Engineering Applications of Artificial Intelligence 25 (2012) 989–996

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Multivariate analysis of human behavior data using fuzzy windowing: Example with driver–car–environment system ☆

Jean-Christophe Popieul [a,b,c,*], Pierre Loslever [a,b,c], Alexis Todoskoff [d], Philippe Simon [a,b,c], Matthias Rötting [e]

[a] Université Lille Nord de France, F-59000 Lille, France
[b] UVHC, LAMIH, F-59313 Valenciennes cedex 9, France
[c] CNRS, UMR 8201, F-59313 Valenciennes cedex 9, France
[d] Université d'Angers, LASQUO EA3858, F-49000 Angers, France
[e] Chair of Human-Machine Systems, Technische Universität Berlin, D-10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 2 May 2011
Received in revised form
13 October 2011
Accepted 29 November 2011
Available online 13 February 2012

Keywords:
Multiple correspondence analysis
Descriptive analysis
Fuzzy windowing
Car driving
Driving simulator

ABSTRACT

In most human component system studies performed in simulators, several factors (or independent variables) (at least two, i.e., individual and time) and many variables (or dependent variables) are present. Large and complex databases have to be analyzed. Instead of using rather automatic procedures, this article suggest that, for a very first analysis at least, the human being must be present and he/she must choose a method being adapted to the data, which is different to run a method supposing that the data fit such or such model. This article suggests starting the analysis while keeping both the multifactorial (MF) and multivariate (MV) aspects. To achieve this aim, with the possibility to show nonlinear relationships, a MFMV exploration of the experimental database is performed using the pair (*fuzzy space windowing, Multiple Correspondence Analysis*). Then may come an inference analysis. This long (due to multiple large graphical views) but rich procedure is illustrated and discussed using a car driving study example.

1. Introduction

Most system empirical studies yield large and complex databases, large because data is recorded using high frequency sampling (compared to the system dynamics) and/or during a long time period and complex because data is heterogeneous. This is particularly true in the areas of human component system studies (psychology, medicine, ergonomics or sociology) with the possible presence of different scale mathematical models (nominal, ordinal and quantitative) (Stevens, 1974), of objective and subjective origins and of time and not-time variables. Whatever the strategy employed in conducting studies (observational, experimental or correlational method) (Sheskin, 2007), the disciplinary fields focusing on data exploitation (e.g., Statistics, Pattern Recognition, Signal Analysis, Data Mining, Artificial Intelligence) propose many methods. Even though these fields are not independent, here are some possible Taxonomic Dimensions (TD) that are often cited:

- TD1, descriptive vs. inferential methods (Sheskin, 2007),
- TD2, monovariate vs. multivariate methods (Jobson, 1991),
- TD3, time vs. not-time methods (Fitzmaurice et al., 2004),
- TD4, supervised vs. unsupervised methods (Pal and Pal, 2001),
- TD5, fuzzy vs. ordinary sets based methods (Komen and Schneider, 2005),
- TD6, probability vs. possibility theory methods (Dubois and Prade, 1988),
- TD7, connexionist vs. analytic methods (Silipo, 2007).

Conceiving and executing an empirical study is a long and complex task so that as soon as all the time data sets have been recorded (or maybe one time dataset only), the system analyst may be in a hurry to get results. Then for questions of speed or lack of knowledge or curiosity, the analyst might choose a fast and well known data analysis path. For instance, in the case of a study performed using an experimental design (e.g., a chemical process in an industrial study), the researcher may rapidly want to test the statistical hypotheses that the experiment was primarily designed to test. The same behavior may occur in the case of a study performed using an observational design (e.g., a system including users and designers in a web use analysis over several years).

Faced with a large and complex database, instead of using rather automatic procedures, we suggest that, for a very first analysis at least, the human being must be present and he/she must choose a method being adapted to the data, which is different to run a method supposing that the data fit such or such model. For instance, with many signals, running the

☆ The material in this paper was partially presented at the 2010 IFAC Conference (HMS 2010), August 31–September 3, 2010, Valenciennes, France.

* Corresponding author at: UVHC, LAMIH, F-59313 Valenciennes cedex 9, France.
E-mail address: Jean-Christophe.Popieul@univ-valenciennes.fr (J.-C. Popieul).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.11.011

arithmetic mean program for summarizing time data has a poor meaning if the signals present sudden changes or a monotone evolution. In the same way, with experimental design data, running the usual variance analysis computing program has a poor meaning if the data is very far from the Laplace–Gauss model or the variances related to the factor levels are very different. Our point of view is to start the data analysis using knowledge and graphics as following.

2. Data analysis methodology

Let us consider this problem through the analysis of a system, where:

- the inputs are the factors, which may be linked to the system or its environment and may have a nominal, ordinal and quantitative scale (in an experimental design, there are specific factors named independent variables) (Sheskin, 2007).
- the outputs are the (measurement) variables, here again possibly with different kinds of scales (in an experimental design, these variables are often named dependent variables) (Sheskin, 2007).

