=V$ and $R=I$ (Benzecri 1992). For other generic cases, Ex.a and Ex.b can be mentioned (Kumar et al. 2014; Guex et al. 2023; Montero et al. 2019). In fact, as stated Diana and Pronello (2010), CA has been little used in travel behavior studies. Furthermore, in any cases, CA is less often used than PCA. A final but important generic case comes when $\boldsymbol{Y}_{RxC}$ is related to $V>2$ QL variables considered for I individuals with the organisation as follows: $R=I$ and $C=J$, the total number of modalities computed across the V variables ($J=\sum J_v$ where J_v is the number of modalities for each variable v, $v=1, \ldots, V$). The generic value of $\boldsymbol{Y}_{RxC}$ is within {0, 1}. The method, named Multiple Correspondence Analysis (MCA), can be seen either as a generalization of CA (with $V>2$) or a specific case of CA: for each individual row, the V sub-totals across the J_v modalities of variable v is 1 (historically, CA computer program was applied to such a table $\boldsymbol{Y}_{RxC}$). A good MCA example comes with variables related to socio-demographic attributes and an activity sequence pattern of travel behavior (Ahmed et al. 2021). MCA examples with QT variables cut using crisp or fuzzy windows ($\boldsymbol{Y}_{RxC}$ is within [0, 1]) have been also considered as an alternative of PCA mainly with the possibility of showing non-linear relationships and to reduce the information less when averaging across a factor, e.g. time factor (Gueguin et al. 2008; Younsi et al. 2011).

Finally, if we consider the presence of set A, there are two main reasons for comparing PCA and CA:

1. while CA was designed for frequency values but not PCA, the latter is sometimes used with such data and
2. these two methods being built for different kinds of input matrix, just for that there are very few comparisons using the same actual data matrix.

Then, given the presence of both sets A and B and that CA input is mainly QL scales, it would be interesting to introduce QT variables related to set B by transforming them into QL variables using a fuzzy windowing.

3 Methods

Traffic data corresponds to a subway network of a town with $S=10$ stations and during $M=4$ months. The corresponding data set is available in a row format (about 32 gigabytes), see *Example 1* above (idem for supplementary data – rain level, temperature and petrol price – that was found on the web, thus with different time scales). The presence of miscellaneous data can be summarized Figs. 1a and 2a. left side, where the blue colour designates the space aspect, the red colour is for time and the green colour stands for the influencing factors and other possible measurement variables, e.g. user rating of the network. The data analysis procedure, Figs. 1 and 2, is as follows.

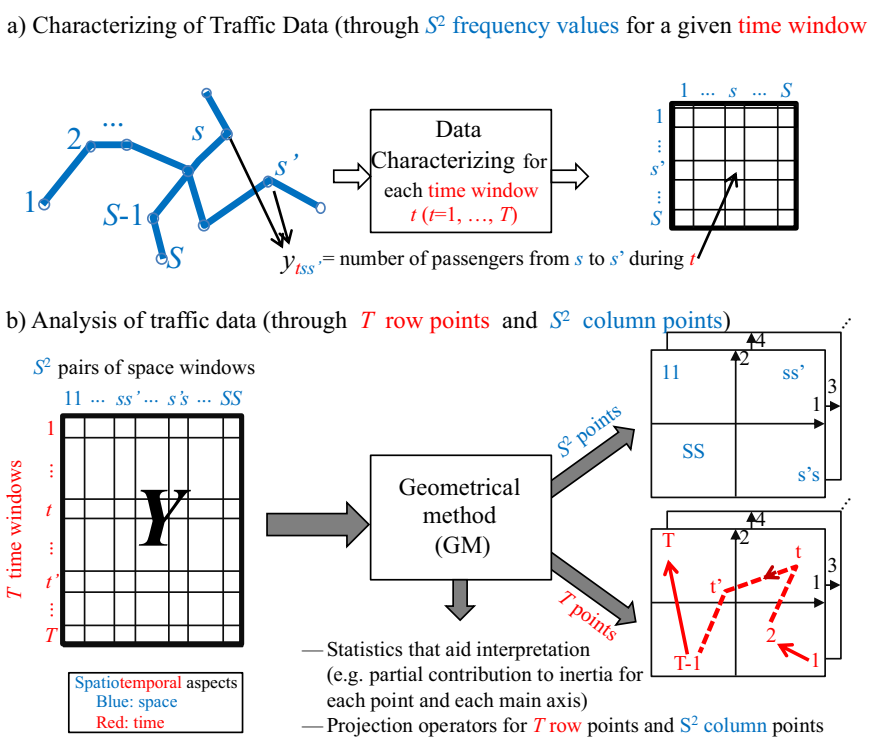

Fig. 1 Data analysis procedure of traffic data (table ***Y***) using Principal Component Analysis (PCA) and Correspondence Analysis (CA) (the procedure is identical, the statistical similarities/differences being shown in Table 1)

3.1 Data Characterizing

As a first step, data characterizing is the most important one because it conditions all the rest of the analysis. Several approaches can be considered, especially if we keep in mind that the statistical indicators can be related to the user (metro subscription cards are often used), the subway station (to have an idea of the rate of presence in a given space s, $s=1,..,S=10$) or the journey in the metropolitan network (thus with places of departure s and arrival s'). This third approach will be used here. Let (s, s') be the generic space pair among the 100 possible pairs (Fig. 1a, $s=s'$ means that an individual remains at a given place e.g. uses a parking space via his/her metro subscription card). If studying the overall statistics related to the 4 months can be interesting e.g. though a 10 by 10 origin-destination matrix, it is preferable at first, to summarize the data less. There remains the choice of the time window which is a compromise between a low value (e.g. 10 min), which can yield both low effectives and complex results, and a high value (e.g. the day), that may not be accurate enough. As a compromise, 1 h is considered. Let $y_{t,ss'}$ be the number of network users moving from s to s' during t ($t=1,..., T$; t is linked to a day d, a week w and a month m but to simplify the notation only t is used, except if required). From this

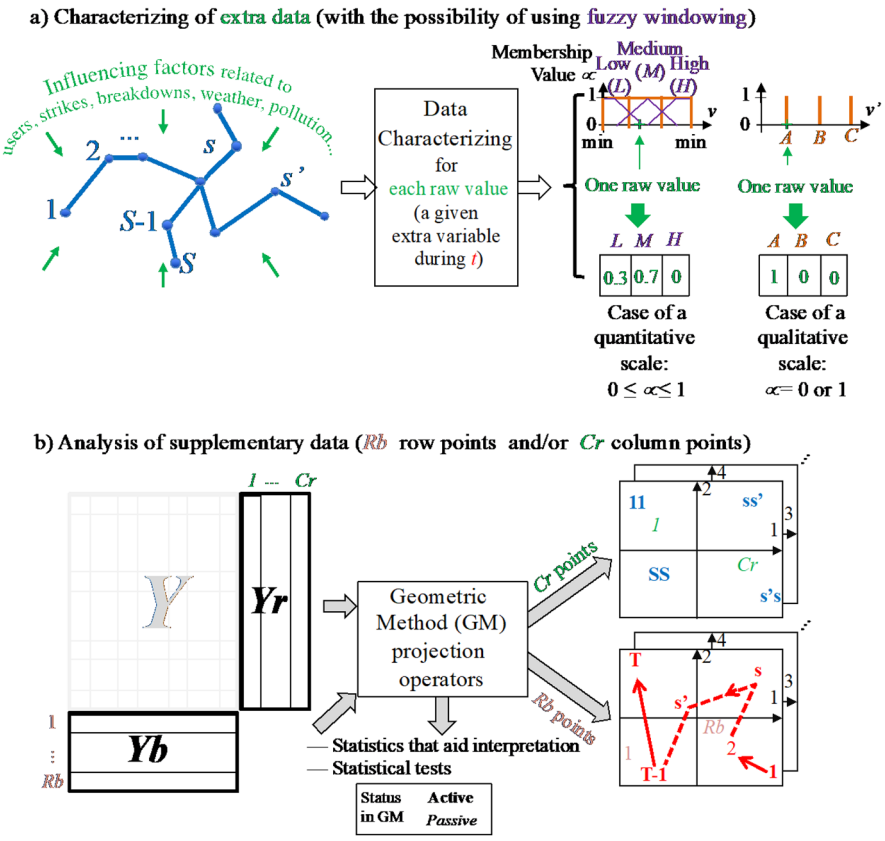

Fig. 2 Data analysis procedure of traffic data (table ***Y***) and extra variables (e.g. with transport network environment data, table ***Y*r**) using PCA and CA

generic duration, many other statistical indicators can be computed with different summarizing levels according to the time and/or space factors:

- for each pair (s, s'), one can compute cumulated frequencies related to the day within the week, the working day vs. the weekend day (or public holiday), the week…;
- for each time window t, on can compute cumulated frequencies related to a subset of stations (north vs. south, city centre vs. suburbs, airport…);
- time and space factors t and s can be combined with different summarizing levels.

Here we suppose that the transport network environmental data is available for the generic time window t. Let $y_{t,e}$ be the generic value for the environment variable e ($e = 1,\ldots, E = 3$, for the rain level, temperature and gas price in our case).

3.2 Data Coding

This stage is required in the perspective of a multivariate analysis and if the measurement scales are different. Here, with the traffic data (set A), there is no specific coding technique (except the coding model intrinsic to PCA and CA).

With extra data (set B), the problem is more complex since there may be QL scales (e.g. sex or job of the transport network user) and quantitative scales (e.g. rating of the user about the network and any weather or pollution indicator). Both PCA and CA allow to consider new variables (either QL or QT) as supplementary points (Husson et al. 2016). An alternative with CA is to get homogeneous data, the QT scales being cut into intervals. Several cutting criteria can be considered as with identical width vs. identical frequencies. Another criterion can be crisp vs. fuzzy windowing.

A first main problem with crisp windowing is that when cutting a range, an interval int_i can be defined using either $[a_i, b_i[$ or $[a_i, b_i]$, the choice being arbitrary. With fuzzy windowing, the idea is that the membership value (μ) when $y = b_i$ is neither $\mu_i(y) = 0$ nor $\mu_i(y) = 1$ respectively but, $\mu_i(y) = 0.5$.

