REGULAR ARTICLE

Strategies evaluation in environmental conditions by symbolic data analysis: application in medicine and epidemiology to trachoma

Christiane Guinot · Denis Malvy · Jean-François Schémann · Filipe Afonso · Raja Haddad · Edwin Diday

Abstract Trachoma, caused by repeated ocular infections with Chlamydia trachomatis whose vector is a fly, is an important cause of blindness in the world. We are presenting here an application of the Symbolic Data Analysis approach to an interventional study on trachoma conducted in Mali. This study was conducted to choose among three antibiotic strategies those with the best cost-effectiveness ratio and to find the demographic and environmental parameters on which we could try to intervene. The Symbolic Data Analysis approach aims at studying classes of individuals consid-

C. Guinot
Laboratoire d'Informatique, Université François Rabelais de Tours, 37200 Tours, France
e-mail: christiane.guinot@univ-tours.fr

D. Malvy
Département de Médecine Interne, CHU St-André and Université
Victor Segalen Bordeaux 2, 33076 Bordeaux, France
e-mail: denis.malvy@chu-bordeaux.fr

J.-F. Schémann
Institut de Recherche pour le Développement, Dakar, Senegal
e-mail: jfschemann@wanadoo.fr

F. Afonso (✉)
SYROKKO, Aéropôle, BP 13918, 95731 Roissy CDG, France
e-mail: afonso@syrokko.com

R. Haddad
SYROKKO and LAMSADE, Paris Dauphine University,
Place du Maréchal De Lattre De Tassigny, 75016 Paris, France
e-mail: haddad@syrokko.com

E. Diday
CEREMADE, Paris Dauphine University, Paris, France
e-mail: diday@ceremade.dauphine.fr

ered as new units. These units are described by variables whose values express for each class the variation of the values taken by each of its individuals. Finally, the results obtained are compared to those previously provided by multiple logistic regression analysis. Symbolic Data Analysis actually provides a new perspective on this study and suggests that some demographic, economics and environmental parameters are related to the disease and its evolution during the treatment, whatever the strategy. Moreover, it is shown that the efficiency of each strategy depends on environmental parameters.

Keywords   Symbolic Data Analysis · Multiple logistic regression · Trachoma

Mathematics Subject Classification   62-07 · 62-09

1 Introduction

Symbolic Data Analysis (SDA) (Bock and Diday 2000; Billard and Diday 2003, 2006; Diday and Noirhomme 2008; Diday 2011) is based on two steps. In the first step, classes of individuals are described by taking into account the variation in the values of the variables that characterize them. This variation is expressed, for example, by a bar chart when a variable is qualitative and by an interval when it is quantitative. As a bar chart and an interval are not numbers, these data are not numerical, therefore they are called "symbolic" data. The resulting data table is called "symbolic" data table where the rows describe the classes and the columns are the variables, which are called "symbolic" variables. This data table contains in each of its boxes a bar chart or an interval, if it is limited (as in this paper) to these two expressions of the variation inside the classes. In the second step, data analysis and data mining methods, such as clustering and principal component analysis extended to this type of data, can be applied.

Symbolic data analysis approach has been applied to an intervention study on trachoma. Trachoma disease caused by repeated ocular infection with Chlamydia trachomatis, which the principal vector is a fly, is a major cause of blindness in the world (WHO 1988). Mass azithromycin distribution is part of current recommended strategies for controlling trachoma. This intervention study was conducted in Mali, in an area where trachoma is endemic, in order to ascertain an efficient strategy model at an acceptable cost. The treatment consisted in the administration of one dose of azithromycin according to three strategies: the first strategy (I) was to treat all residents, the second (II) was to treat all children under 11 years and all women aged 15 to 50 years, finally the third (III) treated all inhabitants of a concession when at least one children had an active trachoma diagnosed. A classical analysis of these data, based on a multiple logistic regression model (Hosmer and Lemeshow 2000), was conducted by Schemann et al. (2007). In order to discover among the demographic and environmental parameters those on which we could try to intervene, it was decided to conduct further analyses using SDA. Our aim is not to extend the logistic regression to symbolic data as done in Souza et al. (2011). The objective of this complementary analysis is to show that a symbolic approach based on bar charts is a smart tool for studying classes of individuals (according to the evolution of the disease during the

study), by enriching the results obtained by classical methods designed for individuals, by additional tools and graphical easily interpretable representations of the classes.

