



Using space windowing for a preliminary analysis of complex time data in human component system studies. Examples with eye-tracking in advertising and car/head movements in driving

P. Loslever^{*}, P. Simon, F. Rousseau, J.C. Popieul

LAMIH, University of Valenciennes, Le Mont Houy, 59313 Valenciennes Cedex 9, France

ARTICLE INFO

Article history:

Received 1 November 2007

Received in revised form 16 May 2008

Accepted 23 May 2008

Keywords:

Fuzzy windowing

Information

Correspondence analysis

Exploratory statistics

Data meaning

Membership value table

Time analysis

Eye movement

Advertising

Car driving

Vigilance

ABSTRACT

Empirical studies of human systems often involve recording multidimensional signals because the system components may require physical measurements (e.g., temperature, pressure, body movements and/or movements in the environment) and physiological measurements (e.g., electromyography or electrocardiography). Analysis of such data becomes complex if both the multifactor aspect and the multivariate aspect are retained. Three examples are used to illustrate the role of fuzzy space windowing and the large number of data analysis paths. The first example is a classic simulated data set found in the literature, which we use to compare several data analysis paths generated with principal component analysis and multiple correspondence analysis with crisp and fuzzy windowing. The second example involves eye-tracking data based on advertising, with a focus on the case of one category variable, but with the possibility of several space windowing models and time entities. The third example concerns car and head movement data from a driving vigilance study, with a focus on the case involving several quantitative variables. The notions of analysis path multiplicity and information are discussed both from a general perspective and in terms of our two real examples.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Time data recorded in human component system (HCS) studies can be relatively complex, depending on the number of steps needed to go from the raw values to final results that can be exploited by potential users (e.g., physicians, psychologists, ethologists, ergonomists, economists or engineers). The primary specificities of the time data generated in HCS studies are described below:

- (1) HCS studies yield intra- and inter-individual differences which are difficult to reduce whatever the method – experimental and observational design [16,17] – used to collect data. For instance a reduction is, maybe, possible when using a very specific individual sample (e.g., high level experts) in an experimental design in psychology or maybe with a very specific population (e.g., 30-year-old finance managers in the USA) in an observational design in sociology.

^{*} Corresponding author. Tel.: +33 6 81 16 55 53; fax: +33 3 27 51 13 16.

E-mail address: loslever@univ-valenciennes.fr (P. Loslever).

- (2) Movements, tasks, or any informational exchanges between the human and/or technological components of the HCS [7,13,15,18,21] yield time data sets which are difficult to analysis together, whatever the *time* notion – duration, chronology and simultaneity [4] – maintained in the analysis.
- (3) Result sets mainly contain relationships between the variables describing the HCS performance and the influences of the factors on the variables which are difficulty to present in a standardized way whatever the modality – mathematical, graphical or verbal [25] – used for modeling.

Given these three points, HCS literature, especially in psychology, ergonomics and medicine, often contains the following peculiarities:

- (1) Whether the study uses an experimental or observational design, there is a disproportionate tendency towards averaging over the individual factor, particularly when considering inter-individual differences, such as random noises in statistical tests.
- (2) Regardless of the three time notions mentioned above, there is a tendency to compute classic indicators (e.g., the arithmetic mean, standard deviation, RMS value, and/or extreme value) to describe and compare time data (e.g., when using a statistical test). Thus, none of the three time notions is evoked. Of course, there are exceptions to this generality, since, for example, when a test is performed to compare two (or more) arithmetic means, with each mean representing a period of observation during the HCS study, the notion of chronology is obviously present.
- (3) Whatever the kind of model used for presenting results, there is a tendency to investigate the variables one by one, thus making it impossible to show how they evolve together (i.e., to take simultaneity into account in the analysis).

Since most HCS studies have at least two factors – individual and time – and that they yield several variables, this paper considers multifactor and multivariate (MFMV) database analysis, in which the three kinds of models are included for the very first analysis of such a database. Before proceeding further, several terms need to be defined:

- The term *complex data* is used for data with MFMV aspects, which are considered in the beginning of the analysis as much as possible. In addition, data heterogeneity may also be considered since there are several scale mathematical models [22].
- The term *analysis* has a different meaning from the term *processing*. Analysis is seen as a series of processing; each processing in the series is preceded and followed by a human-centered phase in which the human chooses the processing method and spends most of the time examining the processing output.
- The term *database* is different from the general sense used in the information/computer sciences. The database generated by an HCS study with an experimental or observational design primarily contains quantitative time data sets that are all situated in the same computer (one may be two) but in different files (i.e., one set of files per individual in most cases). This is quite different from a database or a data warehouse that has many, very large, well organized and distributed data sets [11].

The rest of this article is organized as follows. Section 2 presents some generalities about the many ways that MFMV data can be analyzed. This data may refer to several individuals and pertain to an experimental or an observational design. Given the large number of data processing methods, we chose to focus on the methods for preliminary analysis, often called *exploratory analysis*. Such data exploration maintains the MFMV aspects of the analysis. Section 3 focuses on the characterization of high complexity data. Sections 4 and 5 introduce two applications, the first one considering just one category variable, the second one considering several quantitative variables. The last Section 6 discusses the analysis methods considered in this paper, providing more information and several methods for continuing the analysis from a more inferential perspective.

To facilitate comprehension, the following notational system is used throughout the paper:

- An outlined letter designates a set (e.g., set \mathbb{E}).
- If an initial set can be transformed, the tag “0” designates this set (e.g., set $\mathbb{E}0$), and the tag “1” designates a set from a first transformation (e.g., set $\mathbb{E}1$), and so on.
- An uppercase letter designates the size of the set (e.g., $E0 = \text{card}(\mathbb{E}0)$), where *card* is the cardinality function.
- A boldface letter designates a vector or a matrix (e.g., $\mathbf{x} = (x_1, x_2, x_3)'$ or $\mathbf{X} = \mathbf{x} \cdot \mathbf{x}'$).

Given these basic notations, $\mathbb{U} = \{X_u, u = 1, \dots, U\}$ and $\mathbb{V} = \{Y_v, v = 1, \dots, V\}$ are, respectively, the set of factors and the set of variables; *i* is the subscript indicating the individual (i.e., a participant to the HCS study); and *s*, *t* and *f* are, respectively, the subscripts indicating a window on the space, time or frequency axis. In the rest of this paper, all magnitudes not linked specifically to the time and frequency axes (labeled *t* and *f*) or to the individual (*i*) are called space variables, or factors.

2. General procedure

HCS studies can be performed in many domains (e.g., medicine, psychology, ecology, ergonomics). Although the measurement devices are quite different, the phases of the procedure remain the same: phase 1 is data acquisition and phase 2 is data analysis.

2.1. Data acquisition

The data acquisition phase has four main steps:

Step 1.1: Define the objectives. The objectives of the study are defined by taking into account the results of previous studies through a review of the existing literature and finding some underlying deficiency in these results. In most cases, the main objective is to determine the influences of the various factors on the HCS behavior. Let \mathbb{U} be the set of factors (except the time factor), where the first element, X_1 , designates the individual factor (e.g., with I participants in the study, the individual i ($i = 1, \dots, I$) yields I levels for X_1).

Step 1.2: Build new measurement devices, or implement existing ones. These devices must be able to give indications about the HCS behavior, so their design is important. In most cases, one measurement device produces one measurement variable, but sometimes this is not the case (e.g., a 3D imaging system used with M markers placed on the human body and its surroundings yields $3 \times M$ variables for the respective positions on the x, y and z axes). Let \mathbb{V} be the set of measurement variables.

Step 1.3: Define the space-time organization. Let us consider an HCS study with an experimental design and two experimental factors in addition to the individual factor. First, the numbers and characteristics of the I, J and K levels must be defined. Then, in the case of a full experimental design in which each individual i tests each pair (j, k) , the way that the $B = I \times J \times K$ triplets are combined must be defined. If a full experimental design is not used, the way that each pair (j, k) is chosen for each individual i must be defined a priori. On the other hand, in the case of an observational design, the factors are most often defined a posteriori; for example, one factor is assigned for the individual's sex, one for age, one for level of experience in the studied activity (e.g., driving, smoking, working), and one for country of residence.

Step 1.4: Record and organize the data sets. If $V_0 > 1$, there are multidimensional signals. Regardless of whether the study has an experimental or observational design, these signals can be organized as a hyperparallelepipedic structure \mathbb{H} in which the directions correspond to the factors. Fig. 1 provides an example with 3 factors. Of course, among the potential $B = I \times J \times K$ triplets, some cells may be empty. Let h ($h = 1, \dots, H_0$) be a cell corresponding to an existing triplet (i, j, k) . In this case, for a complete factorial experiment involving repeated measurements, $H_0 = B$; otherwise, $H_0 < B$. Let $y_{ijkn,v}$ ($n = 1, \dots, N_{ijk,v}$) be the current value of the n th time sample of the v th measurement variable. In some cases, n may range from 1 to an identical value, named N_0 , for all the H_0 cells. Please note that H_0, V_0 or N_0 can be reduced in the statistical analysis for a variety of reasons. For instance, due to a malfunction of the measurement device, one variable could always give the wrong data, making $V_1 = V_0 - 1$. Or, because of their very odd behavior, two individuals might have to be removed from the analysis, in which case $H_1 = (I - 2) \times J \times K$.

To summarize this 4-step procedure, the input is a set of study objectives and the output is a hyperparallelepiped \mathbb{H} with MFMV aspects. The second (and final) phase involves analyzing the H_0 data sets within \mathbb{H} .

2.2. Data analysis

Due to the large number of data sets (H_0 is often greater than 100) and their complexity (the data sets often contain several variables and factors, including the individual and time factors), several steps and loops are needed for the data analysis phase. The main steps are:

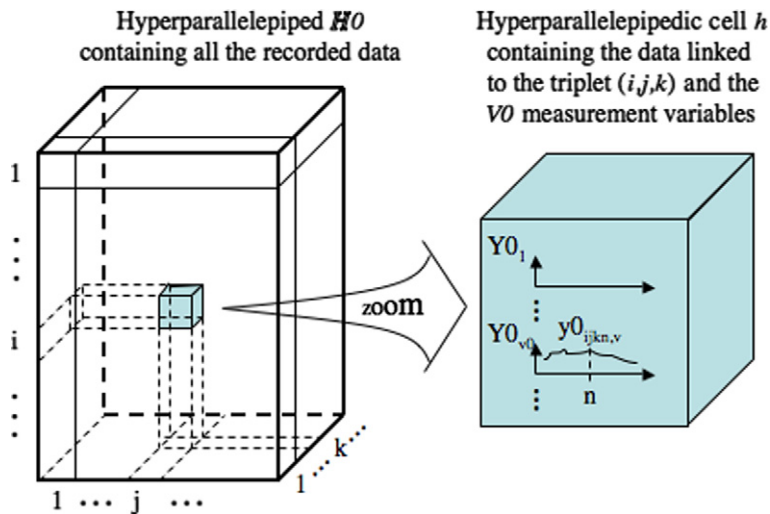


Fig. 1. Example of a (hyper)parallelepipedic data structure in which 3 factors, among them the 'individual' factor, are considered. The time factor, with its reference to *chronology*, is represented through the subscript n .