A second main problem with crisp windowing is that with 3 values so that $y" < y < y'$,

- if y and y' are very close but on either side of the border b_i ($y < b_i$ and $b_i < y'$), the membership values related to the interval i are fully different: $\mu_i(y) = 1$ and $\mu_i(y') = 0$ respectively but,
- if y and $y"$ are very far and within the interval i with $y\text{-}y" > > y'\text{-}y$, the memberships values related to the interval i are identical: $\mu_i(y) = 1$ and $\mu_i(y") = 1$ respectively.

To reduce such an information loss, the fuzzy cutting alternative is preferable, even if it is more complex: many choices are possible (Arslan 2009) and the computing time is higher. For a very first analysis, here is the principle of fuzzy windowing:

1. three intervals (or crisp windows) are considered with an identical width w;
2. the 'low' fuzzy window is considered as follows: from the minimum value (left side of the range) to $w/2$, the membership remains at 1, then it decreases linearly so that the middle of the range yields a null membership value. The symmetric reasoning is performed with the 'high' window;
3. to remain in a statistical context, the sum of the membership values is 1, yielding a triangular pattern for the 'medium' window.

Figure 2a portrays the fuzzy segmentation principle in purple. A case using a QL scale with 3 levels is also presented. Therefore, with a QT scale, the input universe of discourse is quantified into subjectively equal subintervals, yielding two trapezoidal and one triangular windows, which are in fact 3 fuzzy sets (FS) (Arslan 2009). It is worth noting that this coding procedure could be seen as a basic input procedure with Correspondence Analysis (AC), just as the usual standardization (with the mean and standard deviation of a variable) is the basic procedure with Principal Component Analysis (PCA). As stated above, with

several QT variables that play an active role in the control of the main axes, MCA is used instead of CA.

3.3 Data Table Shaping

Generally speaking, the table can have two or more entries. For the traffic analysis, the table can mainly have 3 entries (for t, s and s') or 2 entries e.g. the row can correspond to t and the column to each possible pair (s, s'). In the perspective of analysing both traffic and extra data, a 2-entry alternative will be chosen, Fig. 1b. Let $\boldsymbol{Y}$ be this table, with $C = S^2 = 100$ columns and $R = 2033$ rows (hours with low effectives during the 4 months being removed), the generic value being $y_{t,ss'}$.

Concerning the transport network extra data, one can consider a table with R rows and E columns. Let $\boldsymbol{Yr}$ be this table where the letter r is used to indicate that this table is placed on the right of $\boldsymbol{Y}$, Fig. 2b. In our case, $E = 3$ for temperature, rainfall and gas price. As stated in the data characterizing stage, several time summaries can be computed from $\boldsymbol{Y}$ (and $\boldsymbol{Yr}$) e.g. a table where each row corresponds to a day. Let $\boldsymbol{Yb}$ be the table containing these summaries where the letter b is used to indicate that this table is placed below $\boldsymbol{Y}$, Fig. 2b.

3.4 Table Analysis

In a multivariate analysis perspective, the bibliography analysis (see §1 and 2) shows that $\boldsymbol{Y}$ can be studied at first using a descriptive approach, thus mainly with a geometric method for feature extraction and dimension reduction (GM). With a frequency table $\boldsymbol{Y}$, PCA and CA can be used as follows: once the main axes from $\boldsymbol{Y}$ have been carefully analysed, supplementary row or column points (from $\boldsymbol{Yb}$ and $\boldsymbol{Yr}$ respectively) can be projected onto a plane obtained by crossing two main axes (these points have a passive status in the building of the main axes).

The overall data analysis procedure is summarized though Figs. 1 and 2. The mathematical principle of these methods is recalled in Table 1, see (Husson et al. 2016) for further details. The informatics tool used to perform these analysis stages is the software environment R; PCA, CA and some graphical results being built thanks to specific packages as FactoMineR (Husson et al. 2016) or FactoExtra (Kassambara 2017), which requires ggplot2 (Wickham 2009).

3.5 Result Presentation

One of the main mathematical aspects of CA and PCA comes from the least square distance optimization yielding eigenvalue and eigenvector computation (Table 1). The analysis of PCA and CA outputs yields both graphical results (mainly projection of row and column points on a plane plus maybe scree plot of eigenvalues or bargraph of contributions) (Husson et al. 2016; Kassambara 2017) and verbal results (a text from the analysis of such outputs). Both results being complex and taking

much place, simpler and more synthetic graphical results can be considered, such as a plot showing the effects of a factor on a variable (remember that time factor recovers different scales e.g. hour, day or weekend vs. not weekend) or relationships between two variables.

4 Results

Let us start with the analysis of traffic data, the statistical unit being a one-hour time window, see Fig. 1b.

4.1 Analysis of Network Traffic Data

Whether with PCA or CA method remember that the input table is $\boldsymbol{Y}_{2033*100}$ with $T=2033$ rows (each row corresponding to a 1 h time window) and $O*D=100$ columns (each column is linked to a pair (o, d), see Table 1 and Fig. 1b).

PCA yields a series of principal axes with 53, 18, 6 and 2% of the total inertia. The first 2 axes showing much larger dispersion than the next axes, the analysis will focus on the first 2 principal components (labelled Axes 1 and 2) which display, collectively, 71% of the total inertia. Axis 1 shows that all space points are on the same side, Fig. 3. Such a result is frequent with PCA and exhibits a *global size effect*: Axis 1 will essentially oppose time windows with a low frequency, on its left side, to time windows with a high frequency, on its right side. The space points with the strongest correlations (about 0.94) with Axis 1 are CD, DA, AD, DB and BD (it is worth noting that DA/AD, DB/BD CD/DC correspond to one direction/opposite direction, DC having a high correlation also). The number of time points being rather large, and the time having several aspects —hour, day of the week, week and month — displaying several subplots is better than only one. An interesting possibility consists in using one subplot for each day of the week, each subplot with as many hour trajectories as there are identical days during the available observation time window i.e. 4 months. Figure 4 clearly shows three main factors: a) the day (within the week, thus with 7 levels), b) the week (within the 4 months, thus with 17 or 18 levels depending on the day) and c) the hour (within the day with 24 levels):

(a) for the day factor, globally, the weekend trajectories are quite different from the other 5 trajectories, Saturday and Sunday trajectories being completely located on the very left side of Axis 1. This indicates low frequencies, compared to the average, on Saturday and Sunday. For each of the 5 working days, the time trajectories are quite identical (the trajectory looks like a "<") but
(b) for the first week of January, shown in red, there is a large difference compared with the other weeks. More particularly, on Thursday, January 1st and Friday, January 2nd, the frequencies are very low, which is consistent with the new year day celebration;

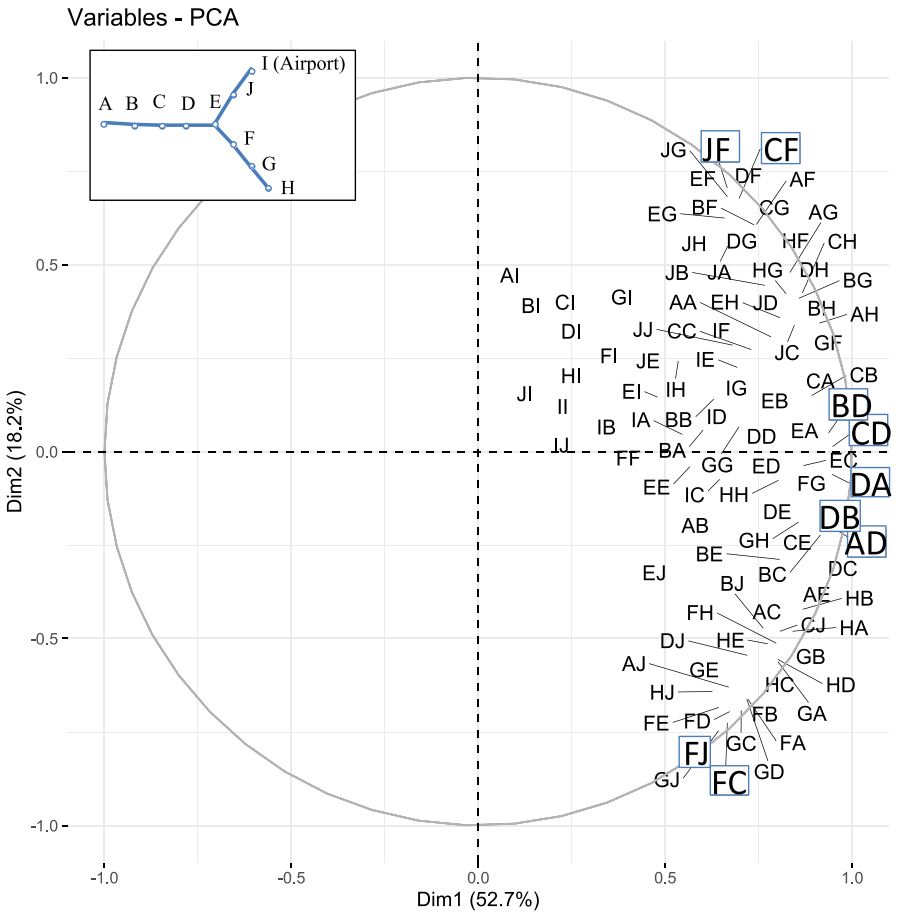

Fig. 3 PCA output of table $\boldsymbol{Y}$ with $T = 2033$ rows (each row corresponding to a 1-h time window) and $O{*}D = 100$ columns (each column is linked to a pair (o, d), see Table 1). Projection of the 100 column points of the main plane

(c) for the hour factor, globally, network user numbers during the time windows [6 h, 8 h] and [16 h, 18 h] are much higher than within the rest of the day, with peaks at 7 h and 17 h, this only for the 5 working days (no difference can be found for Saturday and Sunday).