2 Symbolic data analysis applied to an interventional study on trachoma

2.1 Reminder of the intervention study and data studied

The intervention study involved 6532 individuals described by 31 variables. These individuals lived in nine villages composed of several hundreds of concessions, each formed by several families. Three villages were selected for each of the three treatment strategies. The villages have been grouped to distribute the disparities of the demographic and environmental characteristics in each treatment strategies in the most homogeneous manner. According to the strategy allocation, adults were eventually given a single dose of 1g azithromycin, and children a unique dose of 20 mg/kg. Due to technical reasons, the inclusion was done according two recruitment periods: the first (May 2000–June 2001) concerns four villages, the second (January 2001–February 2002) the five remaining villages. The data collected were age, gender and weight. Moreover, cleanliness and washing of children's faces were assessed, and additional questions were addressed about education, environmental and socio-economic conditions for each household at baseline, such as the number of children attending school, the distance to the water of each concession, the existence of garbage dumps, of latrines, of sheepfold, of cowshed, of ploughs, the number of sheep or goat, of cows, of donkeys, the existence of inner well, and of borehole well. Ophthalmic examination was performed at baseline and one, six and twelve months after inclusion. The outcome variable was clinically active trachoma frequency, twelve months after intervention, among children under 11 years of age.

This study was conducted in accordance with the Helsinki declaration, good epidemiological practice and the Malian legislation. In addition, the proposal was approved by a Malian ethical committee supported by the National Institute of Health (NIH, USA).

2.2 Choice of classes

The first step in SDA approach is to choose the classes of individuals to be studied. We have chosen a partition into four classes according to the evolution of the disease during the study. These classes have been noted 0×0, 0×1, 1×0, 1×1. The class 0×0 contains all individuals free of trachoma at baseline who were still free of trachoma at the end of the study, and the class 0×1 all individuals free of trachoma at baseline who developed trachoma during the study. Besides, the class 1×0 contains all individuals with trachoma at baseline who became free of trachoma during the study and the class 1×1 all individuals with trachoma at baseline, who were still with trachoma at the end of the study.

3 Methods and tools used in this work and advantages

We mainly use three methods. The first one is a discretization method that allows for the transformation of a quantitative variable into a qualitative variable. This discretiza-

Fig. 1 Characterization by relative frequency of *bar charts* categories of the classes according to the evolution of trachoma: individuals free of trachoma at baseline and at the end of the study (0×0), individuals free of trachoma at baseline who developed trachoma during the study (0×1), individuals with trachoma at baseline who became free of trachoma during the study (1×0), and individuals with trachoma at baseline and at the end of the study (1×1)

tion produces bar charts, induced by the given classes, when initial individual data are aggregated up to symbolic data. Thus, based on the principle of classical supervised discretization methods, the method implements discretization criteria which take into account the different classes and find the bounds that give the most discriminant bar charts for these classes. In this application, the criterion is based on L_1 distance between bar charts variables. We can consider that it is an extension of Fisher's algorithm for the maximization of the L_1 distance between bar charts induced by the given classes instead of unsupervised minimization of the within-class inertia (see Diday et al. 2013 for more details).

The second method produces from the individual data, a symbolic data table, where each row is associated with a class of individuals, and each column is associated with a symbolic variable. For example, in Fig. 1, the value of the symbolic variable called "Strategy", for the class 0×0, is the bar chart of the initial standard variable giving the strategy used for each individual free of trachoma at baseline and at the end of the study.

The third method (see Diday 2013) is an extension of principal component analysis (PCA) to symbolic data. It is based on the following principle: first, a standard PCA is applied on the symbolic data table where each category (i.e., bin), of each symbolic variable describing the classes, is considered as a numerical variable. Then, it is possible to represent a factorial plane where each point is associated with a class as well as a correlation circle where each variable is associated with a bin.

Concerning the factorial plane, it is possible to represent each point by the value taken by a symbolic variable on the class associated with this point. Hence, depending of the kind of chosen symbolic variable, we can obtain a point representation by a bar chart or an interval.

Concerning the correlation circle, it is possible to project not only the bins but also the symbolic variables themselves considered as principal components restricted to their bins in case of symbolic bar chart value variables (see Diday 2011, 2013). Moreover, in Diday (2013) it has been shown that these symbolic variables can only be projected inside the hypercube of p dimensions whose projection on each factorial plane is a square with vertices $(1, 1)$, $(-1, 1)$, $(-1, -1)$, $(1, -1)$.