Step 2.1: *Data characterization*. Summarizing indicators are computed. These indicators can be grouped in two principal families: statistical and signal analysis. For each set of NO_{ijk} time samples, either one or several indicators may be considered. For example, the data analyst can consider T indicators with time windowing (if each time window t yields one indicator, such as the arithmetic mean), S indicators with space windowing (from a magnitude histogram or any binning technique [27]), F indicators with frequency windowing (from a amplitude spectrum), $S*T$ indicators with space-time windowing or $T*F$ indicators with time-frequency windowing [1,2,28].

Step 2.2: *Scale transformation*. The scales are transformed so that all the indicators computed in step 2.1 can be compared, for example, using the classic standardization technique “ $(Y_v - a)/b$ ”, where a and b correspond to the arithmetic mean and the standard deviation. Of course, this step may be skipped if all the data from step 2.1 can be compared without transformation (e.g., all the values are probabilities within a given space window s).

Step 2.3: *Table building*. In most cases, a two-entry table with R rows and C columns is built, where the rows correspond to the $H0$ triplets (i,j,k) and the columns correspond to the $V1$ indicators obtained in the previous steps.

Step 2.4: *Table analysis*. The tables built in the previous step are then analyzed to highlight the most salient phenomena. The analysis methods can be grouped according to three main taxonomic dimensions, each one with two levels: mono-variate vs. multivariate, descriptive vs. inferential, and temporal vs. non-temporal (here, time refers to chronology, as in time series). By combining these 3 pairs of levels, 8 families can be distinguished. For instance, if each set of NO_{ijk} time samples is summarized using an arithmetic mean, and each variable is considered one by one, classic variance analysis can be used to assess the influences of the factors; this method is a monovariate inferential non-temporal method. Please note that, in most cases, if this triplet is used with such global indicators, both the time and individual factors disappear. For example, in the ANOVA model, the variation due to individuals is considered to be residual. However, if moving averages are considered individual by individual, the time and individual factors remain.

Step 2.5: *Result presentation*. Obviously, each of the methods referred to in step 2.4 have their own presentation models and can be presented mathematically (e.g., $Y = f(X)$, $Y = f(\text{time})$), graphically (e.g., histogram, scatterplot, dendrogram, factor plane, gray level picture [2,23,24]) or verbally (e.g., the answer to a hypothesis test) [10,12,17]. Nevertheless, such specific outputs are rarely considered just as they are. For example, several outputs of a principal component analysis – the tables that aid in the interpretation of the main factorial planes – are explained verbally.

To summarize this 5-step procedure, the input is an hyperparallelepiped $H0$ containing MFMV aspects (including both the individual and time factors) and the output is a set of results (whose MFMV aspects may be obvious or subtle) that can be presented mathematically, graphically or verbally. This general procedure can be used to analyze simple databases, but it is also applicable in more complex cases, such as the one presented in Section 3, in which each hyperparallelepipedic cell of $H0$ contains complex time data.

3. The case of complex time data

Complexity is the rule whenever the HCS study attempts to analyze the choices, strategies and behaviors [9,10] of the human component in terms of the MFMV data. In this context, the nuances of the space variable $Y0_v$, the chronological aspect of the time factor t , and the individual factor i come into play. First let us consider a didactic example to illustrate the advantages of fuzzy space windowing.

3.1. Didactic example

We consider the didactic Anscombe data set [3] shown in Table 1a. The initial Anscombe set was a quartet of 11 situations with 2 variables (X and Y), where (1) $(X,Y) = (V1,V2)$, (2) $(X,Y) = (V1,V3)$, (3) $(X,Y) = (V1,V4)$ and (4) $(X,Y) = (V5,V6)$.

Though the four bivariate sets are unique, the following data summaries for all four of these sets are identical, namely:

- (1) Arithmetic mean (am) of X_s = 9.0, arithmetic mean of Y_s = 7.5.
- (2) Standard deviation (sd) of X_s = 3.16, standard deviation of Y_s = 1.94.
- (3) Regression equation: $Y = 3 + 0.5X$, correlation coefficient = 0.82 (further results and commentary can be found in [3]).

Since the time factor was not included in the initial Anscombe data set, the $N=11$ observations were rearranged according to the increasing order of $V1$ with $I1$ corresponding to the first time sample and $I11$ to the last time sample. In addition, instead of considering 4 sets of two variables, a set of $V=6$ variables was considered (see Fig. 2). Please note that this example can be seen as an illustration of a set of five quantitative variables and one qualitative variable. Instead of summarizing the data using the classic arithmetic mean and standard deviation, the data was characterized using scale windowing. The 6 ranges were cut into 3 space windows of identical width, first using crisp membership functions and then using fuzzy membership functions (Fig. 3). For a given time sample n ($n=1,\dots,N=11$) and a given variable v ($v=1,\dots,V=6$), the output of the space windowing was $\mu_s(v,n)$, where s indicates the space window ($s=1,\dots,S=3$ for each variable). To maintain the statistical context, the membership values had to add up to 1 across all levels of the S windows for each time sample n .

Table 1

Data set inspired from the Anscombe example [3]: (a) 6 quantitative variables, (b) crisp windowing, (c) fuzzy windowing indicated in Fig. 3 (am = arithmetic mean, sd = standard deviation, MVA = membership value average)

(a) Initial data																		
IND	V1			V2			V3			V4			V5			V6		
I1	4.00			4.26			3.10			5.39			19.00			12.50		
I2	5.00			5.68			4.74			5.73			8.00			6.89		
I3	6.00			7.24			6.13			6.08			8.00			5.25		
I4	7.00			4.82			7.26			6.42			8.00			7.91		
I5	8.00			6.95			8.14			6.77			8.00			5.76		
I6	9.00			8.81			8.77			7.11			8.00			8.84		
I7	10.00			8.04			9.14			7.46			8.00			6.58		
I8	11.00			8.33			9.26			7.81			8.00			8.47		
I9	12.00			10.84			9.13			8.15			8.00			5.56		
I10	13.00			7.58			8.74			12.74			8.00			7.71		
I11	14.00			9.96			8.10			8.84			8.00			7.04		
a.m.	9.00			7.50			7.50			7.50			9.00			7.50		
s.d.	3.16			1.94			1.94			1.94			3.16			1.94		
(b) Membership values with crisp coding																		
ND	V1	V1	V1	V2	V2	V2	V3	V3	V3	V4	V4	V4	V5	V5	V5	V6	V6	V6
	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
1	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	1
2	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
3	1	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0
4	1	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	1	0
5	0	1	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0	0
6	0	1	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0
7	0	1	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0	0
8	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	0	1	0
9	0	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0
10	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	0	1	0
11	0	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0
a.m.	0.36	0.27	0.36	0.27	0.45	0.27	0.18	0.09	0.73	0.73	0.18	0.09	0.91	0.00	0.09	0.55	0.36	0.09
(c) Membership values with fuzzy coding																		
ND	V1	V1	V1	V2	V2	V2	V3	V3	V3	V4	V4	V4	V5	V5	V5	V6	V6	V6
	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
1	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00
2	1.00	0.00	0.00	0.85	0.15	0.00	0.70	0.30	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.82	0.18	0.00
3	0.90	0.10	0.00	0.14	0.86	0.00	0.02	0.98	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
4	0.60	0.40	0.00	1.00	0.00	0.00	0.00	0.48	0.53	1.00	0.00	0.00	1.00	0.00	0.00	0.40	0.60	0.00
5	0.30	0.70	0.00	0.27	0.73	0.00	0.00	0.05	0.95	0.94	0.06	0.00	1.00	0.00	0.00	1.00	0.00	0.00
6	0.00	1.00	0.00	0.00	0.43	0.57	0.00	0.00	1.00	0.80	0.20	0.00	1.00	0.00	0.00	0.01	0.99	0.00
7	0.00	0.70	0.30	0.00	0.78	0.22	0.00	0.00	1.00	0.66	0.35	0.00	1.00	0.00	0.00	0.95	0.05	0.00
8	0.00	0.40	0.60	0.00	0.65	0.35	0.00	0.00	1.00	0.51	0.49	0.00	1.00	0.00	0.00	0.17	0.83	0.00
9	0.00	0.10	0.90	0.00	0.00	1.00	0.00	0.00	1.00	0.37	0.63	0.00	1.00	0.00	0.00	1.00	0.00	0.00
10	0.00	0.00	1.00	0.00	0.99	0.01	0.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.48	0.52	0.00
11	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.07	0.94	0.09	0.91	0.00	1.00	0.00	0.00	0.76	0.24	0.00
a.m.	0.35	0.31	0.35	0.30	0.42	0.29	0.16	0.17	0.67	0.67	0.24	0.09	0.91	0.00	0.09	0.60	0.31	0.09

For each variable, the $N \times S = 33$ membership values and the $S = 3$ membership value averages (MVA) are indicated in Table 1. Though the classic pair (am, sd) was originally identical for some variables (Table 1a), the triplets of membership value averages (MVA1, MVA2, MVA3) all ended up being different. Thus, space windowing appears to characterize a data set better than the classic statistical indicator pair. Let us examine the three multivariate data set of Table 1. Table 1a can be investigated using principal component analysis (PCA), and Tables 1b and c using multiple correspondence analysis (MCA) [5,12].

3.1.1. PCA results

The input table has $R = N = 11$ rows and $C = V = 6$ columns. The variables being normalized ($am = 0, sd = 1$), the total variance is $V = 6$ (or 100%). Remember that if the 6 variables were linearly independent, each principal component would have shown an identical variance (i.e., $100/6 = 16.6\%$). The higher the relative variance of a principal component, the stronger the relationships (keeping in mind that strength of the relationship is explained using the linear correlation coefficient lcc). The first principal component here explains 65% of the total sample variance, which means that this component is a linear combination of more than two variables. The first two principal components, collectively, explain 85% of the total sample variance. Consequently, sample variation is well summarized by 2 principal components and a reduction of the data from 11

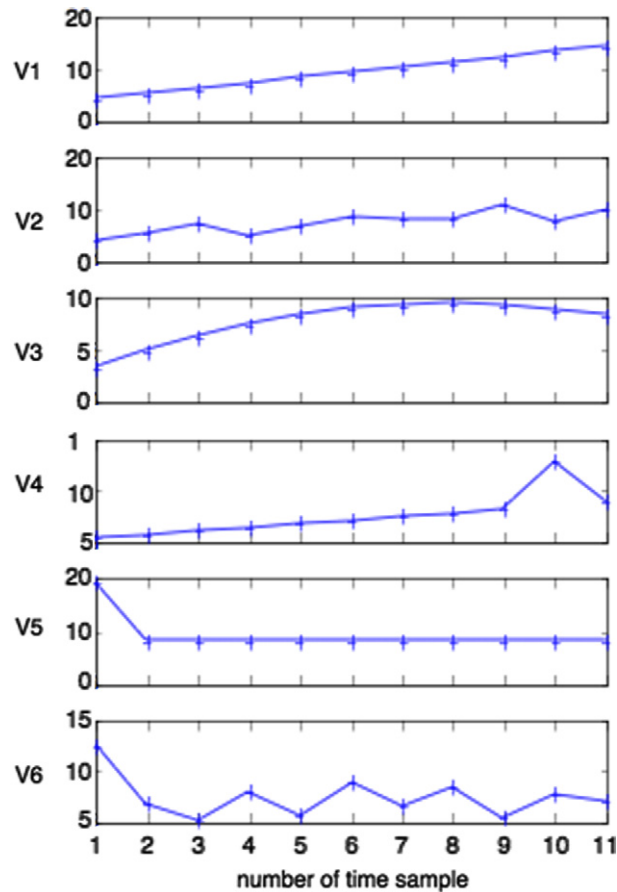


Fig. 2. Time evolutions for the 6 variables.