With the effects of these 3 time factors, the presence of 100 variables, and the multiple statistical *variable(s) vs. factor(s)* plots, many figures can be considered to show how PCA outputs are data consistent. Let us focus on the hour factor and variables CD and DC who played a main role in Axis 1 control. Each hour being present several times during the 4 months (between 7*17 = 119 and 7*18 = 126 times), the corresponding data set can be shown through a boxplot (rather than

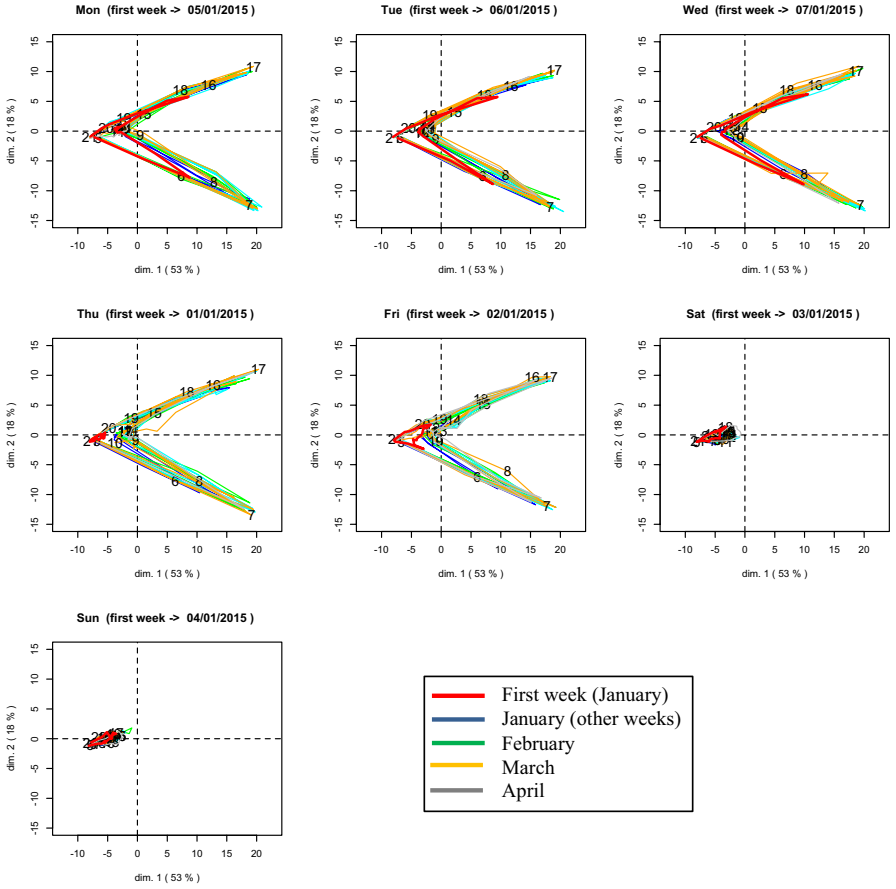

Fig. 4 PCA output of table $\boldsymbol{Y}$ with $T = 2033$ rows (each row corresponding to a 1-h time window) and $O{*}D = 100$ columns (each column is linked to a pair (o, d), see Table 1). Projection of the 2033 row points of the main plane with one weekday per subplot (for each subplot, the points are linked in the chronological order; there are 18 weeks, thus 18-time trajectories)

the arithmetic mean, possibly with more or less the standard deviation). The result figure, Fig. 7a and b, shows that PCA output is data consistent: for both trips CD and DC, the highest frequencies correspond to 7 h and 17 h (these two-time levels are displayed using a green, vertical and dotted line).

Concerning Axis 2, the space points with the strongest positive correlations (about 0.68) are JF, EF, DF, JG and CF); the points with the strongest negative correlation (about -0.72) corresponding to the opposite directions (FJ, FE, …). Figure 4 shows that these positive and negative positions correspond to time windows [16 h, 18 h] and [6 h, 8 h] respectively. Figure 7e and f show that the opposition underscored by PCA Axis 2 is data consistent. It is worth noting that the maximum frequency is much higher for JF (or FJ) than CD (or DC).

Let us go to the CA method (CA output will be described in the same way as with PCA, CA vs. PCA comparison will be then made).

Beforehand, the chi-square is 2.2 10^{-6} which is associated to a p-value being much lower than the usual reference i.e. 0.05. Although the limits of use of the chi-squared test are far from being verified (many cells have small frequencies), the very low p-value yields that table $\boldsymbol{Y}$ is far to the independency case (Husson et al. 2016). CA finds a series of principal axes with 47, 15, 6 and 3% of the total inertia. The first 2 axes showing values much larger than the following, the analysis will focus on Axes 1 and 2 (about 62% of the total inertia). Axis 1 is mainly controlled by space points FJ and FC, see Fig. 5 left side, and points JF and CF, on the opposite side. For these 4 points, as with the next points in order of decreasing contributions, Axis 1 shows an opposition between one direction/opposite direction, e.g. FJ/JF and FC/CF (thus between clusters C1 and C2). The time windows during which these oppositions are present can be found from the projection of the row points with a large contribution to Axis 1 control. Figure 6 shows factor effects as follows:

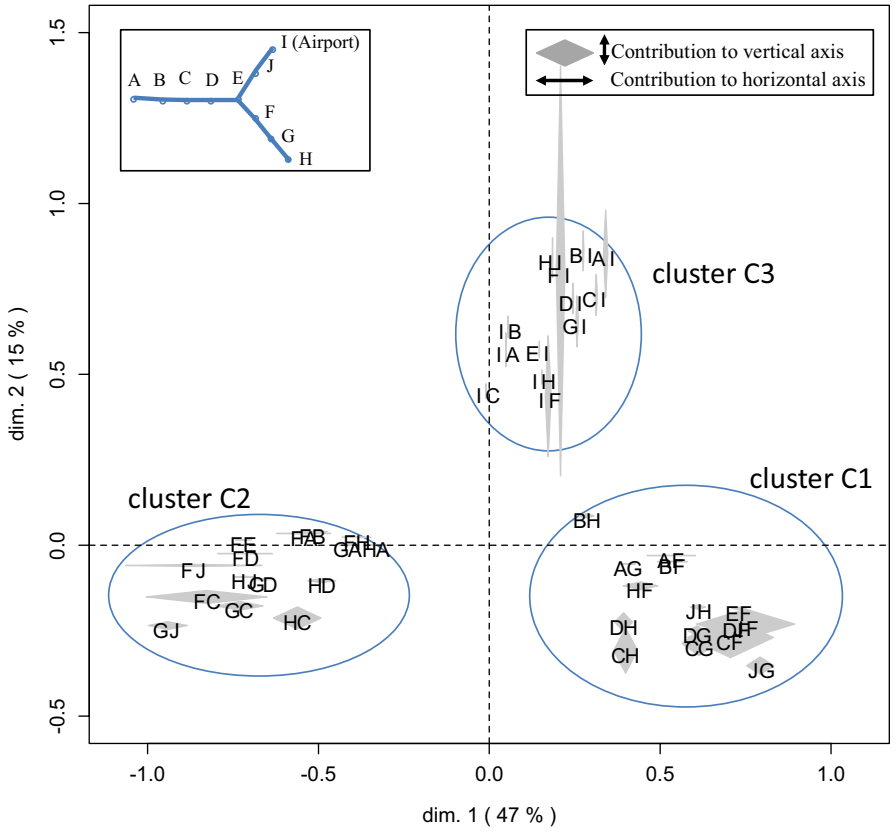

Fig. 5 CA output of table $\boldsymbol{Y}$ with $T=2033$ rows (each row corresponding to a 1-h time window) and $O*D=100$ columns (each column is linked to a pair (o, d), see Table 1). Projection of the 100 column points of the main plane (only points with a strong contribution in the control of Axes 1 or 2 are shown)

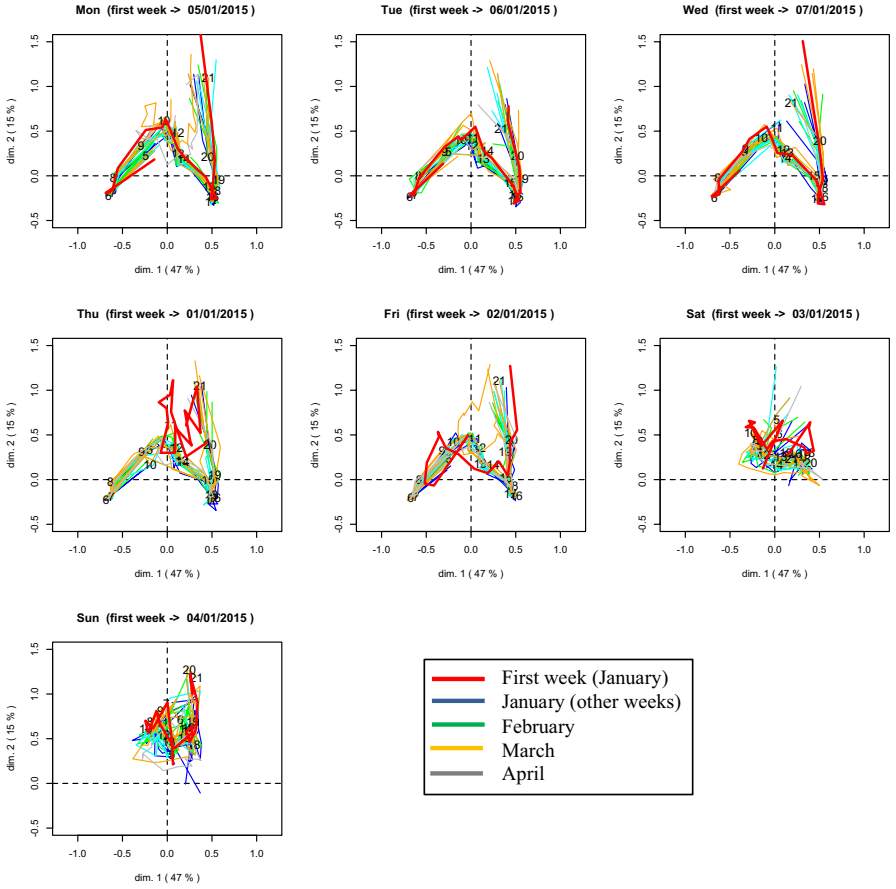

Fig. 6 CA output of table $\boldsymbol{Y}$ with $T = 2033$ rows (each row corresponding to a 1-h time window) and $O{*}D = 100$ columns (each column is linked to a pair (o, d), see Table 1). Projection of the 2033 row points of the main plane with one weekday per subplot (for each subplot, the points are linked in the chronological order; there are 18 weeks, thus 18-time trajectories)

(a) for the day factor, the weekend trajectories are quite different from the other 5 trajectories;
(b) for the week factor, the level of repeatability is strong from week to week, except for the first 2 days of January (see red trajectories);
(c) for the hour trajectory, three main time window subsets emerge during the 5 working days: [6 h-8 h] on the Axis 1 very left side, [17 h, 21 h] on the opposite side, and the remaining hours near the 0 position (i.e. the gravity centre) of Axis 1.

