w

LAMSADE CEREMADE

DATA SCIENCE: NEW DATA, NEW PARADIGMS

From data to classes and classes as statistical units

RECHERCHE - FORMATION
UNIVERSITE PARIS-DAUPHINE
22-23 January 2018

Venue, Lieu des journées : Université Paris-Dauphine

Website : http://vladowiki.fmf.uni-lj.si/doku.php?id=sda:meet:pa18

La numérisation croissante de notre société alimente entre autres des bases de données ouvertes (« Open Data »), de taille grandissante (Big Data). Ces données sont souvent complexes (hétérogènes et multi-tables, munies de variables non appariées) mais peuvent être la source de création de valeur considérable pour la société à condition qu'elles soient exploitées avec des méthodes d'analyse adéquates.

Ces journées ont justement pour objectif de centrer cette fois le débat vers l'analyse de ces données en pensant en termes de classes. Les classes réduisent la taille des données et constituent souvent un pivot central incontournable de l'analyse. Ces classes obtenues par apprentissage non-supervisé permettent d'obtenir une vue concise et structurée modélisant les données, en apprentissage supervisé elles permettent de fournir des règles de décision efficaces.

Une troisième voie consiste à les considérer comme de nouvelles unités statistiques et à les décrire par des données symboliques (i.e. toute forme d'expression permettant de prendre en compte la variabilité interne des classes). On décrira ainsi les classes par des vecteurs de lois jointes ou marginales, d'intervalles, d'histogrammes (issus d'ondelettes, par exemple), de diagrammes de fréquence (d'utilisation courante dans les Instituts Nationaux de Statistique pour décrire des régions), de distributions, etc. . Cela permet de fusionner les données complexes et massives (en résolvant entre autres le problème des variables non appariées), pour pouvoir les étudier et les comprendre dans un cadre explicatif adéquat (i.e. à contrario des approches « boîte noire » de type « réseaux neuronaux »). L'analyse des données symboliques multidimensionnelles qui décrivent les classes peut aussi considérablement enrichir les interprétations classiques unidimensionnelles de ces classes.

L'objectif de ces Journées est de laisser la parole à des spécialistes de l'extraction de connaissances à partir de données de toutes sortes et de réfléchir ensemble aux orientations et tendances de la théorie et de la pratique de l'analyse de ces nouvelles données dans le contexte de la révolution numérique.

A "Data Scientist" is someone who is able to extract new knowledge from Standard, Big and Complex Data: unstructured data, unpaired samples, multi sources data (as mixture of numerical, textual, image, social networks data). The fusion of such data can be done into classes of row statistical units which are considered as new statistical units. Classes can be obtained by unsupervised learning giving a concise and structured view on the data or by supervised learning in order to produce efficient rules (as by deep learning). A third way is to consider classes as new statistical units described by vectors of intervals, probability distributions, weighted sequences, functions, and the like, in order to express the within-class variability. One of the advantages of this approach is that unstructured data and unpaired samples at the level of row units, become structured and paired at the classes' level.

The objective of this Workshop is to let speak the specialists of knowledge extraction from all sorts of data, and to think together about the orientations and trends of the theory and the practice of the analysis of these new data, in the context of the digital revolution.

THEMES

- **Theoretical foundation of classes and Symbolic Data**
- **Linear models for symbolic data**
- **Clustering for symbolic data**
- **Symbolic networks**
- **Dimensionality reduction**
- **Applications in socio-demography and ecology**

International Scientific Committee

L. Billard (UGA, USA)
P. Cazes (CEREMADE, University Paris-Dauphine)
D. Colazzo (LAMSADE, University Paris-Dauphine)
S. Pinson (LAMSADE, University Paris-Dauphine)
M. Ichino (College of Science and Engineering, Tokyo Denki University, Japan)
M. Noirhomme (Namur University, Belgium)
S. Sisson (UNSW Sydney, Australia)
H. Wang (School of Economics and Management, Beihang University, China)

Local Organizing Committee

P. Bertrand (CEREMADE, University Paris-Dauphine)
E. Diday (CEREMADE, University Paris-Dauphine)
W. Litwin (LAMSADE, University Paris-Dauphine)