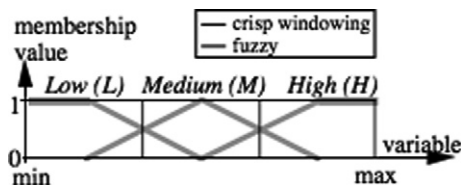


Fig. 3. Crisp and fuzzy windowing.

observations on 6 variables to 11 observations on 2 principal components is reasonable. Figs. 4a and 5a confirm this result since the correlation between each of the 6 variables and either axis 3 or axis 4 is rather low.

The first component mainly decreases with time (Fig. 4b), while the second component appears to be essentially an opposition between a subset containing the time samples 1 and 10 and a subset containing the time samples 2 and 3.

3.1.2. MCA results

The input table has $R = N = 11$ rows and $C = V \times S = 18$ columns. The notion of variance is replaced with the notion of inertia: for the R points corresponding to the rows, the mass of a point r is μ_r , the center of gravity is $G_R = (\mu_c, c = 1, \dots, C)$, the distance between two points is based on the chi-squared, and the total inertia is I_R . The same is true for the C points corresponding to the C columns. Moreover, $I_C = I_R$.

For each variable v ($v = 1, \dots, V = 6$), the $S = 3$ space windows are linked in increasing order. The 6 trajectories are named *space trajectories*, which are comparable to the *time trajectories* obtained when linking the time samples. Most space trajectories present a “V” or “Λ” pattern (Fig. 4c). This is logical with MCA when quantitative scales are changed into ordinal scales: the first axis opposes the two extreme windows (here *Low* and *High*), while the second axis opposes these windows to the medium window (such a window effect is often named Guttman's effect [5]). As stated in the level arm principle in

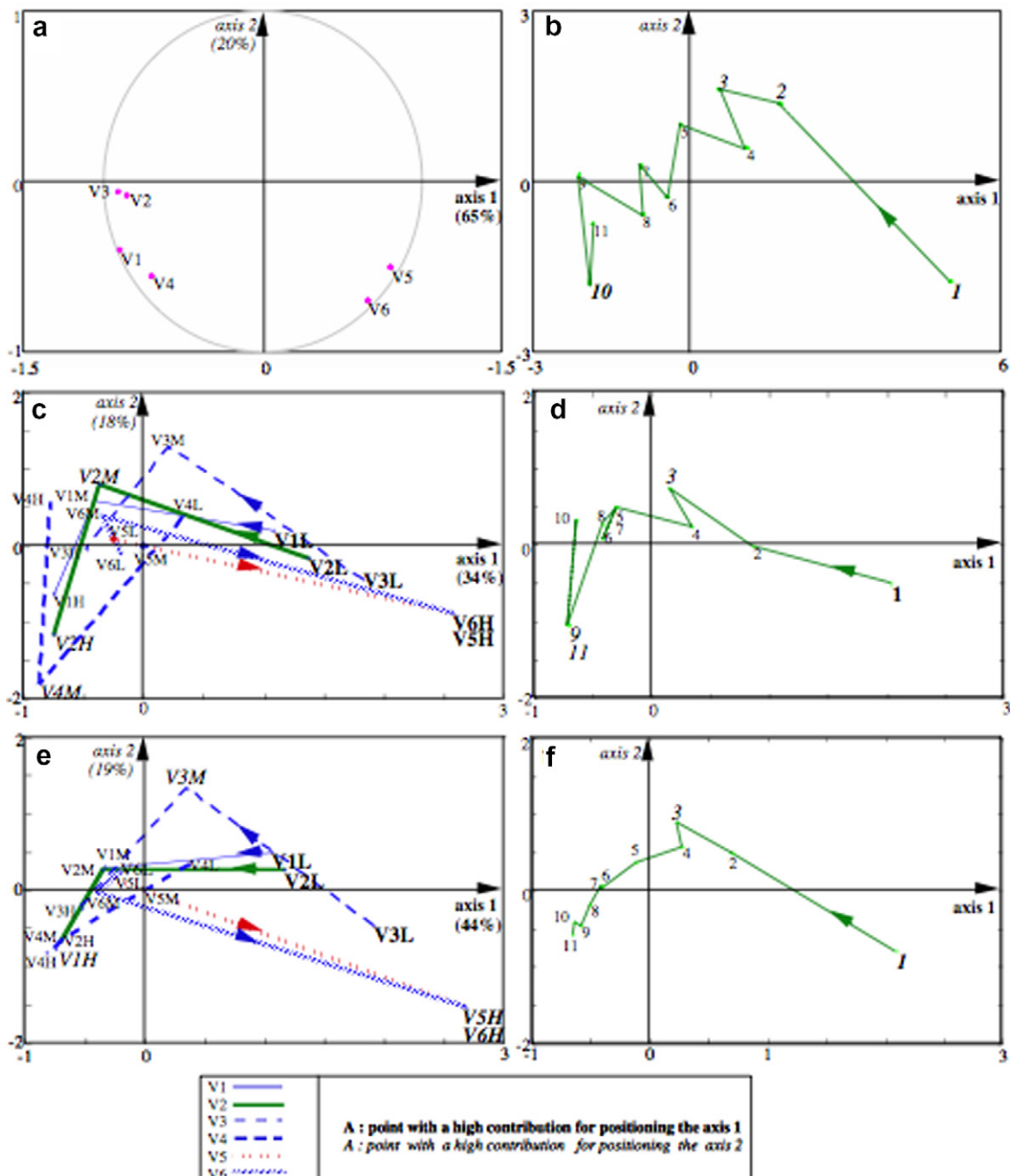


Fig. 4. Main plane crossing axes 1 and 2. (a) and (b) projections of the 6 variables and the 11 time observations for the PCA, (c) and (d) projections of the 6×3 space windows and the 11 time observations for the MCAc, (e) and (f) idem with MCAf.

mechanics, the closer a point is to the gravity center (the point where the main axes cross), the higher its mass (here, its membership value). For instance, the point $V5H$ is far from the gravity center because the corresponding space window occurs rarely (see Table 1b). A point situated just on the gravity center means that the corresponding window never occurs (as with $V5M$).

The first axis is mainly controlled by the *High* window of the variables $V5$ and $V6$ and the *Low* window of the variables $V1$, $V2$ and $V3$. These 5 windows are situated on the positive side of axis 1, which means that if a time sample n is situated on the positive side of axis 1 and is well connected to axis 1, this time sample n should have rather high values for variables $V5$ and $V6$ and rather low values for $V1$, $V2$ and $V3$. The same interpretation can be used to explain the next axes, except that the

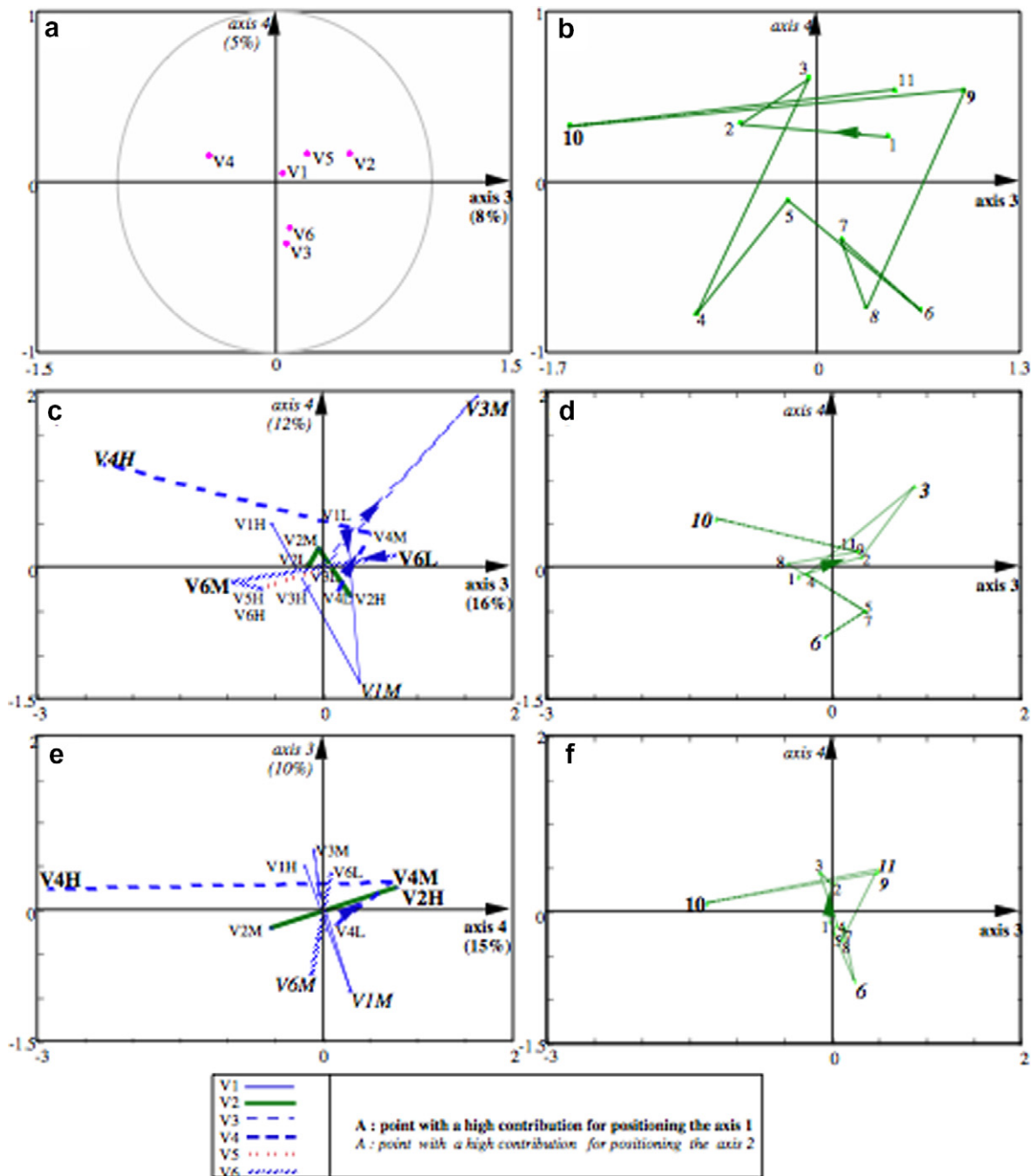


Fig. 5. Main plane crossing axes 3 and 4. (a) and (b) projections of the 6 variables and the 11 time observations for the PCA, (c) and (d) projections of the 6×3 space windows and the 11 time observations for the MCAf, (e) and (f) idem with MCAf.

statistical phenomena are more and more local (in space and in time). For instance, axis 3 shows that the 10th time observation occurs with a high value for V4 (Fig. 5b and c).

3.1.3. Comparison of PCA, MCAC (with crisp windowing) and MCAf (with fuzzy windowing)

3.1.3.1. Comparison between PCA and MCAC. In PCA, the points corresponding to V2 and V3 are very close and both situated near the circle, which should mean a strong linear correlation. The scatterplot of $V2 \times V3$ (not shown) confirms this fact, and the linear correlation coefficient is rather high ($lcc(V2, V3) = 0.75$). In the same way, V5 and V6 should be well linked linearly.