Figure 7e and f show that the correspondences underscored by CA are data consistent e.g. there is a high correspondence between the space window FJ and the time window [6 h-8 h], see Axis 1 left side, and a high correspondence between JF and [16 h-18 h], see Axis 1 right side.

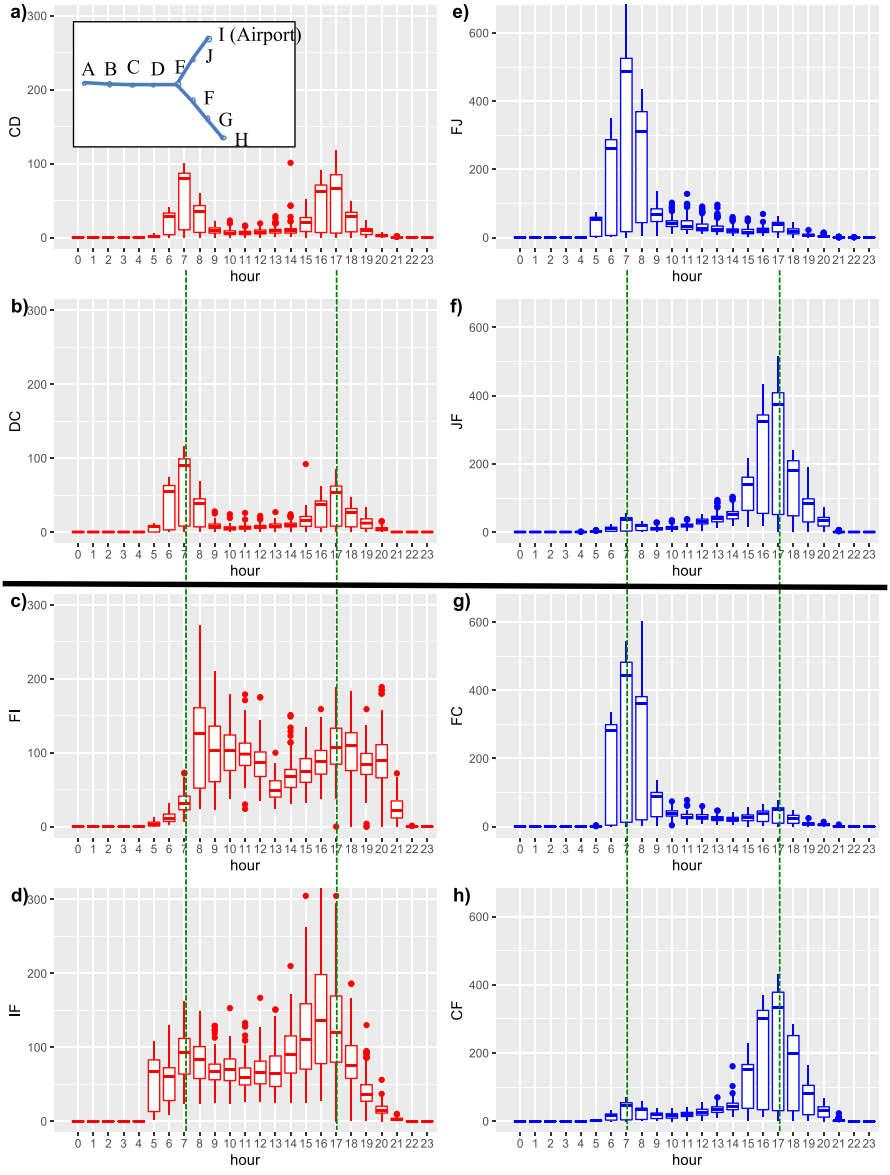

Fig. 7 Height time plots (4 links with one direction and opposite direction, e.g. CD and DC) to show how PCA and CA outputs are data consistent (if a link is more frequent in the morning, this one is displayed above, e.g. FJ is above JF; left and right subplots have different vertical scales): **a** and **b** two column variables (CD and DC) that played a main role in PCA Axis 1 control (Figs. 3 and 4); **c** and **d** idem for CA Axis 2 (top of Figs. 5 and 6); **e** and **f** idem for PCA Axis 2 (Figs. 3 and 4) and CA Axis 1 (Figs. 5 and 6); **g** and **h**, as with **e** and **f**, also show that two close points have identical profiles (see FJ/FC in the cluster C1 and JF/CF in cluster C2 in Fig. 5 and in low and top position in Fig. 3)

Axis 2 is mainly controlled by points FI, IF and AI; then come points all related to station I, see Fig. 4 top. Thus Axis 2 displays travel behaviours related to the airport (cluster C3), these behaviours being independent to the results highlighted above (from Axis 1) e.g. go to or from the airport is not necessarily done during time windows [6 h-8 h] or [16 h-18 h]. Figure 7c and d show that this result is data consistent.

As shown in Fig. 2b, PCA and CA can be continued by studying the projection of supplementary points in relation to the initial column and row points (Figs. 3, 4, 5, and 6).

4.2 Analysis of Time and Environmental Variables as Extra Data

With PCA, the effects of the time factor (hour within the day, day within the week and week), can be obtained using graph customization or hypothesis tests, see Table 1i and j respectively. A graph customization possibility is shown Fig. 8a where the points linked to a given hour (between 5 and 21 h) yield a cloud with a specific colour (see the legend); the 17 gravity centres are also displayed and connected in the chronological order (in blue). This yields a time trajectory that is very close to the trajectories of the first five days of the week, see Fig. 4. Figure 8a displays the high effect of the hour factor for both Axes 1 and 2 (day and week factors having less influence). The same customization procedure can be performed with CA, Fig. 9a, yielding results that are consistent with Fig. 6.

Given the large number of points (especially 17+7+18 points for hour, day and week modalities) to test positions when the inference statistical approach is used, results will be shown through tables instead of text. Such a presentation mode will facilitate PCA vs. CA comparison.

With PCA, only one extra variable is far from the correlation circle centre (Fig. 8b) i.e. Temperature, which is well correlated to Axis 2. This result is data consistent (see the time trajectory shown Fig. 8a). With CA, there are three points (one for each fuzzy window) for each environment variable which are connected in the increasing order (Fig. 9b). The three trajectories show that Temperature presents 3 points with different positions along Axis 1. These positions are data consistent. The 3 points of the Rainfall show a high correlation but with Axis 2, yielding more rain late in the day. Crossing Axes 1 and 2 yields that Temperature and Rainfall are quite independent.

4.3 PCA vs. CA Comparison

Whether for column or row points, PCA and CA clouds are very different. Concerning the column points, PCA shows no cluster for the 100-space window pairs, Fig. 3, whereas CA yields 3 clusters C1, C2 and C3, Fig. 5. Cluster C3 is mainly related to the airport (either as origin or destination station), and clusters C1 and C2 contain pairs with opposite directions (without the airport station); more particularly the departure stations of cluster C1 are mainly towards the south-east

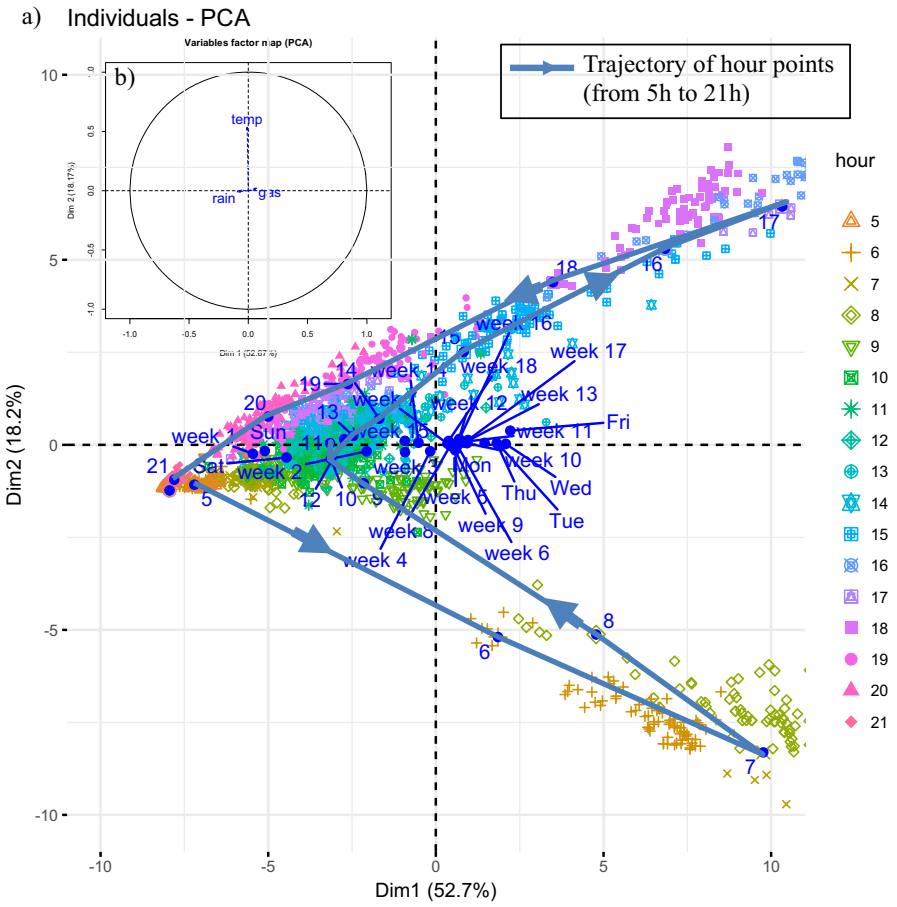

Fig. 8 Time and environmental factor effects with PCA **a** the 2033-h points are shown through different colours; the hours (from 5 to 21 h), the days (from Monday to Sunday) and the weeks (from 1 to 18) are shown through the modality points of an illustrative qualitative variable (some points with quite identical positions are slightly displaced); **b** the temperature, rain level and gas price are shown through quantitative supplementary variables in relation to the correlation circle