For information, these three methods have been implemented in the SYR software (Afonso et al. 2014) which has been applied on these trachoma data.
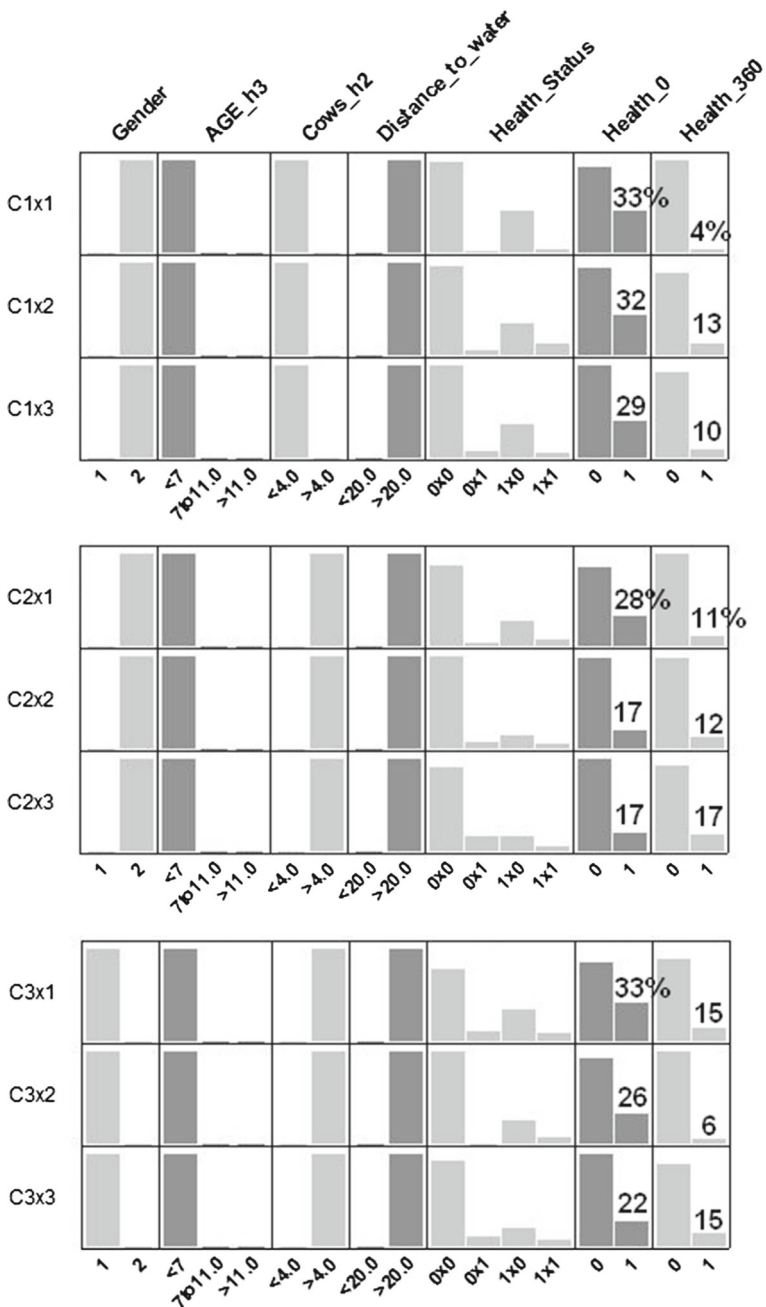
4 Results

4.1 Description of the classes

The symbolic description of the four classes is shown in Fig. 1 where each box contains a bar chart. In rows 0×0 to 1×1, the relative frequencies of categories for each variable are represented. In this table, the last row presents the variation between the smallest and largest values of frequency of each category of each symbolic variable. The symbolic variables are ordered from left to right according to the decrease of these variations.

The first two columns of Fig. 1 show that the class of individuals free of trachoma at baseline and at the end of the study (0×0) is mainly composed by individuals over 11 years old (category 3 of the variable AGE_h3), while individuals with trachoma at baseline and at the end of the study (1×1) were younger and lighter (category 1 of AGE_h3 and weight_h3). As indicated in the third column, the strategy III appears more frequently in the class of the individuals free of trachoma at baseline who developed trachoma during the study (0×1). For the last class, individuals with trachoma at baseline who became free of trachoma (1×0), strategies I and II are more frequent than in the other classes.