The scatterplot $V5 \times V6$ exhibits a situation of independence (here, $lcc(V5, V6) = 0.82$ is large due to an outlier). This underscores the limitation of PCA.

In MCAC, each variable is not represented with a single point but with $S = 3$ points (due to the presence of 3 space windows). The graphic output is much more complex than the PCA output. Though the first plane (crossing axes 1 and 2) often exhibits trajectory patterns that form a "V" or a "Λ", the next planes can exhibit miscellaneous patterns. This complexity goes with the complexity of the data, and thus MCAC shows more local relationship phenomena than PCA does. For instance, the two points corresponding to $V2$ and $V3$ are very close for the first plane of the PCA (Fig. 4a) but the corresponding space trajectories are rather different for the first plane of the MCA (Fig. 4c). The same remark can be made for $V5$ and $V6$: $V5H$ and $V6H$ have the same position due to the outlier (i.e., the first time observation), but $V5M$ and $V6M$ have quite different positions, with the null coordinates of $V5M$ due to the null occurrence of this space window. Again, MCAC provides more specifics than does PCA. One drawback to MCA often cited in the literature is that the inertia percentages of the main axes are much lower than the variance percentages of the main axes of the PCA. However, this is totally logical since the crisp coding yields an artificial orthogonality between the columns of the table (see [5,12] for other measures of the information shown with a main axis).

3.1.3.2. Comparison between MCAC and MCAf. The inertia percentage of the first axis of MCAf is much larger than the corresponding value of MCAC. The space trajectories and the time trajectory with MCAf are more smooth than the corresponding trajectories with MCAC. Some space trajectories are closer (e.g., space trajectories of $V1$ and $V2$), while some trajectories are farther away (e.g., those of $V1$ and $V3$) (Fig. 4c and e). Finally, the first plane exhibits 10 time observations with different coordinates (Fig. 4f), while points 9 and 11 have identical coordinates with MCAC (Fig. 4d). All these differences are due to the fuzzy space windowing, which lessens the information loss when transforming a quantitative scale into a qualitative scale.

Given the advantages of space windowing over using the initial scale, let us continue more generally with such a scale transformation.

3.2. Some aspects of space windowing

Consider a set of $V0$ measurement variables, $\{Y0_v(t) (v = 1, \dots, V0)\}$, and suppose that there are only two factors, in addition to the intrinsic individual factor. Thus, the generic value is $y0_{ijkn,v}$, where i, j and k represent the modalities of the 3 factors and n represents the time sample. The main idea is to consider the variables using verbal nuances – for example, *low*, *medium*, *high* for a position or a speed; *towards the left*, *towards the upper right* for a direction or a trend; and *accurate*, *good*, *bad* for a behavior or a choice – while keeping in mind the possibility of using fuzzy membership functions and combining two or more measurement variables (Fig. 6).

When the space windowing is performed variable by variable (Fig. 6a), the following transformation takes place:

$$Y0_v \mapsto Y1_v \quad \text{with} \quad \{y0_{ijkn,v}\} \rightarrow \{y1_{ijkns,v} = \mu_s(y0_{ijkn,v}), s = 1, \dots, S1_v\} \quad (1)$$

where $S1_v$ is the number of space windows; the arrow " \mapsto " denotes the new variable $Y1_v$, which is associated to the measurement variable $Y0_v$, thus changing the scale from quantitative to qualitative; the arrow " \rightarrow " denotes the computation of a new set of values from an initial generic value ($y0_{ijkn,v}$), in which the tag "1" indicates a data analysis step with "1" corresponding to data characterization; and the μ_s denotes the membership value of the space window s .

When the space windowing combines several measurements variables (Fig. 6b), the transformation becomes:

$$(Y0_v, Y0_{v'}) \mapsto Y1_{v''} \quad \text{with} \quad \{y0_{ijkn,v}, y0_{ijkn,v'}\} \rightarrow \{y1_{ijkns,v''} = \mu_s(y0_{ijkn,v''}), s = 1, \dots, S1_{v''}\} \quad (2)$$

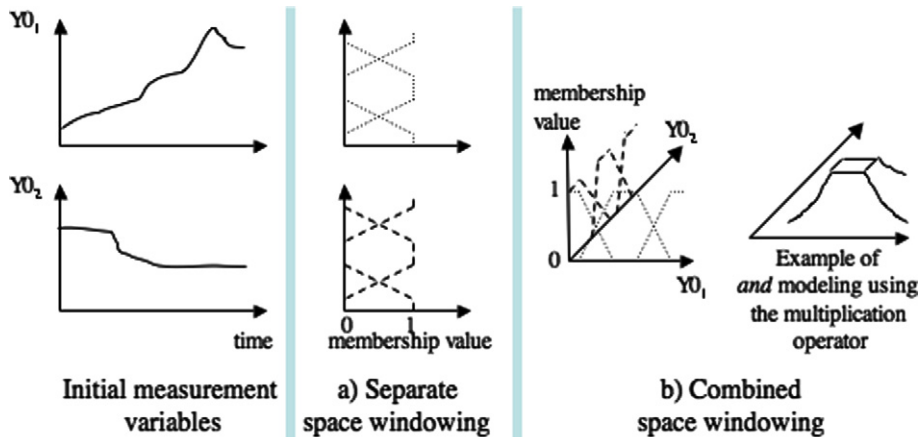


Fig. 6. Example of space windowing with measurement variables considered: (a) one by one and (b) in 2-variable combinations.

where $S_{v''}$ is the number of space windows of the new variable v'' . In most cases, two variables, such as a position and a speed or a horizontal and a vertical position, are combined. In the specific bivariate case shown in Fig. 6b, $S_{1_{v''}} = S_{1_v} * S_{1_{v'}}$. However, as will be shown later with the oculometric data examined in Section 4, many other cases may also be considered.

From the membership values thus generated, statistical summaries can be computed. For instance, with membership values obtained from (1), the classic overall weighted arithmetic mean can be computed for each space window. Thus, for each cell (i, j, k) and each measurement variable v , the set of $NO_{ijk,v} * S_{1_v}$ membership values can be replaced with a set of S_{1_v} membership value averages (MVA) as follows:

$$\left\{ \begin{array}{l} y1_{ijkns,v}, \quad n = 1, \dots, NO_{ijk,v} \\ s = 1, \dots, S_{1_v} \end{array} \right\} \rightarrow \left\{ \begin{array}{l} y2_{ijk,s,v} = \frac{1}{\sum_{n=1}^{NO_{ijk,v}} w1_{ijkn,v}} \sum_{n=1}^{NO_{ijk,v}} w1_{ijkn,v} \times y1_{ijkns,v} \\ s = 1, \dots, S_{1_v} \end{array} \right\} \quad (3)$$

where w denotes the weight value. The weight values may or may not be identical within the same cell (i, j, k) . Moreover, MVA can be computed more locally in time (T time windows are considered, again with either crisp or fuzzy functions). Designing membership functions is a rather complex process, mainly depending on the data and on the objectives of the data analysis. The next two sections consider two real examples of study data. Only the elements necessary for understanding the space-time windowing will be presented. Specifically, this means that no details about the experimental conditions, the measurement device or the pre-processing stages will be discussed.

4. Example with one category variable: area of interest in an eye scan of an advertisement

The aim of the study was to analyze how individuals look at advertising images. A set of 30 pictures was taken from magazines and transformed into slides, which then were shown to 47 individuals using a slide projector. The 30 slides were projected continuously, each presentation lasting 5 s, in the same order for all participants.

4.1. Data acquisition

Data was acquired using an eye tracker [8], which, thanks to a specific calibrating procedure, provides the gaze position in terms of the surrounding environment. The sampling frequency was 60 Hz, with about a 1 visual degree of measurement uncertainty. The parallelepipedic structure shown in Fig. 1 was used to organize the data, but with only two factors, the first linked to the individual i ($i = 1, \dots, I = 47$) and the second linked to the advertising image j ($j = 1, \dots, J = 30$). With $VO = 2$ measurement variables, representing the horizontal (x) and vertical (y) positions, each of the $I * J$ cells contains a multidimensional signal with $VO = 2$ components.

In the following, the point $(x = 0, y = 0)$ corresponds to the bottom-left corner of the slide and the point $(100, 100)$ to the top-right corner. Given the 1° of measurement uncertainty mentioned above, the size of the slide, and the distance between the slide and the individual, the eye position uncertainty is about 2% for both axes. The eye tracker produced a rather complex set of values for these x and y components: most of the values fell between 0 and 100; some were less than zero or more than 100 (e.g., -8% or 120%, indicating that the gaze position was outside the image rectangle); and a very small number were the result of a coding error (which means that, for varied reasons, the eye tracker could not find the eye position). Each cell (i, j) contains $NO_{ij,v} = NO = 60 \text{ Hz} * 5 \text{ s} = 300$ time samples, with more than 90% of them being exploitable.

Since the primary aim of this paper is to show how complex time data can be investigated, only one image will be considered here (i.e., the subscript j will be omitted). Thus, the generic value is $y0_{in,v}$ ($i = 1, \dots, I = 47$; $n = 1, \dots, NO = 300$; $v = 1, \dots, VO = 2$).

4.2. Data analysis

In order to perform the analysis, some eye movement characteristics must be understood. First, images that humans are aware of perceiving are stable and continuous. In fact, the eye in its socket makes rapid irregular movements, or saccades, from 2 to 5 times/s; these saccades have a non-negligible magnitude (between 4° and 15°), and a high speed (up to 600°/s). As Rayner [20] stated, “we do not obtain new information during a saccade”, which should not be taken to mean that no cognitive processing takes place during a saccade. Between two saccades, several kinds of movement can occur, mostly “fixations”, which is actually a misnomer since the eye is never really still. (There is a small constant tremor of the eye, called *micro-nystagmus*. This term is used with either normal or pathological movements. See reference [14] for further details.)

Visual perception is performed through two complementary channels: central vision (materialized by a cone of about 3°) and peripheral vision. The first channel permits detailed visual perception; the second channel is needed to comprehend the global environment. The duality between central vision/peripheral vision means that, in most cases, the individual's attention corresponds to his/her gaze direction. But in some situations (e.g., car driving), the attention is shared between the central vision and the peripheral vision. When ocular activity is studied, only the gaze direction is recorded (i.e., the central visual field). For the reasons evoked above, this measurement can easily lead to interpretation errors, and thus it is better to corroborate any inference result with other measurements.

Now that the principal eye movement characteristics have been explained, let us move on to analyze the data, using the five steps presented in detail in Section 2.2.

The characterization method encountered in the eye movement literature, specifically in ergonomics and psychology, is usually fixation characterization, mainly involving determining the duration of the fixation (e.g., arithmetic mean of the fixation durations is computed for a given individual). The duration indicator is considered to provide information about the mental workload. In addition, this indicator is fairly easy to compute and test using a statistical inference procedure, which is not the case for the position indicator because it is often based on a qualitative scale. The next most popular method attempts to characterize the saccade movement, mainly by examining its magnitude (The interested reader can consult reference [14] for other types of eye movements).

Careful examination of x and y signals recorded for the image allows a fixation to be considered as a class of a bidimensional signal, this class lasting more than 100 ms [14] and containing close points. In our case, this signal visual analysis led us to choose points that are separated by 8% max (NB: the size of the image is 0–100 by 0–100). Once the fixations have been found, a distinction that can be made about the *space entity*, which is either a geometric or a semantic zone. The former is a geometric area without reference to its content (e.g., the central zone [40,60] for x and [40,60] for y). The latter refers to the elements of the advertising image: the product (e.g., car, dishwashing liquid, perfume), the individual (e.g., a beautiful woman), the slogan or text providing the product characteristics, or the brand logo. For the following analysis, let us consider the example given in Fig. 7.