(thus these stations correspond to the destination for cluster C2). Combining Figs. 5 and 6 yields that there are two main ways to distinguish the displacement behaviours:

1. trips related to the job (or school and university): users start early in the morning and mainly from the south-east, then go back home in the late afternoon-early evening. One must acknowledge that this result is also present with PCA, but with Axis 2 only, see for instance

   – trips FJ and FC and hours 6, 7 and 8 on its negative side, Figs. 3 and 4;

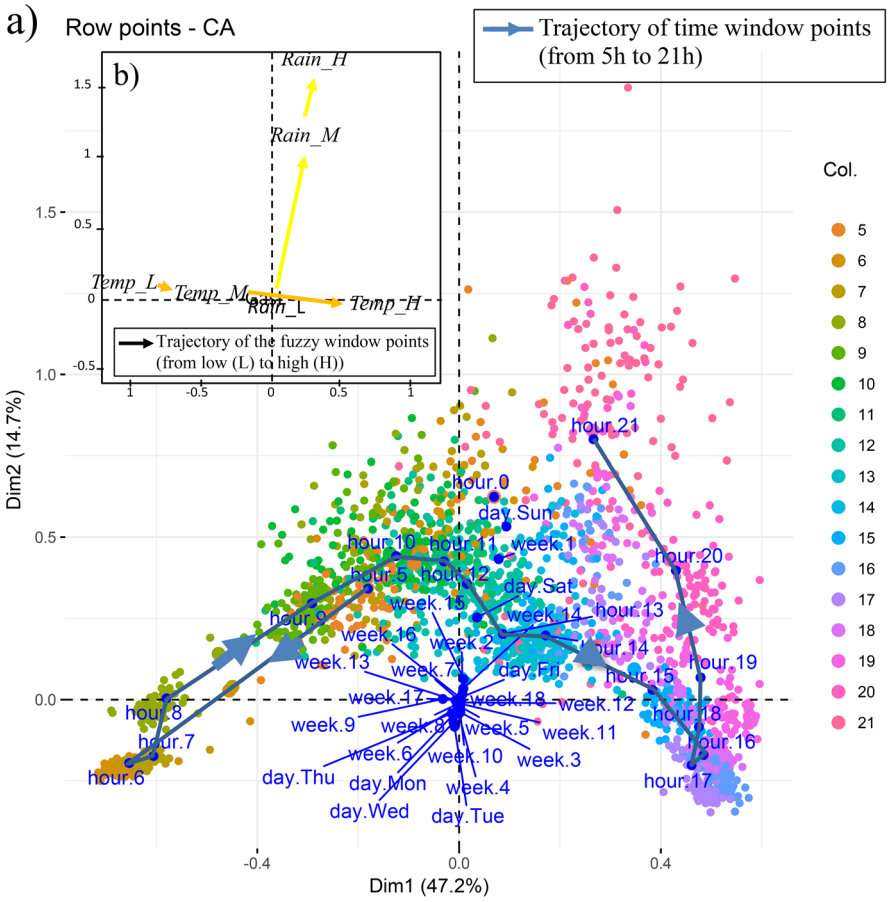

Fig. 9 Time and environmental factor effects with CA **a** the 2033-h points are shown through different colours; the hours (from 5 to 21 h), the days (from Monday to Sunday) and the weeks (from 1 to 18) are shown through the modality points of an illustrative qualitative variable (some points with quite identical positions are slightly displaced); **b** each of the temperature, rain level and gas price variable is shown through 3 supplementary modalities (each modality corresponds to a fuzzy window, see Fig. 2a)

– trips JF and CF and hours 16, 17 and 18 on its positive side (see also Fig. 7 right side).

Given the main effect of the hour factor (Figs. 4 and 6), the temperature is significantly correlated with PCA Axis 2 (Fig. 10d) and CA Axis 1 (Fig. 10i). The rainfall effect is only present with CA only and with more rain in the evening. One could reproach these two methods for highlighting trivial phenomena, however these methods would have been very bad if the relationships between time of day and temperature had not been shown. Consistently, as suggested in Figs. 8b and 9b, both methods show that time of the day is not connected to the gas price (this variable appears nowhere in Fig. 10).

PCA

Column (space) Points (with an active status in the main axis control)

a)	Axis 1	Axis 2
+	CD, DA, BD, DB, AD, CE...	JF, EF, DF, JG, CF, EG...
-		FJ, GJ, FC, FD, GC, FE...

Illustrative Qualitative variables

b)	Axis 1	Axis 2
	Hour Day Week	Hour

Illustrative modalities (from above variables)

c)	Axis 1	Axis 2
+	6...8, 16...18h Mon... Friday	15...20h
-	5, 9...14, 20, 21h Sat, Sun Week 1, 2	5...21h

Illustrative quantitative variables (environment)

d)	Axis 1	Axis 2
+		Temperature
-		

CA

f)	Axis 1	Axis 2
+	JG, JF, EF, DF, CF, EG...	BI, AI, HI, FI, CI, DI, GI...
-	GJ, FJ, FC, GC, FE, FD...	JD, JG, CH, CG, DJ, CD...

g)	Axis 1	Axis 2
	Hour	Hour Day Week

h)	Axis 1	Axis 2
+	14...20h Sunday	5, 9...15, 20...21h Sun, Sat Week 1
-	6...10h	6, 7, 16...19h Mon...Thur

i)	Axis 1	Axis 2
+	Temp_ Low	Rain_High, Rain_Medium
-	Temp_High	

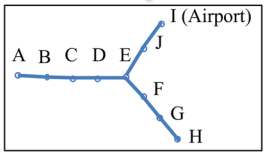

Fig. 10 Points with either a strong contribution or leading to rejection of the null hypothesis (p-value lower than 0.05, see Table 1j) for Axes 1 and 2 of PCA and CA

2. trips to and from the airport. Most trips are present, firstly during the hours 9 to 15 or 20 to 21, see the two top peaks according to Axis 2, Fig. 6, and, secondly, during the weekend (Fig. 10h). This second behaviour is not underscored by PCA, see Fig. 4.

To summarize, CA shows more displacement behaviours, which could be named 1) *daily work-related behaviour* (from Axis 1) and 2) *holidays-related behaviour* (from Axis 2, knowing that we can fly to work). If we pay attention to the differences in scales between the left and right parts of Fig. 7, the following remark should be made concerning CA. Count data in the right column are about twice as large as those in the left column; this is consistent with the number of employees or students going to work, see Fig. 7e to h, which is higher than the number of people who take the plane, Fig. 7c or d. This higher value for the frequency data is not shown in CA because this method deals with profiles (thus no absolute frequencies), see Table 1e. For instance, Fig. 5 shows that points FJ and FC are closed, this is because the profiles are quite identical, see Fig. 7e and g (but the frequencies also). If we divide the frequencies of the column FJ by 2 and restart CA, points FJ and FC remain closed.

5 Discussion

Considering the overall procedure shown in Figs. 1 and 2, let us discuss the PCA and CA outputs.

5.1 Traffic Data (Set A)

With our actual example, it is undeniable that the graphic outputs related to the frequency data are more complicated with CA than PCA, see Fig. 3 vs. Fig. 5 and Fig. 4 vs. Fig. 6. One reason is linked to the *global size effect* shown by PCA: with this method, the table columns are viewed as QT variables (the frequencies related to a link in our case) and when values related to a variable *v* increase (e.g. for a trip (*s*, *s'*)), the growing effect is more or less present with another variable *v*' (e.g. a trip (*s"*, *s'"*)). This effect yields that all the variables are situated on the same side of Axis 1. Obviously, these positive positions along Axis 1 do not involve that all the linear correlation coefficients (*cor*, see Table 1, part f) are positive, e.g. *cor*(DA, AD) = 0.92, *cor*(DA, FJ) = 0.63, *cor*(DA, JF) = 0.60, but *cor*(JF, FJ) = -0.12 (checking the scatterplot for these 4 bivariate series yields that *cor* value is data consistent, knowing that this is not always the case, see (Anscombe 1973) for some examples). The positions of points DA, AD, FJ and JF, Fig. 3, are consistent with these correlation values and partly explain the "<" pattern of the time trajectory for the 5 working days of the week, see Fig. 4: the left and right sides of "<" yield low and high frequencies respectively. As stated in Section 2 and Table 1, row points and columns points don't play symmetrical roles and therefore the results would have been different with the $\boldsymbol{Y}$ transpose as PCA input.

This *global size effect* is not present with CA since each point (either row or column) is considered through a frequency profile and the distance used to compare two points (either 2 rows or 2 columns) is the Chi-square, thus a distance aimed to compare frequency profiles, see Table 1. It is for this reason that CA was able to highlight the traveller behavior linked to the use of the airport line (including station I), see Cluster C3 in Fig. 5, this use being not the most frequent, see Fig. 7.

The advantage of CA over PCA is consistent with results found from a biology study performed on count data (Hsu and Culhane 2023).

5.2 Weather and Gaze Price as Extra Data (Sets A and B)

First of all, it is worth noting that we could have given an active role to the variables of the two sets. Since the 3 variables of set B are QT and PCA can be conventionally regarded as a suitable choice for QT variables, PCA could have been used with a table ***Y*** with 103 columns, see Fig. 1b. In the same way, since a possible extension of CA is to incorporate variables with different scale models, MCA option could have been used with a table ***Y*** with 109 columns: 100 links and 3*3 = 9 windows, with 3 space windows per QT variable. Here our main objective being the space-time analysis of traffic data (knowing that this objective is quite complex, see Figs. 3, 4, 5, and 6), we preferred to incorporate the 3 variables with a passive status (as supplementary columns) in the building of the main axes.

If it was not possible to find highly interesting results from a supplementary table ***Yr*** (Fig. 2) related our 3-network environment variables (temperature, rain level and gas price, see Figs. 8b, 9b and 10i), an observation window larger than 4 months should achieve this. It would also be interesting to have other indicators, see Fig. 2a left side, because PCA and CA showed the possibility to display and test the connections between the factors and the main axes, this with either the connection presence (as with the temperature or rainfall) or the absence (as the gas price), see Figs. 8, 9 and 10i.