Lecturers in the order of the program

G. Saporta (CNAM, Conservatoire National des Arts et Métiers, France) gilbert.saporta@cnam.fr

S. Sisson (UNSW Sydney, Australia) scott.sisson@unsw.edu.au

R. Emilion (MAPMO, Université d'Orléans, France) richard.emilion@gmail.com

E. Diday (CEREMADE, Université Paris-Dauphine) diday@ceremade.dauphine.fr

B. Beranger (UNSW Sydney, Australia) b.beranger@unsw.edu.au,

Z. Wang (SEM, Beihang University, China) wangzc1415@buaa.edu.cn

F. De Carvalho (CIn-UFPE, Recife, Brazil) fatc@cin.ufpe.br

T. Huang (SEM, Beihang University, China) 1521750380@qq.com

M. Nadif (MI, Université Paris-Descartes, France) mohamed.nadif@mi.parisdescartes.fr

O. Rodriguez (Costa-Rica University, Costa Rica) oldemar.rodriguez@gmail.com

A. Irpino (Univ. of Campania L. Vanvitelli, Caserta, Italy) antonio.irpino@gmail.com

W. Litwin (LAMSADE, Univ. Paris Dauphine, France) witold.litwin@gmail.com

R. Verde (Naples University, Italy) rosanna.verde2@gmail.com

V. Batagelj (FMF, Ljubljana, Slovenia) vladimir.batagelj@fmf.uni-lj.si

M. Malek (EISTI, Cergy-Pontoise, France) mma@eisti.eu

P. Brito (University of Porto, Portugal) mpbrito@fep.up.pt

F. Afonso (SYMBAD Symbolic Data Lab, Roissy-Pôle, France)
filipe.afonso@symbolicdata.com

C. Biernacki (Lille Univ., INRIA, France) christophe.biernacki@math.univ-lille1.fr

F. Lebaron (ENS, Paris-Saclay Cachan, France) flebaron@yahoo.fr

C. Toque (DGALN/SAGP/SDP/BCSI (Ministère de la transition écologique et solidaire. Ministère de la cohésion des territoires) carole.toque@developpement-durable.gouv.fr.

==========

PROGRAM

Monday, January 22nd

9h00 – 9h15 WELCOME ADDRESS

First Session

THEORETICAL FOUNDATION OF CLASSES AND SYMBOLIC DATA

9h15 – 9h45 G. Saporta (Conservatoire National des Arts et Métiers, France)

Paul Lazarsfeld and latent classes: some history

Résumé : L'analyse en classes latentes est à la classification ce que l'analyse en facteurs communs et spécifiques est à l'ACP. L'hypothèse fondamentale est que les p variables observées (en général qualitatives) sont indépendantes conditionnellement aux classes. Cas particulier d'un modèle de mélange, l'analyse en classes latentes est très utilisée dans les sciences sociales et on la trouve désormais dans de nombreux logiciels. L'exposé présentera également la personnalité exceptionnelle de son inventeur, Paul Lazarsfeld (1901- 1976) considéré comme le promoteur de la pensée mathématique en sociologie.

Abstract: Latent class analysis is a clustering method similar in its spirit to factor analysis when compared to PCA. The fundamental hypothesis is that the p manifest categorical variables are independent conditionally to the classes. LCA is a particular mixture model. LCA is still frequently used in social sciences and included in numerous softwares. The talk will also focus on the personality of its inventor Paul Lazarsfeld (1901-1976) considered by many as the father of mathematical thinking in sociology.

References

Bartholomew, D.J. et Knott, M. [1999], "Latent Variable Models and Factor Analysis", Arnold.

Bourdieu, P, [2004] "Esquisse pour une auto-analyse", Raisons d'agir,

Droesbeke, J.J. , Thomas-Agnan, C., Saporta, G. [2013]. "Modèles à variables latentes et modèles de mélange", EditionsTechnip,

Lautman, J. et Lécuyer, B.P. (éditeurs) [1998], "Paul Lazarsfeld (1901-1976), la sociologie de Vienne à New-York", L'Harmattan.

Lazarsfeld, P. F., et Henry, N.W. [1968]," Latent Structure Analysis", Houghton Mifflin.

9h45 – 10h15 S. Sisson, B. Beranger, J. Lin, T. Whitaker, X. Fan (UNSW Sydney, Australia)

A general framework for constructing symbolic likelihood functions

Abstract: Existing methods for constructing likelihood functions for symbolic data focus on building direct generative models for the symbol. This approach is ok, but the construction is inadequate if interest is in fitting likelihood models to the classical data underlying the symbols, while only observing the symbols themselves. In this talk I will outline a general method of fitting classical data likelihood functions given observed symbols, with of focus on random intervals and histograms. I will illustrate this through simulated examples, analyses of crop satellite data and mixture models.