The next stage should be the scale transformation. Since the characterization method is based on space windowing, all the data are homogeneous (i.e., membership values between 0 and 1), making it unnecessary to transform the data scale. Let us now consider how to build the data tables.

For a multidimensional descriptive analysis, the data obtained during this study can be used to build a series of two-entry tables. Let us suppose that the idea of “fixation/semantic zone” pair is retained and that a comparison of the individual behaviors is the goal of the analysis. For length considerations, only two cases are considered here. The main aim of the statistical analysis is the comparison of the individual profiles, thus the chronological aspect of time is not introduced in the analysis (the problem is to know which parts of the ad are seen, regardless of when). In this case, $I = 47$ individuals must be compared globally. For each of them, the 5 MVA are placed side by side, yielding, for each individual, a table with $R = 47$ rows and $C = 5$ columns, denoted $T_{47 \times 5}$.

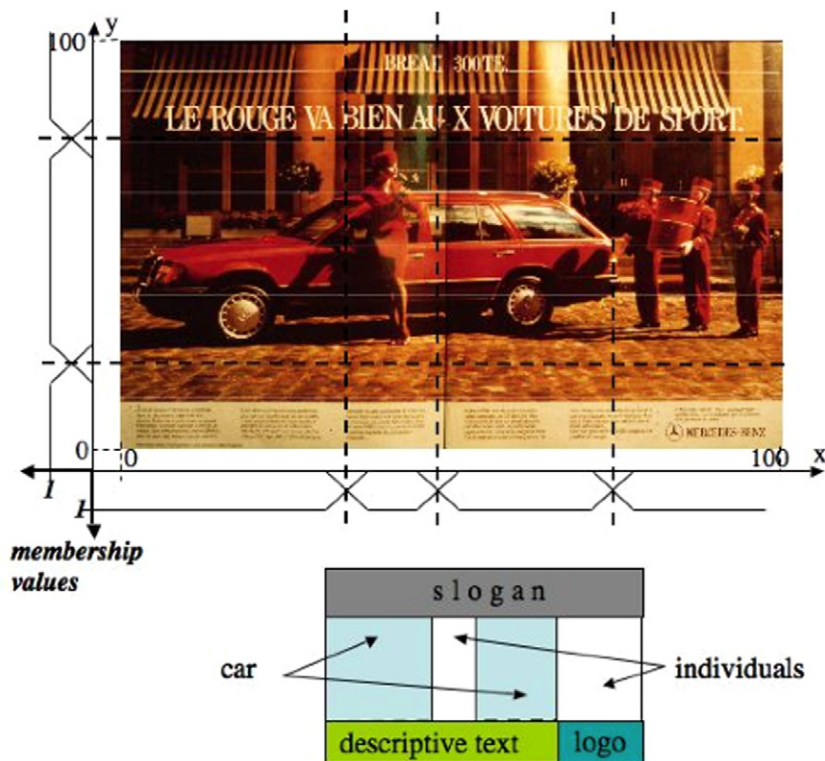


Fig. 7. Example of an advertising image and fuzzy windowing with $S1 = 5$ semantic zones corresponding to the car, the individuals, the slogan (the line in large characters situated at the top), the descriptive text and the logo (at the bottom-right corner). (Each of the 5 semantic zones is obtained by combining the $5 \times 3 = 15$ trapezoidal zones).

Time windowing (thus chronology aspect of time) can now be introduced into the analysis, for instance with $T = 5$ windows of 1 s. To do sequential analysis, each set of 5 MVA corresponding to each individual time window can be placed in one row, yielding a table with $R = 47 \times 5 = 235$ rows and $C = 5$ columns, encompassing all 47 individuals. This table is denoted $\mathbf{T}_{235 \times 5}$. Since there are three subscripts – i , t and s – other tables can obviously be built. Let us consider a table whose membership values are averaged over the time factor, yielding table $\mathbf{T}_{147 \times 5}$ in which each row r corresponds to an individual i , and another table whose membership values are averaged over the individual factor, yielding table $\mathbf{T}_{5 \times 5}$ in which each row r corresponds to a 1 s time window t .

These tables can be studied using multidimensional exploratory methods, either based on the classification principle (e.g., either rows or columns are grouped) or the factor analysis principle (spectral decomposition)[12]. Here again, to keep this paper at a reasonable length, only the factor analysis principle will be considered. Since the tables contain membership values, multiple correspondence analysis (MCA) can be used [5]. Again, several analysis possibilities exist; however, we chose to use MCA to examine only the case with one-second time windows.

MCA of $\mathbf{T}_{235 \times 5}$ yields two principal axes with relative inertia values of 37% and 31%, respectively; these values are much larger than the values for the secondary axes, which are 18% and 15% respectively. Thus only the first main plane will be considered here. The principal graphic output is shown in Fig. 8. The set of 235 row points (Fig. 8a) has a very particular pattern: most points (subset A) are situated on a line running from the point $(-0.5, -1)$ to the point $(1, 0.5)$, while the other points (subset B with $B \ll A$) are situated above and to the left of this line. This is because only three main zones – individual, car and slogan – were observed (see Fig. 8b). The points in subset B mainly correspond to the time windows 4 or 5 (i.e., during the 4th or 5th second), with the corresponding individuals tending to fixate on the logo or the text. This is notably the case for individual 39. The point labeled 39-5 is situated at the top-left corner of Fig. 8a. This specific behavior during the 5th second is shown more quantitatively in Fig. 9a.

In order to highlight the influence of both the individual and the time factors, the 47×5 row points of tables $\mathbf{T}_{147 \times 5}$ and $\mathbf{T}_{5 \times 5}$ were projected as supplementary points [5,12] onto the planes obtained from the CA of $\mathbf{T}_{235 \times 5}$ (Fig. 8c and d, respectively). Fig. 8c shows that the differences between the 47 overall individual profiles are rather large. These differences are

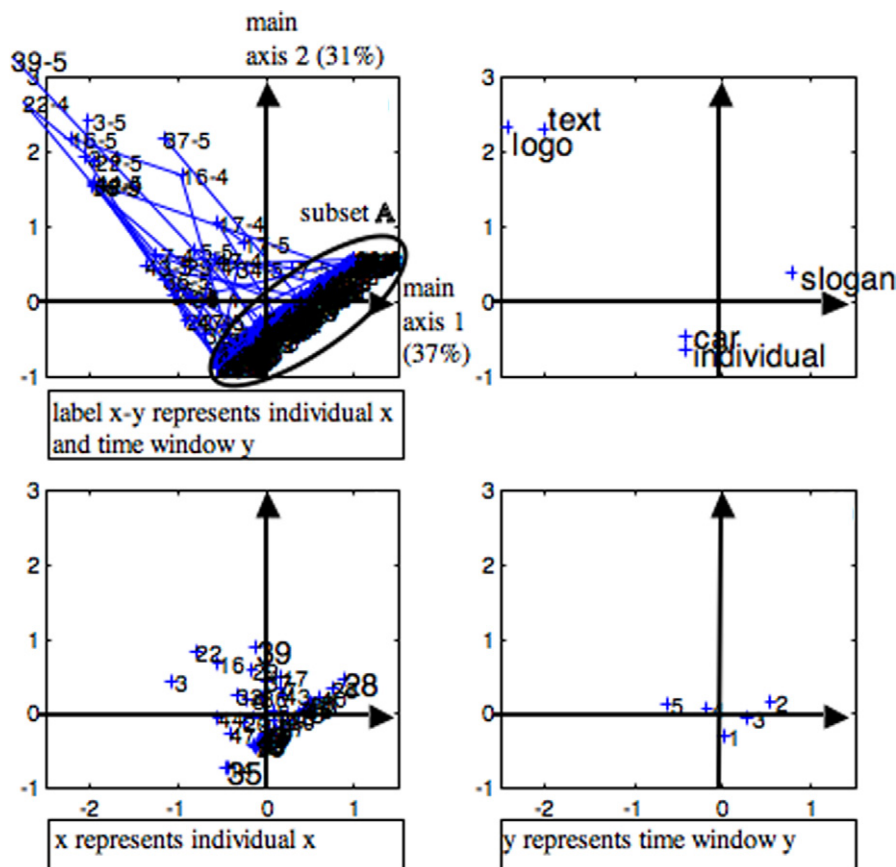


Fig. 8. Graphic MCA output for table $\mathbf{T}_{235 \times 5}$ with axes 1 and 2: (a) projection of the $47 \times 5 = 235$ row points corresponding to the 47 individual time trajectories (with 5 time windows for each trajectory), (b) projection of the 5 column points corresponding to the 5 semantic zones, (c) projection of the 47 supplementary row points corresponding to 47 individual average profiles (rows of table $\mathbf{T}_{147 \times 5}$), and (d) projection of the 5 supplementary row points corresponding to 5 time window average profiles (rows of table $\mathbf{T}_{5 \times 5}$).

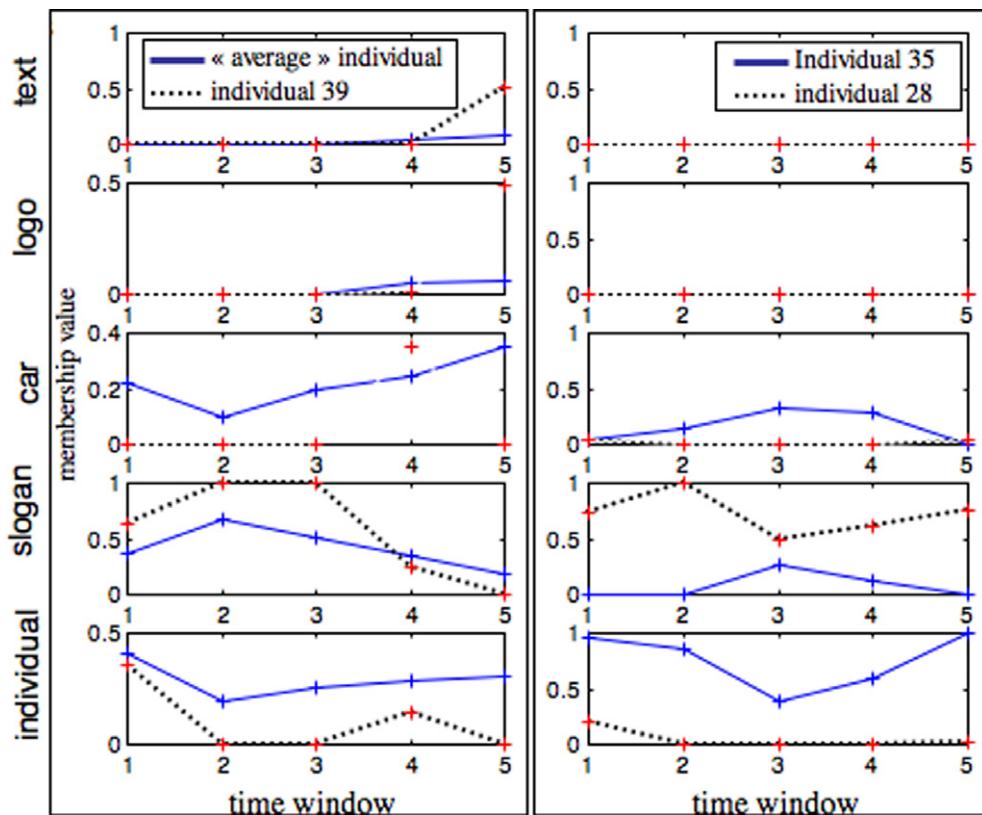


Fig. 9. Examples of time evolutions: (a) a comparison of the profiles of the “average individual” (the data are contained in $T_{3_{5 \times 5}}$) and individual 35 (who appears to behave differently from the average during the 5th time window, as shown in the left-top corner of Fig. 8a and b, a comparison of the profiles of individual #28 and individual #35 (who appear to behave quite differently according to Fig. 8c).

presented more quantitatively in Fig. 9. Fig. 9b presents two individuals: individual #35, situated on the bottom-right part of subset \mathbb{A} , and individual #28, situated on the top-left part of subset \mathbb{B} . As Fig. 8d shows, the 5 points corresponding to the 5 one-second time windows are not so far from one another, meaning that there are several time evolutions. Nevertheless, in general, the text or the logo tends to be read after the slogan or scanning the images of the individuals or the car.