To illustrate this, we can consider simulated data, for example representing subjective scales. Let us suppose that each user gives some opinions during his/her trip about the comfort, cleanliness and the punctuality of the train. If these 3 aspects are assessed using a rating between 0 and 10 or by putting a cross on a segment showing two opposed nuances (as with 'low comfort' and 'high comfort'), it could be interesting to show complex relational phenomena. To achieve this, let us consider, as possible data, a first variable where the individual rating is proportional to the hour within the day (let X be this variable), then a variable with a linear relation with X (let $2X$ be this variable where $2X = 2*X$ plus a random noise with a Laplace-Gauss model where the arithmetic mean (am) is 0 and the standard deviation (sd) is 1) and finally a variable $X2$ with a quadratic relational model $(X2 = (X\text{-}am(X))^2$ plus a random noise with a Laplace-Gauss model with am = 0 and sd = 1). These 3 variables can be cut using fuzzy windows (Fig. 2a), yielding a total of 9 columns, and the corresponding membership values is analysed using MCA (instead of CA). To better understand how MCA works with such relational phenomena, let us first consider a table ***Y*** (Fig. 1b) with these 9 columns and 2033 rows. Figure 11a show MCA graphical output for Axes 1 and 2 which collectively display about 85% of the initial inertia. The trajectories corresponding to X and $2X$ (both in red) are very close which is data consistent. These two trajectories are very far from that of $X2$ (in purple; the point X2L is closer to the gravity center than points X2M and X2H because its weight is higher, see Table 1h)). Axis 1 exhibits a main correspondence between XM, $2XM$ and $X2L$ (see the green ellipsis on the negative side) which is data consistent (see Fig. 11, top-left corner). On the posited side, a second correspondence

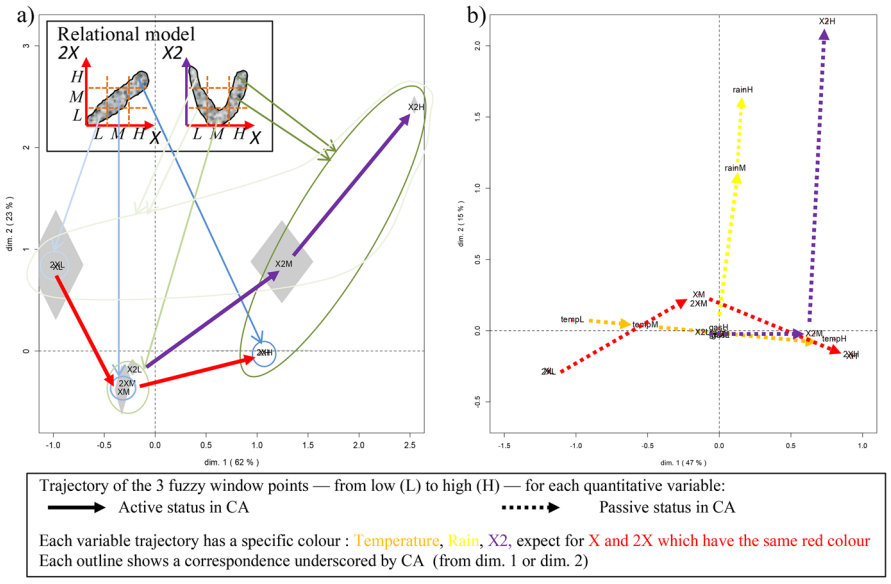

Fig. 11 How CA shows linear and quadratic relational phenomena using three simulated variables —*X*, *2X* and *X2*— that could represent subjective assessments. **a** Case of Fig. 1b with ***Y*** having 9 columns (one for each fuzzy window) and 2033 rows; the sizes of the lozenges indicate the contributions, see Fig. 5; the outlines show correspondences between the 9 fuzzy window points); the top-left rectangle displays the relational models and correspondences between two sets of three windows for each bivariate plot. **b** Case of Fig. 2b with actual and simulated data (***Yr*** has 18 columns and ***Y*** has 100 columns)

is present with XH, 2XH, X2M and X2H, here again data consistent. The same approach can be used with Axis 2.

It is worth noting that PCA Axis 1 (not shown) is controlled by *X* and *2X*, while Axis 2 by X2, which could suggest that the latter is independent from the first two variables (here again, the graphical output of PCA is simpler than with MCA; nevertheless it accounts less well for real phenomena). Let's go back to the initial approach i.e.

1. to perform the analysis of table ***Y*** corresponding to traffic data (table ***Y*** has 100 columns, see Figs. 5, 6, and 9 for the main results);
2. to consider a table ***Yr*** where the columns correspond to the fuzzy windows of all the extra variables (***Yr*** has 6*3 = 18 columns). Figure 11b shows the 18 supplementary points. Obviously, the relative positions of the 9 fuzzy windows corresponding to the simulated subjective assessments are different than when they had an active status in CA, but the main trends remain.

In fact, the procedure looks like that found in (Ahmed et al. 2021) where

1. active variables correspond to socio-demographic attributes. Examples for columns of ***Y*** are modalities of qualitative variables (e.g., gender or driving license

holder) and modalities obtained from a crisp windowing of a quantitative variable (e.g., age with '0 to 22', '23 to 47', '48 to 64' and '65 and above').
2. supplementary variables correspond to activity sequence pattern such as 'Home-Work-home' or 'Home-Leisure-Home'.

If the two databases are different (e.g., the statistical unit is the individual in (Ahmed et al. 2021) and one-hour time window in our study), idem form data coding (crisp and fuzzy coding respectively), both studies show that using a two stage-based approach is quite interesting for a preliminary analysis with many heterogeneous variables and observations. The first interest of scale windowing is to be able to consider both qualitative and quantitative scales in such an analysis (CA is replaced by MCA). The second interest is to reduce the information loss as soon as we have to summarize data through a factor (time, space, individual…). For instance, suppose a quantitative subjective scale between 0 and 10 and that two individual values x_1 and x_2 are available. If these two values are characterized using a central trend indicator (e.g., the arithmetic mean $x.$) we have, for two pairs of values, the following results

—x_1=1, x_2=9 -> $x.$=5
or
-> triplets $\boldsymbol{\mu}_1$=(1, 0, 0) and $\boldsymbol{\mu}_2$=(0, 0, 1) with crisp windowing using Fig. 2.a
-> $\boldsymbol{\mu}$= (0.5, 0, 0.5) which says that there were a 'low' and a 'high' values.

—x_1=4, x_2=6 -> $x.$=5 (as with the previous pair)
or
-> triplets $\boldsymbol{\mu}_1$=(0, 1, 0) and $\boldsymbol{\mu}_2$=(0, 1, 0) with crisp windowing using Fig. 2.a
-> $\boldsymbol{\mu}$= (0, 1, 0) which says that there were always 'medium' values.

With fuzzy coding using Fig. 2a, the information loss is even lower. This gives MCA another big advantage over PCA in addition to that of showing complex relational phenomena, Fig. 11a (it is interesting to note that ***Yr*** columns could be obtained both from fuzzified quantitative variables and qualitative variables linked to the space, e.g. presence/absence of accidents, failures or strikes on a given geographical area of the transport network). On the other hand, it must be recognized that fuzzy sets-based MCA is more complicated due to

1. the design of the membership functions (here, to compare with PCA, a simple approach was used, see Fig. 2a, but more sophisticated model could have been chosen),
2. the computation of the membership values (the usual standardization technique used with PCA, see Table 1b, is much faster, both for the coding choice, often a default choice, and the computing time),
3. the presence of several modality points for each variable, Fig. 9b (instead of one point for PCA, Fig. 8b).

Finally, the global principle shown Fig. 1 could be used in quite real time: CA being carried out up to the date D_{end} (corresponding frequencies are placed within a large table ***Y***, Fig. 1b), the data obtained after, for example hour by hour, could be placed in table ***Yb***, Fig. 2b, and shown as supplementary row points. Using bottom plot of Fig. 2b, it would then be possible to see if the traffic during a new hour h is close or not to the traffic obtained up to the date D_{end} for the same hours and days of the week (see Fig. 6 for possible examples obtained up to the date d). Let us suppose that during a given Friday after our studying period, there is no train departing from the station

- A at 8 a.m. and this for 1 hour,
- I (Airport) at 8 a.m. and this for 1 hour,
- I at 8 p.m. and this for 1 hour.

It is worth reminding that, for a given hour, there are 100 frequencies (***Y*** as 100 columns). With these 3 perturbation examples, only some columns are affected. In such simulated cases, table ***Yb*** has 3 rows which can be placed below the initial data table; see Fig. 2b for the general principle and Fig. 6 for CA results concerning the different Fridays. Figure 12 shows how the 3 points move from the initial position i.e. the last Friday of April. We could take the Average Trajectory of all Fridays $(AT_{(\alpha,\alpha')}(d)$; $(\alpha, \alpha') = (1, 2)$ to indicate the first main plane; $d = 5$ for 'Friday') after removing those being very far from this average as with the red

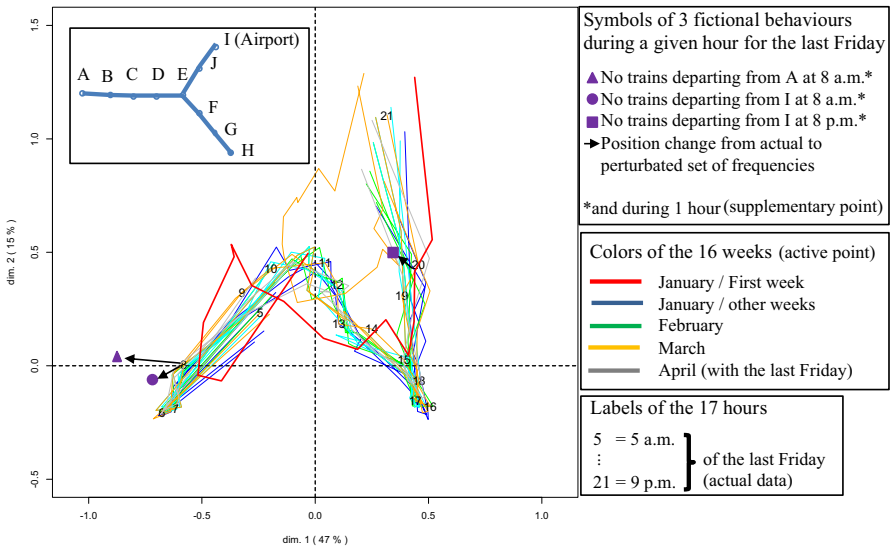

Fig. 12 How CA shows new time data in quasi-real time (example with a Friday after the 4 months). The three points in purple correspond to 3 supplementary row points of table ***Yb*** in Fig. 2b. All the other points correspond to CA active points (see table ***Y*** in Fig. 1b where only row points linked to Friday are displayed, see Fig. 6 for the 7 days)