This notion of input and output sets can be used to attain four main objectives for the multifactor–multivariate (MFMV) data analysis:

- to determine the effects of the input variables on the output variables (objective named O1).
- to discover the relationships between the output variables (O2).
- to find classes of variables and/or of empirical situations (human beings, countries, plants, jobs, system behavior states, …) (O3).
- to merely find data summaries (about failures, diseases, incomes, costs, …) (O4).

Let us first focus on the case where the time factor is present in the database. To reach these objectives with MFMV data, the statistical analysis process cannot generally give results with a single stage. For instance, instead of using hypothesis testing statistical tests (i.e., inference statistics), as usually with MFMV experimental data, we suggest starting this process using the pair *space-time windowing (STW)/Analysis showing factors and variables behaviors (ASFVB)* and then, ending with such tests.

Let V be the measurement variable set where V, the size of V, is large, e.g., $V>10$ when mechanical and electric variables are present in a mechatronical system study or when physiological and action variables are present in a human component system study. Let U be the factor set which often contains more than $U=4$ elements, e.g., individual (human being, tool, machine tool, plant …), time and manipulated variables in an experimental design. As stated above V, U may contain several scale models. To reach results with a heterogeneous MFMV database, up to 5 data processing stages may be necessary (Loslever, 2001):

1) Data characterization. The V time variables are characterized using V' synthetic indicators. With heterogeneous MFMV data, the main idea is to carefully analyze the magnitude histogram of each time variable, either globally or for specific time windows, which allows doubtful data to be differentiated (mainly shown in modes or in the lateral sides of the histogram). This analysis is used to build the indicator set V', which becomes the analysis variable set.
2) Data coding. The V' variables are coded to make the variables compatible either with one another (particularly when both quantitative and qualitative scales are involved) or with a specific method (particularly with quantitative variables). For instance, a rank data set becomes the input of the Principal Component Analysis in the descriptive context (Jobson, 1991) and the input of a nonparametric test in the inferential context (Sheskin, 2007).
3) Data organization. For a preliminary analysis, the data is organized so that it can be investigated using multivariate techniques. In most cases, the data is arranged in one table whose rows correspond to the statistical units (e.g., individuals, experimental situations, time windows) and whose columns correspond to the analysis variables.
4) Relation, effect, class or summaries obtaining (objectives O1 to O4). For instance, the data may be analyzed to determine relationships either between the V' analysis variables and the U factors effects or between the V' analysis variables; the relations and/or the effects can also be found using classification and discrimination techniques (Han and Kamber, 2006; Jobson, 1991; Pal and Pal, 2001; Fitzmaurice et al., 2004).
5) Results presentation. The results are presented in up to three kinds of models: graphic (histograms, time excursions, scatterplots, spectrums) (Tufte, 1983; Pao and Meng, 1998), verbal (the conclusion to a statistical hypothesis test) (Sheskin, 2007) or mathematical (multivariate regression, time series models) (Jobson, 1991).

Thus a data analysis path (DAP) is a succession of A stages ($A\leq 5$) with a specific method m ($m=1, \ldots M_a$) for each stage a ($a=1, \ldots, A$). Facing a heterogeneous MFMV data set, the data analysis can be performed taking several DAPs. Obviously, at each stage, several loops may be necessary. For instance, in the choice of DAP set, three main "tactics" are possible:

- Descriptive analysis only, DAPs are mainly based on graphic and data summaries,
- Inferential analysis only, DAPs are based on hypothesis tests, or
- Combination of the two, (1) descriptive, (2) inferential, with several loops maybe.

Of course, the stage four mainly conditions a DAP. For instance, to show relationships between variables and the time influence in the descriptive DAP, the basic principle is often the singular value decomposition (Jobson, 1991), which yields methods such as Principal Component Analysis (PCA) or Multiple Correspondence Analysis (MCA). Then, for each method, several options are available, conditioning the choices of stages 1 to 3. For instance, PCA can be used with several coding techniques (Jobson, 1991) and MCA with either crisp or fuzzy windowing (Benzecri, 1992). The choice of a method and the respective options for the stage 4 is quite difficult. For instance, MCA with space windowing shows nonlinear relationships, but the output is much more complex than PCA output. Let us now consider the pair *STW/ASFVB* with actual data as the main elements of the descriptive DAP.

3. Didactic example

The example concerns the Driver–Car–Environment System (DCES). The aim of the study considered here was to analyze the effect of road geometry—roads with or without curves, with or without a constant lane width—on car driving behavior, while taking time and individual differences into account. There are thus $V=4$ factors, one corresponding to the individual, two corresponding to the road geometry and the last one corresponding to the time. The experiment was conducted using a driving simulator (designed by PSA Peugeot–Citroën). A two-lane road