We can see in the table of Fig. 1 that the strategy I leads to improvement as it is the less frequent in class 0×1 and has high frequency in class 1×0, the strategy II to stabilization as it is the most frequent in classes 0×0 and 1×1 , or even to improvement as it is the most frequent in class 1×0, and the strategy III to degradation as it is the most frequent in class 0×1 and the less frequent in class 1×0. But is it the third strategy or the demographic and environmental circumstances that makes degradation (i.e., persistence or development of trachoma during the study)?

4.2 Efficiency of the strategies in some environmental situations

In the same way as we have looked, in Fig. 1, for the variables that generate the most variation between the four classes 0×0, 0×1, 1×1 and 1×0, we have looked for the variables that generate the most variation between classes 0 (without trachoma) and 1 (with trachoma) at baseline. These variables are Gender, Age (< 7, 7–11, >11 years old including adults), Cows (less or more than 4 cows), Distance to water (less or more than 20 km). Then, in order to analyze the efficiency of the three strategies for different environmental situations, we consider classes of environmental contexts by

Fig. 2 Comparison of the three strategies (1,2,3) for the three different environmental situations (C1–C3 defined by the variables Gender, Age_h3, Cows_h2 and Distance_to_Water) with the most disease problems. For instance, C1×2 gives the results of strategy 2 for environmental class C1 when we compare Health_0 with Health_360 and Health_Status

performing the cartesian product of all the categories of these variables. Thanks to these cartesian product we obtain 20 environmental classes.

We retain the three classes (C1–C3) with the highest trachoma frequency at the beginning of the treatment (variable Health_0). In Fig. 2, we can see that C1 is the class of girls, <7 years old, with less than 4 cows in their environment and distance to water over 20 km. We have respectively 132, 166 and 129 children for each strategy. We see thanks to Health_0 (category 1) that nearly a third of children are suffering from the disease at the beginning of the treatment. C2 is the class of girls, <7 years old, with more than 4 cows in their environment and distance to water over 20 km. We have respectively 93, 91 and 87 children for each strategy. C3 is the class of boys, <7 years old, with more than 4 cows in their environment and distance to water over 20 km. We have respectively 97, 95 and 107 children for each strategy.

For each class of environmental situation, we look at the results obtained with each strategy (1–3) one year after the treatment (Health_360 and Health_Status for more details). For instance, C1×2 gives the results of strategy 2 for environmental class C1 when we compare Health_0 with Health_360 and Health_Status.

When we look at the improvement of the frequency of children with trachoma between Health_0 and Health_360 (category 1), we note that, for the environmental situations C1 and C2, strategy 1 obtains better results than strategy 2 and 3 but for C3, strategy 2 obtains better results than the others. Moreover, strategy 3 is clearly less efficient than the others for C3 and not efficient at all for C2.

Other classes have been studied, in particular according to the relative frequency of the degradation class (0×1) in each village. To do this, we grouped the people of the nine villages into two classes depending of the relative frequency of this degradation class. The villages were firstly reordered according to the increasing of this relative frequency. We obtain the order: 8, 2, 3, 6, 4, 1, 9, 7 and 5. Then, the villages were grouped into two classes: class C1 that contains the five first villages with the lowest relative frequency of the degradation class: (8, 2, 3, 6 and 4), and class C2 that comprised the remaining villages. We note from Fig. 3 that the class with the lowest relative frequency of degradation (Partition_2_C1) has been processed less often by strategy 3, and that it is associated with fewer borehole wells, deposits of garbage, cows, ploughs, donkeys and oxen.


Fig. 3 Comparison of the Partition_2_C1 class, that groups the five villages with the lowest frequency of degradation (0×1), against Partition_2_C2 containing the four villages with the highest frequency of degradation (0×1). Variables are ordered by decreasing variation

4.3 Principal component analysis of the symbolic data

Now, let us examine the results of PCA of the symbolic data described in Fig. 1 where the 4 classes, defined according to the evolution of trachoma (0×0, 0×1, 1×0 and 1×1), are described by the following selected symbolic variables: Gender, Garbage, Age_ h3, Oxen, Water distance, Latrine, Cows, Ploughs, Borehole well, Strategy.