Table 1 Basic elements of PCA (Principal Component Analysis) and CA (Correspondence Analysis) for a network traffic data set designed through several Origin-Destination subsets, one subset for a given time window (the pair (s, s') of Fig. 1 becomes here (o, d) for (origin, destination))

Features	PCA (with standardized variables)	CA
a) Generic values	$t = 1, \ldots, T$: time window $o = 1, \ldots, O$: origin; $d = 1, \ldots, D$: destination ($O = D$) $y_{t,od}$: number of network users moving from o to d during t $\boldsymbol{Y_{RxC}}$: matrix containing $y_{t,od}$ with $R = T$ rows and $C = O \times D$ columns (a comma is used to separate the row and column subscripts)	
b) Generic row point P_t in $\mathbb{R}^{O \times D}$ (named *time point*)	$P_t = \left(\frac{y_{t,od} - m_{od}}{s_{od}}, o = 1, \ldots, O; d = 1, \ldots, D \right)$ where m_{od} and s_{od} are the arithmetic mean and the standard deviation of column od	$P_t = \left(\frac{y_{t,od}}{y_{t,**}}, o = 1, \ldots, O; d = 1, \ldots, D \right)$ where $y_{t,**} = \sum_{o=1}^{O} \sum_{d=1}^{D} y_{t,od}$ is the weight of the point P_t contains a row frequency profile
c) Distance between 2 row-points: $d^2(P_t, P_{t'})$	$\sum_{o=1}^{O} \sum_{d=1}^{D} \left(\frac{y_{t,od} - y_{t',od}}{s_{od}\sqrt{T}} \right)^2$	$\sum_{o=1}^{O} \sum_{d=1}^{D} \frac{1}{y_{*,od}} \left(\frac{y_{t,od}}{y_{t,**}} - \frac{y_{t',od}}{y_{t',**}} \right)^2$
d) Generic column point Q_{od} in $\mathbb{R}^T$ (named *space point*)	$Q_{od} = (y_{t,od}, t = 1, \ldots, T)$	$Q_{od} = \left(\frac{y_{t,od}}{y_{*,od}}, t = 1, \ldots, T \right)$ where $y_{*,od} = \sum_{t=1}^{T} y_{t,od}$ is the weight of the point. Q_{od} contains a column frequency profile
e) Distance between 2 column points: $d^2(Q_{od}, Q_{o'd'})$	$2(1\text{-}cor_{od,o'd'})$ where $cor_{od,o'd'}$ is the linear correlation coefficient between the two columns od and $o'd'$	$\sum_{t=1}^{T} \frac{1}{y_{t,*}} \left(\frac{y_{t,od}}{y_{*,od}} - \frac{y_{t,o'd'}}{y_{*,o'd'}} \right)^2$
f) Generic (gen) value of the matrix on which the singular decomposition is performed (we suppose that $T > O \times D$)	$gen_{od,o'd'} = cor_{od,o'd'}$ which is the generic value of the correlation matrix $\boldsymbol{C}$.	$gen_{od,o'd'} = \sum_{t=1}^{T} \frac{y_{t,od} \times y_{t,o'd'}}{y_{t,*} \times y_{*,o'd'}}$ Rows and columns play symmetrical roles.
g) Tabular outputs	Eigenvalues (absolute value and % of the total inertia) and tables that aid interpretation (for each point P_t or Q_{od} and each main Axis α, statistics as the squared cosine or the contribution)	

Table 1 (continued)

Features	PCA (with standardized variables)	CA
h) Graphical Outputs	Projection of each point P_t or Q_{od} on a main plane obtained when crossing two main Axes α and α' in the corresponding space (if $\alpha=1$ and $\alpha'=2$, the name *first main plane* is used)	
	Q_{od} points are shown through the correlation circle (2 points close and P_t points through their projections (but biplots showing both sets can be used) (Kassambara 2017)	P_t and Q_{od} playing symmetrical roles, the 2 sets of points can be shown using the same plot (but if the number of points is large, two plots must be used) A point with a low weight may be far from the gravity center (lever arm principle)
i) Supplementary elements (represented as points in the joint row and column space, but are not used while determining the location of active points P_t and Q_{od})	*Individuals*: from initial row points P_t, many summaries can be computed and placed in table ***Yb*** (Fig. 2b), e.g. one arithmetic mean for each day or two means (before noon and afternoon). *Quantitative variables*: combinations of the initial variables or other variables can be computed and placed in table ***Yr*** (Fig. 2b), e.g. variables linked to the transport network environment (meteorological, socio-economic, dysfunction, accident, …). The corresponding points are shown through the correlation circle. *Qualitative variables*: such variables are present in ***Yr*** but the M modalities (or levels) are shown through individuals e.g. each of the $M=7$ day modalities (Monday to Sunday) yields 7 gravity centres computed from the initial row points with the corresponding modality. Each row subset that corresponds to a given modality can be shown through a specific colour.	
		Modalities: other modalities can be placed in ***Yr***.
j) Inference context	For a *qualitative variable* with M modalities (see above) and a given main Axis α, a one-way analysis of variance (ANOVA) can be performed to test whether the M gravity centres are identical or not (Husson et al. 2016). Then, for each modality m, a Student's test is performed to compare the average coordinate of individuals with modality m with the global average (of the M modalities). Finally, the p-values are sorted to obtain the most different modalities from the zero point (in both directions i.e. modalities with either very negative or very positive coordinates along Axis α). For a *quantitative variable*, it is possible to test the significance of the correlation coefficient between this variable and Axis α. Such a technique can be used with a supplementary variable with both CA and PCA or an active variable with PCA. In the latter case, sorting the contributions, see (g), of points Q_{od} to control Axis α or the p-values allow understanding which points have a main role in the positioning of a main axis For both cases, the usual level of 0.05 will be used.	

trajectory, Fig. 12. This approach can therefore be used to show changes in quite real-time.

Another possibility is to use CA main planes to show the goodness of fit of an analytical time model of the count data (here the first main plane represents 62% of the total inertia, which is a rather high value for a table with 2033 rows and 100 columns). For instance, for each day d of the week ($d = 1, \ldots, 7$), the row set from the hourly frequencies yielded by the model is placed in table $\boldsymbol{Yb}$, then the distance between the time trajectory pattern from estimated data and $AT_{(\alpha,,\alpha,')}(d)$ is appreciated using Fig. 6.

6 Conclusion

While PCA was specifically designed for several QT variables and CA for 2 QL variables, the bibliographic analysis showed that these two methods were used for frequency data relating to origin-destination links. Little comparative analysis having been carried out in this field, it seemed interesting to perform such an analysis using actual data. The main result is that CA graphical output is more complex than PCA output but reveal more interesting result, e.g. a specific behaviour linked to the airport with CA and low vs. high frequency clusters with PCA. In fact, it could be interesting to perform both PCA and CA but it would generate twice as much time (approximately since CA takes more time than PCA).

Another interesting point of CA was the possibility to incorporate QT variables related to the network environment as extra data (here weather and gas price) using fuzzy windowing, such a scale change being much less used that with crisp windowing using Multiple Correspondence Analysis.

Obviously, it would be necessary, on the one hand, to extend the comparative analysis with other frequency data (e.g. by taking the day instead of the hour as time window or by considering another data set) and, on the other hand, to add experts to assess the pros and cons of PCA and CA (e.g., by using the Kappa Statistics).

Author Contributions Pierre Loslever analyzed the literature, the data and wrote the manuscript.

Funding None.

Availability of Data and Materials Data can be obtained from the author.

Declarations

Ethical Approval Not applicable.

Competing Interests The authors declare no competing interests.

References

Ahmed U, Moreno AT, Moeckel R (2021) Microscopic activity sequence generation: a multiple correspondence analysis to explain travel behavior based on socio-demographic person attributes. Transp (Amst). https://doi.org/10.1007/s11116-020-10103-1

Alonso B, Ibeas A, Musolino M, Rindone C, Vitetta A (2019) Effects of traffic control regulation on network macroscopic fundamental diagram: a statistical analysis of real data. Transp Res Part A Policy Pract 126:136–151

Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27(1):17–21. https://doi.org/10.1080/00031305.1973.10478966.JSTOR2682899

Antolini F, Giusti GA (2021). Tourism of Italians in Italy through crisis and development: the last 15 years, region by region. In: Bertaccini B, LFabbris, Petrucci (eds) ASA 2021 statistics and information systems for policy evaluation, pp 239–244. https://doi.org/10.36253/978-88-5518-461-8.45

Arslan T (2009) A hybrid model of fuzzy and AHP for handling public assessments on transportation projects. Transp (Amst). https://doi.org/10.1007/s11116-008-9181-9

Awad FA, Graham DJ, AitBihiOuali, Singh R, Barron A (2023) Benchmarking the performance of urban rail transit systems: a machine learning application. Transp A: Transp Sci. https://doi.org/10.1080/23249935.2023.2241566

Beh EJ, Lombardo R (2014) Correspondence analysis. Theory, practice and new strategies. Wiley, Chichester

Bellanger L, Tomassone R (2014) Exploration de données et méthodes statistique: data analysis & data mining avec le logiciel R. Ellipses, Paris

Benzecri JP (1992) Correspondence analysis handbook. Marcel Dekker, New York

Blasius J, Greenacre M (2006) Correspondence analysis and related methods. In: Blasius J, Greenacre M (eds) Multiple correspondence analysis and related methods. Chapman and Hall, London, pp 3–40

Burges CJ (2009) Geometric methods for feature extraction and dimensional reduction - a guided tour. In: Maimon O, Rokach L (eds) Data Mining and Knowledge Discovery Handbook, Springer. https://doi.org/10.1007/978-0-387-09823-4_4