To show the influence of the time factor more generally, the evolution over time of the “average individual” can be studied by performing MCA on $T_{3_{5 \times 5}}$ instead of on $T_{2_{35 \times 5}}$. The first main plane (not shown) is consistent with the average profile shown in Fig. 8a.

The primary aim of presenting this eye movement example was to show the *multiplicity* of data analysis paths that can result from a MFMV database containing complex time data. Only two of the possible paths have been presented here, and this rather quickly: a path yielding the result set \mathbb{E}_{51} , obtained from the MCA of table $T_{2_{235 \times 5}}$, with the rows of $T_{1_{47 \times 5}}$ and $T_{3_{5 \times 5}}$ projected as supplementary points, and a path yielding the result set \mathbb{E}_{52} , obtained from the MCA of table $T_{3_{5 \times 5}}$. The two sets of results from these two paths mainly generate graphic illustrations showing the influence of both the individual factor and the time factor on the eye scanning behavior, as shown in Figs. 8 and 9. Some of the other possible paths are summarized in Fig. 10.

Our second reason for presenting this eye movement example was to introduce the notion of *information*, for which one meaning could be “what result can be obtained about eye scan paths using one or another data analysis path”. The notions of *multiplicity* and *information* are discussed in more detail below in order to justify our choice of analysis path both generally and in the specific case of eye gaze data. Let us now consider the case in which there are several quantitative variables in the analysis.

5. Example with several quantitative variables: car and head movements in driving

The aim of the study was to investigate highway driving vigilance using a simulator. A set of five elementary sections of 2 km length was designed, each section being rather “boring” (no other cars on the road and a monotonous landscape) with one of the following profiles: straight line (labeled *s*), right and left turns with a *low* road curvature (5 km radius, labels *r* and *l*), right and left turns with a *high* road curvature (2 km radius, with labels *R* or *L*). A standard highway module was composed

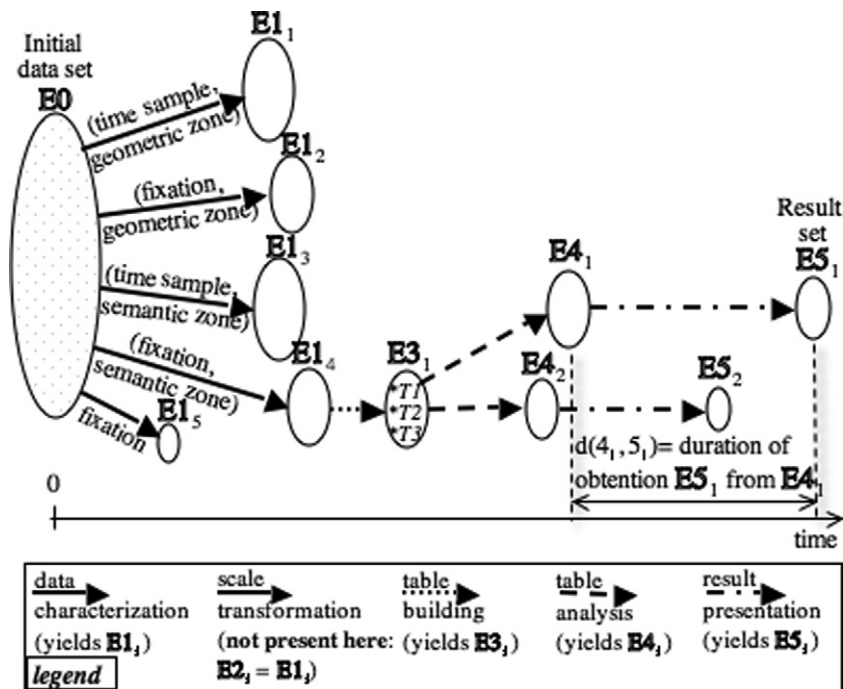


Fig. 10. Examples of possible data analysis paths (the size of the oval reflects the size of a set, and the relative position of the oval on the time axis reflects the duration of a set acquisition).

of 12 sections so that the chronology would seem to be random (i.e., a sequence such as *rsRlsslsrsLs*. yielded a 24 km lap). The minimum period duration without a rest was 3 laps (i.e., $24 \times 3 = 72$ km), plus a 7 km section to start the driving task and a 5 km section to end it (i.e., $72 + 12 = 84$ km total). Movements were recorded only during the 72 km. Four driving periods were considered with a 5 min. rest between two periods (i.e., $4 \times 84 = 336$ km total). The instructions given to the 32 subjects were to drive as they normally did, paying attention to the highway code.

5.1. Data acquisition

Data was acquired using the simulator computers and specific measurement devices, yielding a total of 11 variables: 5 variables related to the car (speed, lateral position on the road, steering wheel angle, steering wheel torque and the relative position of the accelerator pedal) and 6 variables related to the driver's head (3 linear and 3 angular positions). The first 11 derivatives were computed, yielding $V0 = 22$ variables globally. Because the device measuring the head position probably changed position during the rest between 2 recording periods, the 6 head positions were centered period by period using the arithmetic mean.

The data structure $H0$ was organized with the individual i ($i = 1, \dots, I = 32$), the driving period j ($j = 1, \dots, J = 4$), the lap k ($k = 1, \dots, K = 3$) and the section l ($l = 1, \dots, L = 12$). Thus the structure $H0$ contains $H0 = I * J * K * L = 4608$ hyperparallelepipedic cells.

5.2. Data analysis

The characterization methods encountered in the car driving literature, specifically in ergonomics and psychology, are usually based on global indicators such as the arithmetic mean or the standard deviation. These indicators are often investigated one by one using a monivariate analysis method, specifically an inference method.

Considering the remarks in Section 3, space windowing seems the best choice, followed by a multivariate analysis method. Thus 22 magnitude histograms must be carefully analyzed. Each histogram is built from all the 6 Hz time data of the $H0$ signals, which yields about 1.5×10^6 values (this number is consistent with an the average driving speed situated about 130 Km/h). There are two main histogram patterns:

- dissymmetric, with one or more modal areas for such variables as driving speed,
- (quasi) symmetric, resembling the Laplace–Gauss model for such variables as road positions.

It is worth noting the presence of very low or very high magnitude values, as compared to the arithmetic mean. Such values occur very rarely, probably due to measurement perturbations. Thus a windowing only based on a space criterion will be

greatly influenced by these extreme and rare values. Furthermore, with such a criterion, the weight of the space windows may be very different in MCA. For both reasons, a space windowing yielding more or less identical membership values averages is preferred. Since this is an exploratory context with a rather large number of time variables ($V0=22$), the space windowing only considers 3 windows (i.e., *low*, *medium*, *high*).

Given these choices, the symmetric aspect of the histogram must be taken into account. For instance, if the zero value corresponds to an absence of physical phenomenon (e.g., a null angle means there is no rotation), the medium space window must be situated just around this null value. In the other cases of symmetric histograms, the medium space window can be adjusted around the median, which is less sensitive to outliers than the arithmetic mean. Fig. 11 provides some examples of

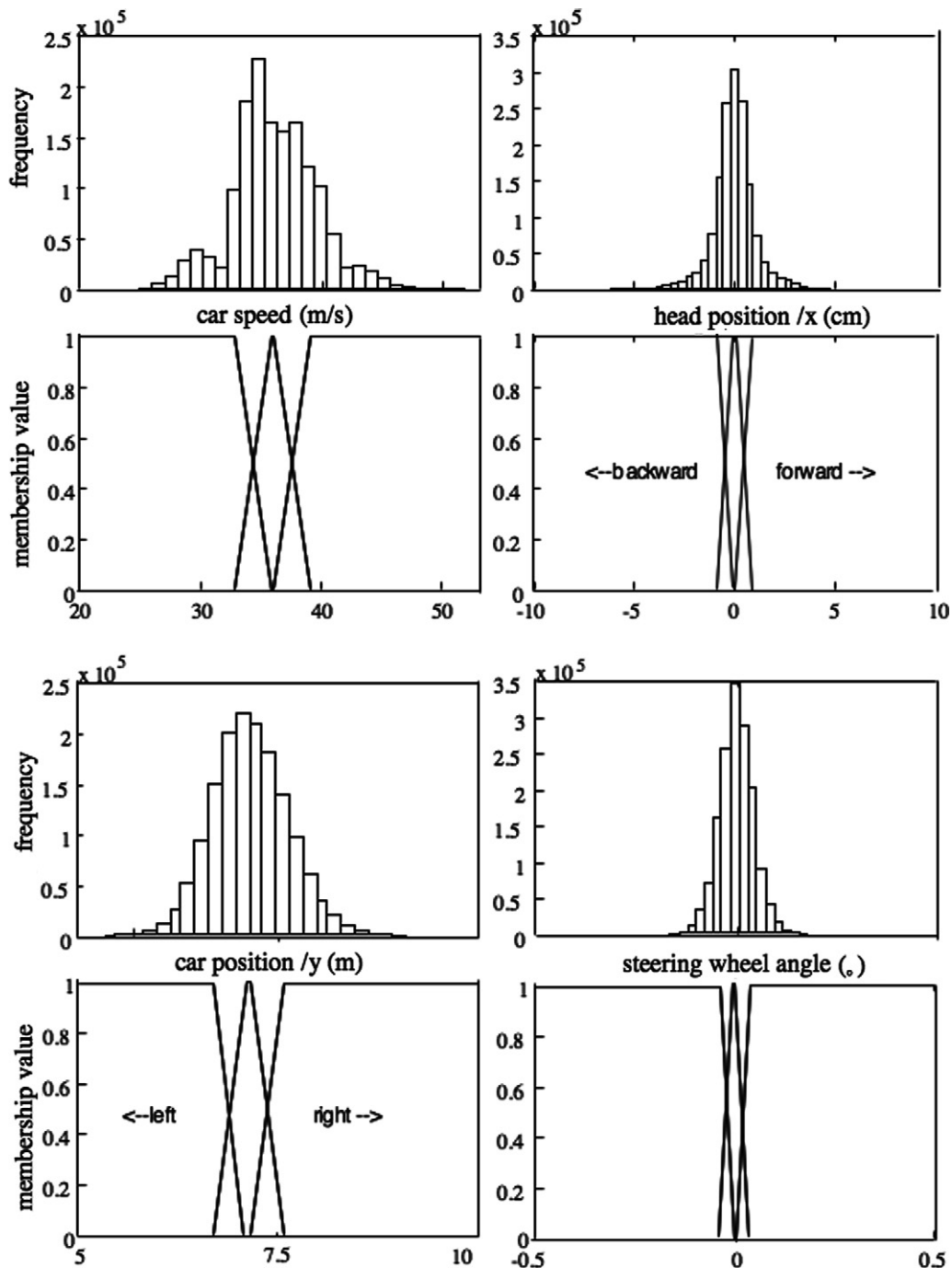


Fig. 11. Some examples of magnitude histogram and scale windowing in the car driving study.

histograms and the corresponding space windowing models. Please note the coherence between the histogram and the data. For instance, the histogram of the car's lateral position on the road presents a Laplace–Gauss pattern, meaning that the individual drives with quasi-stationary central trend and symmetric fluctuations around this trend. The nominal position can be characterized by the median indicator, and its position value corresponds to the maximum of the 'medium' function membership (about 7.4 m, this position being given in relation to the reference axes). The symmetric aspect can be explained by the same number of right and left turns (with a symmetric distribution for the curvatures).