Figure 4 shows the map of proximities between the classes. The first and second principal components of the PCA were used as an axis system (in the left of Fig. 4). The first component explains 53 % of the total variance, and the second component 28 %. On this figure, the first axis opposes individuals free of trachoma at baseline who developed trachoma during the study (0×1), i.e., the degradation class, to the classes of individuals free of trachoma at baseline and at the end of the study (0×0), and the class of the individuals with trachoma at baseline who became free of trachoma during the study (1×0). The second axis opposes the class of individuals free of trachoma at baseline and at the end of the study (0×0) at the top, to the class of the individuals with trachoma at baseline who became free of trachoma during the study (1×0), at the bottom. The remaining class (1×1), i.e., individuals with trachoma at baseline and at the end of the study, is located in the middle of the map because it is not well represented on the first factorial plane but on the second one (in the right of Fig. 4) by using axes 1 and 3. Moreover, we can also see the size of each class and notice that
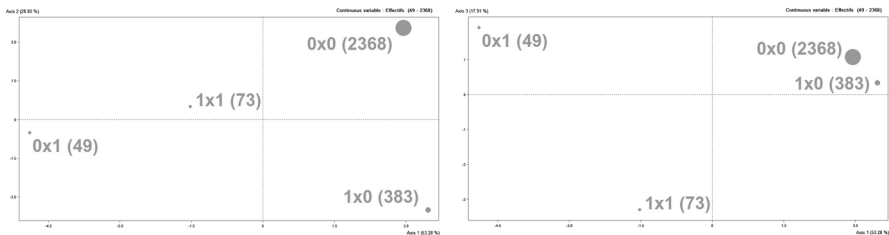
Fig. 4 Map of proximities between the four classes according to the evolution of trachoma on the first (*left*) and second (*right*) factorial plane, and representation of the size of each class (figures in parentheses = numbers of individuals)
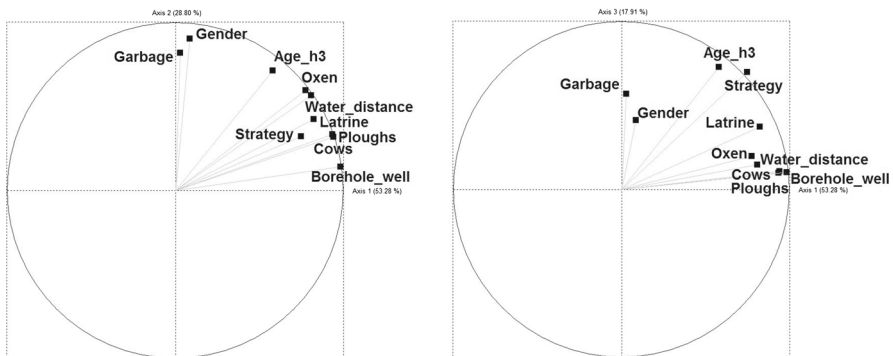

Fig. 5 Representation of symbolic variables on the first (*left*) and second (*right*) factorial plane

the class of individuals free of trachoma at baseline and at the end of the study (0×0) is happily the largest class.