Cabral L, Kim A (2022) An empirical reappraisal of the level of traffic stress framework for segments. Travel Behav Soc. https://doi.org/10.1016/j.tbs.2021.09.007

Card S, Mackinlay J, Shneiderman B (1999) Readings in information visualization: using vision to think. Academic Press, San Diego

Chahouki M (2012) Classification and ordination methods as a tool for analyzing of plan communities. In: de Freitas L, de Freitas A (eds) Multivariate analysis in management, engineering and the sciences. Intech. https://doi.org/10.5772/54101

Chawla S, Zheng Y, Hu J (2012) Inferring the root cause in road traffic anomalies. In: 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, pp 141–150. https://doi.org/10.1109/ICDM.2012.104

Chen RB (2018) Models of count with endogenous choices. Transp Res Part B Methodol 117:862–875

Chen W, Guo F, Wang FY (2015) A survey of traffic data visualization. IEEE Trans Intell Transport Syst 16(6):2970–2984

Chen W, Lei Y (2017) Path analysis of factors in energy-related CO2 emissions from Beijing's transportation sector. Transp Res Part D Transp Environ 50:473–487

Cheng Z, Wang W, Lu J, Xing X (2018) Classifying the traffic state of urban expressways: a machine-learning approach. Transp Res Part A Policy Pract. https://doi.org/10.1016/j.tra.2018.10.035

Chrisman N (1998) Rethinking levels of measurement for cartography. Cartogr Geogr Infor Sc 25:231–242

Cornillon P (2018) R pour la statistique et la science des données. Presses Universitaires de Rennes, Rennes

Cottrill C, Thakuriah P (2015) Location privacy preferences: a survey-based analysis of consumer awareness, trade-off and decision-making. Transp Res Part C Emerg Technol 56:132–148

Crawford F, Watling DP, Connors RD (2017) A statistical method for estimating predictable differences between daily traffic flow profiles. Transp Res Part B Methodol 95:196–213

Diana M, Pronello C (2010) Traveler segmentation strategy with nominal variables through correspondence analysis. Transp Policy 17:183–190

Djukic TG, Lint FH, Hoogendoorn S (2012). Efficient real time OD matrix estimation based on principal component analysis. In: 15th international IEEE conference on intelligent transportation systems. pp 115–121. https://doi.org/10.1109/ITSC.2012.6338720

Egu O, Bonnel P (2020) Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon. Travel Behav Soc 19:112–123

Exel NJ, Rietveld P (2009) When strike comes to town… anticipated and actual behavioural reactions to a one-day, pre-announced, complete rail strike in the Netherlands. Transp Res Part A Policy Pract 43(5):526–535

Farber S, Ritter B, Fu L (2016) Space–time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: a social equity perspective. Travel Behav Soc 4:40–48

Fekih M, Bellemans T, Smoreda Z et al (2021) A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France). Transp (amst) 48:1671–1702

Friendly M, Meyer D (2016) Discrete data analysis with R: visualization and modelling techniques for categorical and count data. Chapman & Hall / CRC, London

Gao J, Zhang YC, Zhou T (2019) Computational socioeconomics. Phys Rep 817:1–104

Guardiola IG, Leon T, Mallor F (2014) A functional approach to monitor and recognize patterns of daily traffic profiles. Transp Res Part B Methodol 65:119–136

Gueguin M et al (2008) Exploring time series retrieved from cardiac implantable devices for optimizing patient follow-up. IEEE Trans Biom Eng 55(10):2343–2352

Guex G, Loup R, Bavaud F (2023) Estimation of flow trajectories in a multi-lines transportation network. Appl Netw Sci 8(1). https://doi.org/10.1007/s41109-023-00570-7

Hong L, Ye B, Yan H, Zhang H, Ouyang M, He X (2019) Spatiotemporal vulnerability analysis of railway systems with heterogeneous train flows. Transp Res Part A Policy Pract 130:725–744

Hsu L, Culhane A (2023) Correspondence analysis for dimension reduction, batch integration, and visualization of single-cell RNA-seq data. Sci Rep 13(1197). https://doi.org/10.1038/s41598-022-26434-1

Husson F, Lé S, Pagés J (2016) Analyse de données avec R. Presses Universitaires de Rennes, Rennes

Iseki H, Liu C, Knaap G (2018) The determinants of travel demand between rail stations: a direct transit demand model using multilevel analysis for the Washington D.C. Metrorail system. Transp Res Part A Policy Pract 116:635–649

Jain D, Tiwari G (2019) Explaining travel behaviour with limited socio-economic data: case study of Vishakhapatnam, India. Travel Behav Soc 15:44–53

Jalayer M, Zhou H (2016) A multiple correspondence analysis of at-fault motorcycle-involved crashes in Alabama. J Adv Transp 50:2089–2099

Jobson JD (2012) Applied multivariate data analysis. Springer, New York

Kassambara A (2017) Practical guide to principal component methods in R. STHDA.com

Khoo HL, Asitha KS (2016) An impact analysis of traffic image information system on driver travel choice. Transp Res Part A Policy Pract 88:175–194

Klingen J (2019) Do metro interruptions increase the demand for public rental bicycles? Evidence from Paris. Transp Res Part A Policy Pract 123:216–228

Kolaczyk ED, Csardi G (2014) Statistical analysis of network data with R. Springer, New York

Krishnakumari P, Cats O, Van Lint H (2020). A compact and scalable representation of network traffic dynamics using shapes and its applications. Transp Res Part C Emerg Technol. https://doi.org/10.1016/j.trc.2020.102850

Kumar A, Vijaya Saradhi V, Venkatesh T (2014) Role of correspondence analysis in network traffic flow analysis. Proceedings of the 6th IBM Collaborative Academia Research Exchange Conference (I-CARE) on I-CARE 2014 (I-CARE 2014), New York, ACM, pp 1–4

Lebart L, Piron M, Morineau A (2006) Statistique exploratoire multidimensionnelle: visualisations et inférences en fouille de données. Dunod, Malakoff

Marcucci E, Gatta V (2017) Investigating the potential for off-hour deliveries in the city of Rome: retailers' perceptions and stated reactions. Transp Res Part A Policy Pract 102:142–156

Mattioli G, Anable J (2017) Gross polluters for food shopping travel: an activity-based typology. Travel Behav Soc 6:19–31

McCarthy O, Caulfield B, O'Mahony O (2016) Technology engagement and privacy: a cluster analysis of reported social network use among transport survey respondents. Transp Res Part C Emerg Technol 63:195–206

Montero L, Ros-Roca X, Herranz R, Barceló J (2019) Fusing mobile phone data with other data sources to generate input OD matrices for transport models. Transp Res Proc 37:417–424

Nishisto S (2007) Multidimensional nonlinear descriptive analysis. Taylor & Francis Group, Boca Raton
Nosal T, Miranda-Moreno LF (2014) The effect of weather on the use of North American bicycle facilities: a multi-city analysis using automatic counts. Transp Res Part A Policy Pract 66:213–225
Pagès J (2013) Analyse factorielle multiple avec R. EDP sciences, Les Ulis
Parry K, Hazelton ML (2012) Estimation of origin–destination matrices from link counts and sporadic routing data. Transp Res Part B Methodol 46(1):175–188
Parry K, Hazelton ML (2013) Bayesian inference for day-to-day dynamic traffic models. Transp Res Part B Methodol 50:104–115
Pruscha H (2013) Statistical analysis of climate series: analyzing, plotting, modelling and predicting with R. Springer, London
Rostami-Nasab M, Shafahi Y (2020) Estimation of origin–destination matrices using link counts and partial path data. Transportation (Amst). https://doi.org/10.1007/s11116-019-09999-1
Ryder B, Dahlinger A, Gahr B, Zundritsch P, Wortmann F, Fleisch E (2019) Spatial prediction of traffic accidents with critical driving events – insights from a nationwide field study. Transp Res Part A Policy Pract 124:611–626
Saporta G (2011) Probabilités, analyse des données et Statistique. Technip, Paris
Schmöcker JD, Su F, Noland RB (2010) An analysis of trip chaining among older London residents. Transp (Amst). https://doi.org/10.1007/s11116-009-9222-z
Stevens SS (1974) Measurement scaling: a sourcebook for behavioral scientists. Aldine Publishing Co, Chicago
Taylor P (2006) Statistical methods. In: Berthold M, Hand DH (eds) Intelligent data analysis. Springer-Verlag, Berlin, pp 69–129
Truong L, Somenahalli S (2015) Exploring frequency of public transport use among older adults: a study in Adelaide. Australia Travel Behav Soc 2(3):148–155
Tsiotas D, Tselios V (2023) Dimension reduction in the topology of multilayer spatial networks: the case of the interregional commuting in Greece. Netw Spat Econ 23:97–133
Tu H, Li H, van Lint H, van Zuylen H (2012) Modeling travel time reliability of freeways using risk assessment techniques. Transp Res Part A Policy Pract 46(10):1528–1540
Van Lierop D, Bahamonde-Birke FJ (2023) Commuting to the future: assessing the relationship between individuals' usage of information and communications technology, personal attitudes, characteristics and mode choice. Netw Spat Econ 23(2):353–371
Volle M (1997) Analyse des données. Economica, Paris
Wan L, Yang T, Jin Y et al (2021) Estimating commuting matrix and error mitigation – a complementary use of aggregate travel survey, location-based big data and discrete choice models. Travel Behav Soc 25:102–111
Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer-Verlag, New-York
Yin B, Leurent F (2022) What are the multimodal patterns of individual mobility at the day level in the Paris region? A two-stage data-driven approach based on the 2018 Household Travel Survey. Transp (Amst). https://doi.org/10.1007/s11116-022-10285-w
Younsi K, Loslever P, Popieul JC, Simon P (2011) Fuzzy segmentation for the exploratory analysis of multidimensional signals. Example from a study on driver overtaking behavior. IEEE Syst Man Cybern (Part A) 41(5):892–904

Authors and Affiliations

Pierre Loslever[1,2]

✉ Pierre Loslever
Pierre.Loslever@uphf.fr

[1] Laboratoire d'Automatique, de Mécanique et d'Informatique industrielles et Humaines, UMR CNRS 8201, Université Polytechnique Hauts-de-France, Campus Mont-Houy, 59313 Valenciennes Cedex 9, France

[2] LAMIH, Polytechnic University of Hauts de France, Valenciennes, France