The presence of a Laplace–Gauss model is less pronounced for the car speed because there are slow and fast drivers (the speed is not much influenced by the road curvature since the road only bends slightly). The membership function figure shows that the median is about 35 m/s (129.5 km/h), which is consistent with the maximal authorized speed in France (130 Km/h on highways).

Once the membership values have been obtained for all the time samples of all the 22 variables, given our MFMV context, many tables can be considered for an exploratory analysis. Since the aim of this study is driving vigilance, showing the influence of the section profile (factor I) is secondary. On the other hand, showing the time influence is important. Thus a table $T1_{128 \times 66}$ whose the rows correspond to the $I \times j = 128$ individual driving periods is built. Several global tables can be obtained, for instance by averaging across periods (table $T2_{32 \times 66}$) or across individuals ($T3_{4 \times 66}$). In addition, due to the presence of a large number of space windows, several MCA may be necessary if the first and/or second axes are built on the strong correspondences between very few space windows.

In the case considered here, the first two axes from $T1_{128 \times 66}$ are due to a high connection between the speed and the accelerator pedal position, which sounds logical. Thus one of these two variables must be removed. Several MCA were considered, including the one presented Fig. 12. This MCA highlights a strong correspondence between the 6 head speeds, opposing the medium window (the speed is close to zero) to the two other windows (see axis 1, Fig. 12). Since the 6 trajectories are close together, to keep the graphic simple, only the one with the highest contribution to position axis 1 is shown (i.e., vertical linear speed). This first axis also exhibits the time influence, especially through the 4 points corresponding to the 4 periods considered as supplementary points. This influence is shown more quantitatively in Fig. 13, which shows the time evolution of the standard deviation (sd) of the vertical head speed. Of course, it is important to remember that an sd value computed from a signal is not the same as 3 MVA computed from the signal derivative.

As with the example with eye movement data, the primary aim of presenting the car driving example was to show the *multiplicity* of data analysis paths that can result from a database containing complex time data. Given this commonality with the first example, the main difference is that, in the first example, there is only one qualitative variable, while in the second, there are several quantitative variables.

As with the first example, the second reason for presenting this example was to introduce the notion of *information* from the perspective of space windowing, since very few data analysts spend their time dividing the variables up when these variables can be considered as they are: quantitative. Let us now consider the *multiplicity* and *information* aspects of the analysis.

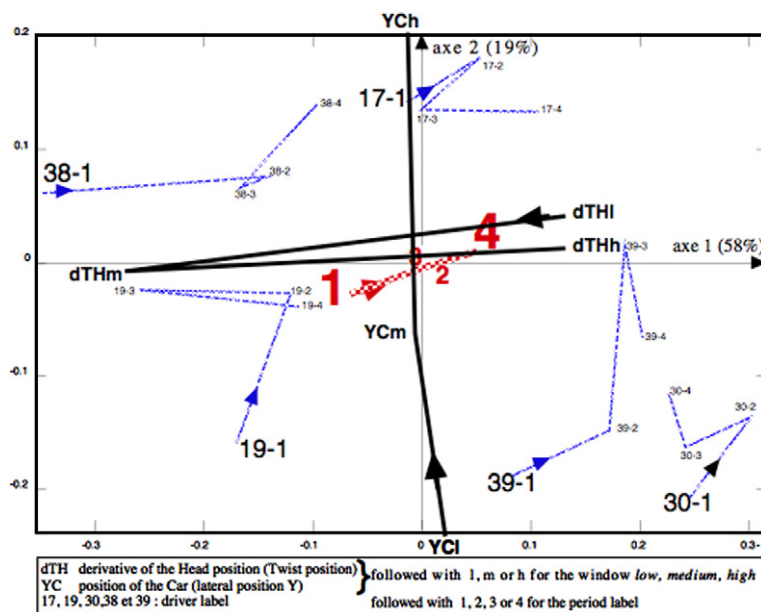


Fig. 12. Graphic MCA output for table $T4_{128 \times 36}$ with axes 1 and 2 showing (a) column points of the space windows with the highest contributions, (b) some row points of specific individual time evolutions and (c) 4 supplementary row points corresponding to 4 time window average profiles.

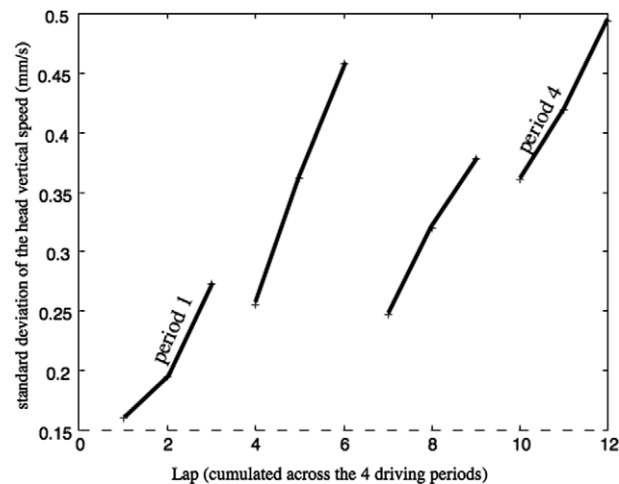


Fig. 13. Evolution of the standard deviation of the vertical head speed throughout the laps.

6. Discussion

Despite the fact that the two examples considered above are completely different, there is an identical triplet of graphic elements:

- (1) A space windowing (respectively, Figs. 7 and 11).
- (2) A rather complex factor plane from a multidimensional descriptive method (i.e., MCA) (Figs. 8 and 12).
- (3) A simpler plot showing the influence of one factor (i.e., time) on a variable (a membership value for Fig. 9 and a standard deviation for Fig. 13), in which the influence is highlighted by both the space windowing and the factor plane.

The notions of *multiplicity* and *information* are present in each of these three basic elements but, as explained in § 2.2, since the data characterization is the first sub-step of an MFMV analysis, it mainly influences the rest of the analysis. For this reason, we have focused here on the space windowing (element 1). Both the multiplicity and information notions are considered first from a general perspective and then in terms of our examples.

6.1. Multiplicity

In HCS studies, the *multiplicity* in the analysis of a MFMV database generated from an experimental or observational design is first due to the multiple ways that time data can be characterized. Even considering only quantitative measurement variables, there are many indicators used in the fields of statistics and signal analysis, as well as many related fields (e.g., econometrics or epidemiology) which deal with time data. In most of these fields, quantitative scale mathematical models (e.g., arithmetic mean, standard deviation, extreme values, energy value, moving average) are used almost exclusively.

For a very first MFMV analysis, we propose using space windowing instead of maintaining the quantitative scale. Obviously, doing so will increase the number of possibilities for the analysis (i.e., its multiplicity). For instance, transforming the initial quantitative scale model into a qualitative model requires that two main alternatives for the number of space windows S be kept in mind: a high value or a low value:

- A large value can be chosen when designing a magnitude histogram (e.g., $S = 20$ intervals or more), subsequently allowing modal areas to be found. However, a large value should be avoided when there are many variables to be studied together (e.g., when $V \gg 5$).
- A low value can be chosen when there are many variables to be studied together. For instance, when $V > 10$, choosing $S = 3$ windows (e.g., *low*, *medium* and *high*) allows main relationship trends to be identified between the variables, with sufficient occurrences to fall within a combination of V space windows. In some cases, particularly when qualitative variables are also present, $S = 2$ windows can be chosen, such as *medium* and *extreme* (in the sense of *low* or *high*).

Given that space windowing increases the *multiplicity* of the analysis, why would we choose to use this method? Our response to this question also responds to the various criticisms that have been made when we have proposed this scale model change in applied fields, such as medicine, ergonomics or psychology. The primary criticisms reflect a lack of understanding of the advantages of the technique (e.g., Why impoverish the scale, given that the measurement device, often expensive,

already provides a quantitative scale?) and concerns about the subjectivity of the windowing choices (e.g., How many windows should be used and which membership function pattern should be chosen for each window?).

In general, the primary advantage of using the windowing technique for the very first analysis lies in the fact that a set of S membership value averages (MVA) identifies the presence of outliers more readily than does the arithmetic mean (or a combination of the arithmetic mean and the standard deviation). This is even more true if the space windowing is done following a careful analysis of the magnitude histogram, which is not done as often when classic indicators are used. The second advantage is that, despite the impoverishment of the scale, individual behaviors are described more accurately than with classic indicators. For instance, the arithmetic mean is only able to demonstrate a general tendency, whereas just $S = 3$ MVA allows the intensity of the space window's influence to be clearly indicated, for example, by showing how often the signal derivative is *positive, close to zero or negative*.

Certainly, analyzing a large MFMV database – like the one in Fig. 1 but with more than three factors and a large multi-dimensional signal within each cell – is much more complicated. Analyzing a set of MVA triplets (or more MVA) takes more time and also seems to be much more subjective than analyzing a set of arithmetic means (or a set of arithmetic means and a set of standard deviations). However, there is often hidden subjectivity with classic indicators (e.g., Why choose the arithmetic mean? Why not the RMS value?). In reality, the subjectivity linked to windowing is partly due to the large number of space windowing possibilities.

To explore the reality of this subjectivity, let us first consider our example concerning eye movement data. The fixations can be assigned to bivariate space windows stemming from geometrical considerations (e.g., the ad is cut into rectangular zones with identical surfaces) or semantic considerations (e.g., the components in the advertising image). Though for evaluating the ad (and maybe comparing several ads), the second approach appears to be better, the advantage of the first approach is the possibility to highlight more global phenomena, such as that the eye scanning mainly starts at the top-left corner and ends at the bottom-right corner. In fact, in the exploratory context, the best way would be to test the two windowing models.

Let us now consider our driving data example, in which many more windowing models can be used. Two main taxonomic dimensions can be considered (i.e., identical width vs. identical MVA windows and few vs. many windows). Although the advantages and disadvantages of each of the two choices may seem obvious, we believe that the space windowing depends first on the type of data and the analysis objectives. For example:

- (1) The space windowing must be performed using the physics of the data. When the scale has an absolute zero (0 m/s^2 for the car speed means there is no acceleration, 0° for the steering wheel angle means the car does not turn), the space window must be built around this zero value.
- (2) The space windowing must be based on the histogram pattern. When the distribution is symmetrical, the space windowing must also be symmetrical, with a window around the mode(s). If the distribution should be symmetrical but is not, due to measurement errors, the windowing must be symmetrical anyway. This is the case when there is a physical zero within the range $[\min, \max]$, but due to the presence of outliers in the left side of the histogram, $|\min| \ll |\max|$. A windowing with identical width windows is obviously more sensitive to outliers than a windowing with identical MVA. But, in both cases, the outliers can be removed first, just as with the trimmed arithmetic mean.