In Diday (2011), it has been shown that the symbolic variables can be projected in the first quadrant of the PCA of the smallest hypercube containing the correlation sphere. The first quadrant of the circle of correlation of the two first axes and those of the first and third axes are given Fig. 5 (on the left and the right of the figure respectively). It can be seen that the symbolic variables borehole well, ploughs, number of cows, latrines, distance to the water characterize the first axis (oxen and strategy also but less than the previous variables), the symbolic variables gender, garbage dumps and age characterize the second axis, and that garbage dumps, strategy and the age characterize also the third axis.
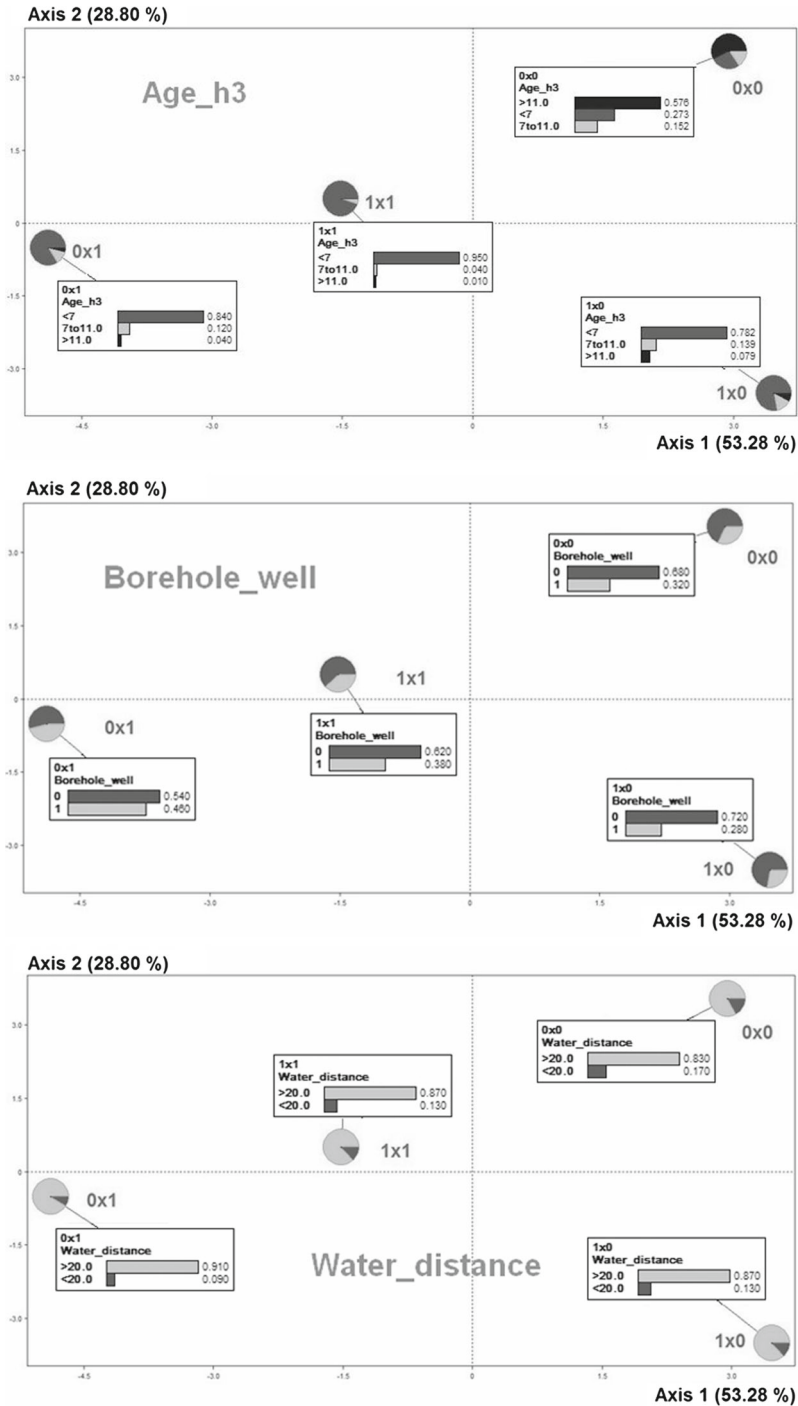
Figures 6 and 7 show the circle of correlations of the categories of the symbolic variables on the first factorial plane and on the second factorial plane, respectively.
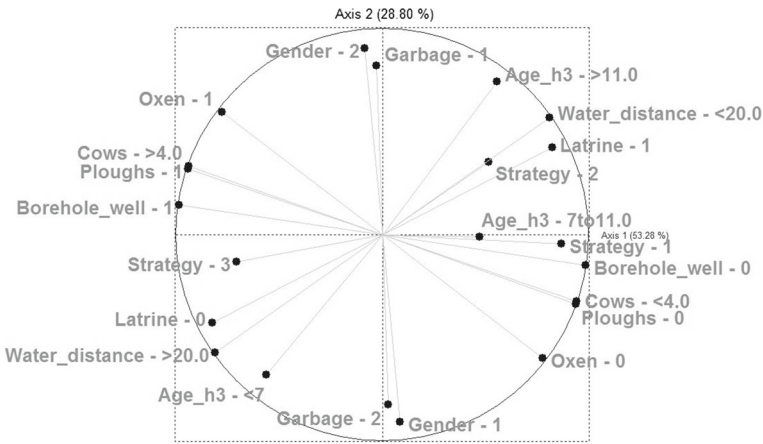
Fig. 6 *Circle* of correlations of the symbolic variables categories on the 1st factorial plane
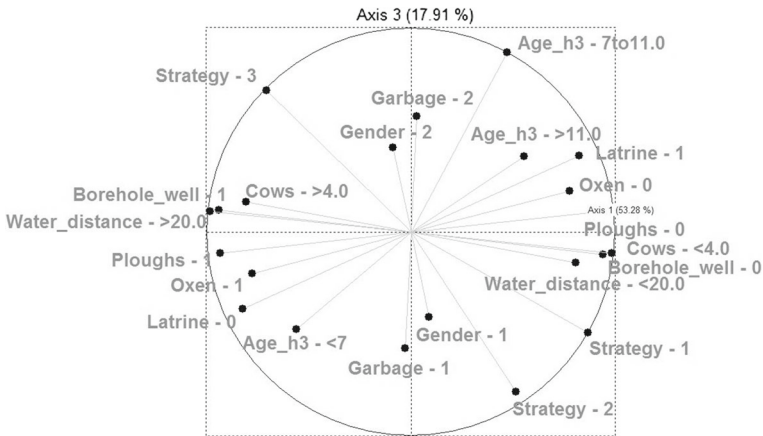
Fig. 7 *Circle* of correlations of the symbolic variables categories on the 2nd factorial plane