The other considerations, such as the number of windows or the mathematical criterion used to design the windows (e.g., obtaining identical MVA) are, in most cases, less important for a very first data analysis. Thus, the *multiplicity* of MFMV database analysis is an important aspect to keep in mind because it provides more specifics to choose from when interpreting the results. In addition, the notion of *information* should also be taken into account, as explained below.

6.2. Information

In HCS studies, the notion of *information* is highly important throughout the analysis process, especially when dealing with a large MFMV database generated from an experimental or observational design. For a very first analysis, cutting the space scales and maybe the time scales (for chronology axis and/or duration axis) leads to the appearance of category scales. (Though frequency windows can also be considered, in most cases, the resulting data are not very helpful when investigating complex HCS time data.). The space-time windowing is linked to the number V of time variables.

If $V = 1$, the case illustrated by the eye movement data is obtained (e.g., one variable indicates the gaze position where S is the number of semantic zones), but many other cases can be imagined. Time windows could correspond to transient and steady-state parts in a manual response to a stimulus during a tracking task with sudden changes of the target, with space windows coming from windowing the peak, mean or RMS values (the values can also be evaluated for error signals). Or, time windows could correspond to electrocardiogram phases obtained during a task involving stress, with space windows coming from windowing the duration, amplitude and area values [6].

If $V > 1$, one set of possible situations corresponds to the examples described in the previous paragraph but for $I = V$ individuals working together (e.g., V individual electrocardiograms must be studied). A second set of situations occurs when V variables are considered together. This could be multidimensional signals as in our driving data example, or it could be the variables corresponding to the positions of several body parts (of one or several individuals) or of objects during a game (e.g.,

chess, tennis, football) or during any task (e.g., engine or process control, sorting or assembling of objects). The interested reader can consult references [9,10] for further examples.

Thus, with space windows, the classic notion of information theory can be introduced in the data characterization step. For instance, via *entropy*, the literature proposes a connection between information and transition matrices (conditional entropy) [10] or between information and correspondence tables (joint entropy) [19]. In addition, the notion of information can be considered in the table analysis sub-step. For instance, if correspondence analysis is used to study a two-entry table, the principle of spectral decomposition is involved, and some authors have proposed *information accounted for by one main component* [19]. Instead of these classic quantitative aspects of information, we consider a more qualitative aspect, that of the *smallest statistical entity chosen to give potential results for the final users* (e.g., physicians, ergonomists, psychologists, engineers). We consider that the smallest statistical entity (*sse*) is a function of the elementary pair “(time sample, empirical situation)”, in which the empirical situation is:

- For an experimental design, a given individual in a given experimental condition (i.e., a cell of the hyperparallelepiped H (Fig. 1)).
- For an observational design, a given individual.

Consequently, for a preliminary statistical analysis of a MFMV database, the *sse* may correspond to a row in a data table. However, with complex time data, one time sample rarely yields one *sse*. Let us consider the case of eye movement data for certain potential users:

- For psychologists, the individual dwells (i.e., successions of fixations on a same semantic zone) could be seen as *sse*. One of the main problems would be to obtain the chronology of the dwells to see if the advertising image was well designed, by answering a number of questions (e.g., Which semantic zone is scanned first? What is the next dwell? How many dwells are performed before reading the descriptive text?) [20].
- For physicians, the individual time samples could be seen as *sse*. One of the main problems for these users would be to obtain information about eye movement abnormalities, for instance, of patients with schizophrenia [26].
- For sociologists, dwells for individual groups could be seen as *sse*. One of the main problems for these users would be building homogeneous groups *a priori*, according to specific factors (e.g., sex, age, or education level).

Since an eye tracker is a rather expensive measurement device, costing about \$20.10³, and is usually bought specifically to establish the fixation position (i.e., what the human being is looking at), defining the notion of *information* precisely in this very specific context of eye movement data could give the idea that the only important piece of information is the location and maybe the duration of the fixation. Still, the eye tracker provides a fixation position, assumed to reflect central field of vision. This device provides no idea about what the brain is really doing, even though an experimenter, carefully examining a replay of the eye movement chronology superposed on the advertising image, may find that it is not inconsistent with the image features, even sometimes thinking that he/she might have made the same movements in the same order.

Let us now consider the car driving experiment. Here again several *sse* can be built:

- For physicians, the individual recording periods (i.e., 72 km driving without any rest), could be seen as *sse*. One of the main problems could be to get connections between the driving data and the data more directly linked to the vigilance, such a physiological data (e.g., heart rate and blood pressure) and subjective data (e.g., fatigue and boring level), these two data-sets being recorded during the rest just after the driving period.
- For psychologists, the individual sections (i.e., 2 km where the road has a specific profile – R, r, s, l or L , see the beginning of Section 5) could be seen as *sse*. Here the main objective could be to see how drivers take risks when they make a specific turn and how this risk taking evolves over time.
- For engineers, the time samples (at 6 Hz) of the best drivers could be seen as *sse* in the perspective of designing an automatic control system that improve driver performance in the turns.

Given the different kinds of information than can be drawn from the *sse*, there are two main considerations for space windowing:

- (1) Space windowing adapted to each variable allows individual behaviors to be compared. This is the case when the stimulus evolves identically for all the individuals (e.g., with eye movement in advertising, whether a mere picture as with our experiment or a set of pictures as with a TV ad). If the evolution of the stimuli is linked to the human component actions (e.g., in our driving example), the main drawback of this kind of space windowing is that obvious results can be found, such as there are fast or slow drivers.
- (2) Space windowing adapted from a set of *sse* often allows individuals strategies to be compared. A significant example is when an individual presents a nominal value for a given variable (e.g., the driving speed) that is different from another individual. In such a case, the space windowing may be adapted to the individual.

7. Conclusion

Given a large MFMV database containing complex time data, we have shown that many statistical analysis paths are available, each one providing different kinds of “information”, with this notion being highly dependent on the domain of the specialist (e.g., engineering, psychology, medicine, sociology, ergonomics). Thus, we feel that, for the very first analysis, it is better to use a procedure based on space windowing (and maybe time windowing also), even though choosing the membership functions for both the space and time axes is rather difficult. If the MFMV database is very large (H0 contains hundreds of cells with dozens of measurement variables), it may be possible to omit time windowing during this very first analysis. As considered in our examples, it is interesting to test several descriptive analysis paths (with and without the chronological aspect of time). Since this very first analysis also helps to highlight doubtful data, the data itself and the results of the analysis are highly trustworthy, and they can be used without hesitation in subsequent analyses involving more local studies. In most cases, such analyses are based on statistical inference, including tests of factor influences (both time and other factors) and tests of inter-variable connections.

References

- [1] R.L. Allen, D.W. Mills, *Signal Time Analysis, Frequency, Scale and Structure*, Wiley-IEEE, 2004.
- [2] A. Andrade, P. Kyberd, S. Nasuto, The application of the Hilbert spectrum to the analysis of electromyographic signals, *Information Sciences* 178 (2008) 2176–2193.
- [3] F.J. Anscombe, Graphs in statistical analysis, *American Statistician* 27 (1973) 17–21.
- [4] H. Barreau, Temps, In: *Encyclopaedia Universalis*, Paris, 1989.
- [5] J.P. Benzecri, *Correspondence Analysis Handbook*, Marcel Dekker, New York, 1992.
- [6] E.J. Berbari, Principles of electrocardiography, in: J.D. Bronzino (Ed.), *The Biomedical Engineering Handbook*, CRC Press, New York, 2000.
- [7] T. Cox, The recognition and measurement of stress: conceptual and methodological issues, in: J.R. Wilson, E.N. Corlett (Eds.), *Evaluation of Human Work*, Taylor & Francis, London, 1990.
- [8] A.T. Duchowski, Eye tracking methodology, in: *Theory and Practice*, Springer, London, 2003.
- [9] J.M. Gottman, A.K. Roy, Sequential analysis, in: *A Guide for Behavioral Researchers*, Cambridge University Press, Cambridge, 1990.
- [10] P. Haccou, E. Meelis, *Statistical Analysis of Behavioral Data: An Approach Based on Time-Structured Models*, Oxford University Press, Oxford, 1992.
- [11] M.-C. Hung, M.-L. Huang, D.-L. Yang, N.-L. Hsueh, Efficient approaches for materialized views selection in A data warehouse, *Information Sciences* 177 (2007) 1333–1348.
- [12] J.D. Jobson, *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods*, Springer, New York, 1992.
- [13] A. Kilbom, Measurement and assessment of dynamic work, in: J.R. Wilson, E.N. Corlett (Eds.), *Evaluation of Human Work*, Taylor & Francis, London, 1990.
- [14] E. Kowler, The role of visual and cognitive processes in the control of eye movement, in: E. Kowler (Ed.), *Eye Movements and Their Role in Visual and Cognitive Processes*, Elsevier, Amsterdam, 1990.
- [15] Z. Ladin, Three dimensional instrumentation, in: P. Allard, I. Stokes, J.P. Blanche (Eds.), *Three Dimensional Analysis of Human Movement*, Human Kinetics Champaign, 1995.
- [16] J.M. Legay, Quelques réflexions sur le plan expérimental, *Statistique et analyse de données* 11 (1986) 51–57.
- [17] R. Mead, *The Design of Experiments: Statistical Principles for Practical Application*, Cambridge University Press, 1988.
- [18] M.R. Neuman, Physical measurements, in: J.D. Bronzino (Ed.), *The Biomedical Engineering Handbook*, CRC Press, New York, 2000.
- [19] S. Nishisato, *Multidimensional Nonlinear Descriptive Analysis*, CRC press, Chapman & Hall, New York, 2006.
- [20] K. Rayner, Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin* 124 (1998) 372–422.
- [21] R.B. Stammer, M.S. Carey, J.A. Astley, Task analysis, in: J.R. Wilson, E.N. Corlett (Eds.), *Evaluation of Human Work*, Taylor & Francis, London, 1990.
- [22] S.S. Stevens, *Scaling: A Sourcebook for Behavioral Scientists*, Aldine Publishing Co., Chicago, 1974.
- [23] E.R. Tufte, *Envisioning Information*, Graphic press, Cheshire, Connecticut, 1983.
- [24] E.R. Tufte, *The Visual Display of Visual Information*, Graphic press, Cheshire, Connecticut, 1990.
- [25] B. Walliser, *Systèmes et modèles, Introduction critique à l'analyse des systèmes*, Ed. Seuil, Paris, 1977.
- [26] A. Wolff, G.A. O'Driscoll, Motor deficits and schizophrenia: the evidence from neuroleptic-naïve patients and populations at risk, *Journal of Psychiatry Neurosciences* 24 (1999) 304–314.
- [27] R. Yager, D. Filev, Summarizing data using a similarity based mountain method, *Information Sciences* 178 (2008) 816–826.
- [28] S.-M. Zhou, J.Q. Gan, F. Sepulveda, Classifying mental tasks based on features of higher-order statistics from EEG signals in brain-computer interface, *Information Sciences* 178 (2008) 1629–1640.