Fig. 8 Maps showing the distribution in the categories of "Age", "Borehole well" and "Water distance" on the first factorial plane

Figure 6 shows that at lower left quadrant, on the side of the class of degradation
(0×1), are located the categories of the symbolic variables that characterize this class:
larger frequency of strategy III, more borehole well, more than four cows, more ploughs
and oxen, less latrines and greater distance to the water. In contrast, the class free of
trachoma at baseline and at the end of the study (0×0) located at upper right quadrant,
and the class with trachoma at baseline who became free of trachoma during the study
(1×0), located at lower right quadrant, are characterised with the followings variables:
higher frequency of strategies I and II, less of borehole well, smaller number of cows,
ploughs, oxen, more latrines, and lower distance to the water. The second axis opposes
the still healthy class (0×0) upwards the healing (1×0) class down: there are more
girls on the side of the still healthy class (0×0), while there are more men and less
garbage on the side of the healing (1×0) class deposits. Besides, Fig. 7 shows that on
the side of the class the individuals with trachoma at baseline and at the end of the
study (1×1) are located the categories of the symbolic variables that characterize this
class: larger frequency of strategy II, more garbage, and age under seven years old.

Figure 8 shows the distribution of the categories of the age on the first factorial
plane. From the representation of the age on the top, it can be seen that the class
(1×1), i.e., individuals with trachoma at baseline and at the end of the study, are
almost always under seven years old. Moreover, in the same figure, we can see that
the class 0×1 of degradation has greater water distance and more boreholes well.

4.4 Comparing logistic regression and symbolic data analysis results

Logistic regression is a well-known statistical method for assessing association
between a risk factor and the probability of a disease occurrence. In Schemann et al.
(2007), the authors tested the associations between each potential risk factor and the
outcome variable, i.e., clinically active trachome occurrence among children under
eleven years of age one year after the drug administration. Then they used a multiple
logistic regression model to adjust simultaneously the potential confounding effects
of all the risk factors and covariates which were found significant at the first step of
the analysis. The results were the following. Even if some covariates showed indi-
vidual significant links, such as the age, the presence of borehole well or showed
non-statistically significant trends, such as the presence of sheep or goat in the house-
hold, the presence of sheepfold and the fact that the child attending school, almost
all these effects disappeared in the multivariate analysis, excepted the age and the
presence of active trachoma at baseline. In short, the multivariate logistic regression
model demonstrated that, after adjustment on age of the children and presence of
active trachoma at baseline (Adjusted Odd Ratio and [95 % confidence interval]: 0.81
[0.75–0.87] and 3.81 [2.70–5.39], respectively), both strategies I and II gave similar
results (Adjusted Odd Ratio and [95 % confidence interval]: 1 and 1.13 [0.74–1.72],
respectively), while strategy III was statistically less effective (1.56 [1.01–2.42]).

However, the logistic regression method has potential drawbacks which are not
generally recognized. In short, logistic regression model assumes that the probability of
the disease occurrence is linearly and additively related to the risk factors on the logistic
scale. Nevertheless, empirical evidence suggests that this assumption is generally not

true, and that the actual relationship between the risk factors and the disease is likely to be nonlinear and non-additive on the logistic scale. (For details, see Lee 1986).

The SDA advantage is to produce a symbolic data table where bar charts describe classes according to the evolution of the disease during the study. Moreover the symbolic variables can be sorted by decreasing discriminating power. It is then easy to see the contrast between the bar charts associated with each class. In this approach the variables are not defined necessarily on the same set of units (in this approach some are defined on individuals, others on households). A scalar approach would represent the classes by means of the numerical variables (on each class), which is much less informative than bar charts where the bins are built in the most discriminating way of the classes.

A scalar approach would consider the bins as standard numerical variables not the symbolic bar chart value variables. Therefore, the bar charts would not appear as visual representation as in the symbolic data table (see Fig. 1). Moreover, they would not appear in the PCA visualizations (see Fig. 8) and inside the correlation circle (see Fig. 5).

The original analysis performed using multivariate logistic regression model showed, after adjusting on age and presence of trachoma at the beginning of the study, that the strategies I and II gave similar results, while the strategy III appeared significantly less effective. Nonetheless, this analysis did not demonstrate any significant effect of the various variables describing the demographic and environmental aspects of trachoma (Schemann et al. 2007). Consequently to the SDA results given in the Sect. 3, we can conclude that the SDA approach applied on the same data set provides new insights into the data, and suggests that some demographic, economic and environmental parameters are actually related to the evolution of the disease. The Symbolic Data Analysis approach showed here its ability to handle classes of individuals, providing to classical statistical analysis, additional results and visualizations with easy interpretation.

5 Conclusion

In this paper we have proposed a SDA framework for the evaluation of different strategies and the study of the influence of the environmental conditions applied to people leaving in a region where trachoma is an important cause of blindness. Hence, the SDA framework appears useful in all domains where the strategy evaluation, depending on some environmental conditions, is needed. These domains are numerous, for sure in medical and sociodemographic epidemiology as shown in this paper, but also in many other domains. In marketing, for example, in order to evaluate (for a new product), several strategies of advertising impact, on a sample of population depending on sociodemographic conditions; in strategy evaluation of agricultural or industrial products quality depending on environmental conditions. More generally, can we say that SDA gives better results than standard approach? The main difference between SDA and the classical frameworks is that in the classical framework the units are individuals whereas in SDA the units are classes of individuals. Therefore, as we cannot say that studying species of birds gives better results than studying birds, we

cannot say that SDA gives better results than standard methods. The only things that we can say is that SDA gives complementary results to the standard approach and moreover SDA methods and tools are more general than standard ones as individuals can be considered as classes reduced to a single unit.

References

Afonso F, Haddad R, Toque C, Eliezer ES, Diday E (2014) User manual of the SYR Software. Syrokko internal publication, p 70

Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98(462):470–487

Billard L, Diday E (2006) Symbolic data analysis: conceptual statistics and data mining. Wiley, Chichester, p 321

Bock H, Diday E (2000) Analysis of symbolic data. In: Bock D (ed) Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg, p 425. ISBN 3-540-66619-2

Diday E (2011) Principal component analysis for categorical histogram data: some open directions of research. In: Fichet B, Piccolo D, Verde R, Vichi M (eds) Classification and multivariate analysis for complex data structures. Studies in classification, data analysis, and knowledge organization. Springer, Heidelberg, pp 3–15

Diday E (2013) Principal component analysis for bar charts and metabins tables. Stat Anal Data Min 6(5):403–430

Diday E, Afonso F, Haddad R (2013) The symbolic data analysis paradigm, discriminate discretization and financial application. In: Advances in Theory and Applications of High Dimensional and Symbolic Data Analysis, HDSDA 2013. Revue des Nouvelles Technologies de l'Information, vol E-25. pp 1–14

Diday E, Noirhomme M (2008) Symbolic data analysis and the SODAS software. Wiley, Chichester, p 457

Hosmer D, Lemeshow S (2000) Applied logistic regression. Wiley, New York. ISBN 0-471-61553-6

Lee J (1986) Insight on the use of multiple logistic regression analysis to estimate association between risk factor and disease occurrence. Int J Epidemiol 15:22–29

Schemann JF, Guinot C, Traore L, Sacko D, Zefack G, Dembele M, Diallo I, Malvy D (2007) Longitudinal evaluation of three azithromycin distribution strategies for treatment of trachoma in a sub-saharan African country, Mali. Acta Trop 101:40–53

Souza RMCR, Queiroz DCF, Cysneiros FJA (2011) Logistic regression based pattern classifiers for symbolic interval data. Pattern Anal Appl 14:273–282

WHO (1988) Programme for the prevention of blindness and deafness. In: Coding instructions for the WHO/PBL eye examination record (version iii). Tech. rep. World Health Organization, Genève