References

Beranger, B., Lin, H. and Sisson, S.A. (2018). A general framework for symbolic likelihood functions. In preparation.

Le-Rademacher J. and L. Billard (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. Journal of Statistical Planning and Inference, 141, 1593-1602.

++

10h15 – 10h45 ***Coffee break***

++

10h45 – 11h15 R. Emilion (Université d'Orléans, France), E. Diday (Université Paris-Dauphine, France)

Likelihood on symbols: probabilistic setting and examples

Abstract: We first propose a probabilistic framework for properly defining a symbolic variable as a function of the conditional distribution of a description variable given a class variable. Then, we consider the problem of defining distributions on symbols of classes considered as statistical units or more simply to propose some likelihood functions for finite-dimensional symbols. The class label space should be an uncountable non-atomic probability space. In the parametric case, examples raise from finite mixtures of distributions on the simplex of probability vectors. In the nonparametric case Kernels derived from such distributions are proposed. Dirichlet Process Mixtures are examples in the infinite mixture case. When the description variable is a random vector, Dependent Dirichlet Processes provide some appropriate models. Such likelihood functions, that assign a numerical value to each symbol, can be of interest for many decision-making problems on symbols, including ranking, classification, outliers' detection, prediction, and so on.

References

Brito, P., Duarte Silva, A.P. (2012). Modelling Interval Data with Normal and Skew-Normal Distributions. Journal of Applied Statistics, Volume 39, Issue 1, 3-20.

Diday E. (2016). Thinking by classes in data science: symbolic data analysis. WIREs Computational Statistics Symbolic Data Analysis, Volume 8, September/October 2016 ©2016 Wiley Periodicals, Inc. 191.

Soule A, Salamatian K, Taft N., Emilion R, Papagiannaki K. (2004). Flow classification by histograms. In: Proceedings of Sigmetrics'04, New York.

Emilion R. (2012) Unsupervised classification of objects described by nonparametric distributions. Statistical Analysis and Data Mining, Vol.5, 5, 388–398.

Le-Rademacher J. and L. Billard (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. Journal of Statistical Planning and Inference, 141, 1593-1602.

11h15 – 11h45 E. Diday (CEREMADE, Université Paris-Dauphine, France)

Some recall on SDA and Basic theory for placing in order classes of complex data and their symbolic descriptive variables

Abstract: First we recall some basic notions in Data Science: what are complex data? What are classes? Which kind of class variability can be considered? What are classes of complex data? Then we define "symbolic variables" and "symbolic data tables" and we give some advantages of such kind of class description. Often in practice the classes are given. When they are not given, clustering can be used to build the classes which yields to a symbolic data table. The quality of the obtained clusters can be measured by different criteria (as the entropy) of the obtained symbolic data table. Finally we suggest a theoretical framework for Symbolic Data Analysis based on three random variables defined on the ground population and we illustrate it by a way for placing in order classes of complex data and their symbolic descriptive variables by a symbolic generalization of the Tf-Idf much used in text mining.

References

Bock H., Diday E. (2000). Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer Verlag, Heidelberg, 425 pages, ISBN 3-540-66619-2.

L. Billard, E. Diday (2003). From the statistics of data to the statistic of knowledge: Symbolic Data Analysis. JASA . Journal of the American Statistical Association. Juin, Vol. 98, N° 462

Brito P, Noirhomme-Fraiture M, Arroyo J.(2015) Special issue on symbolic data analysis. Editorial. Adv. Data Anal Classif 2015, 9:1–4.

Diday, E. (2016) Thinking by classes in data science: symbolic data analysis. WIREs Computational Statistics Symbolic Data Analysis, Volume 8, September/October 2016 ©2016 Wiley Periodicals, Inc. 191.

Emilion R. Unsupervised classification of objects described by nonparametric distributions. *Stat Anal Data Mining* 2012, 388–398.

11h45 – 12h15 B. Beranger, T. Whitaker, S. Sisson (UNSW Sydney, Australia)

Extreme value analysis using symbolic data

Abstract: The analysis of Spatial Extremes has been given a growing interest over the last decade in a broad range of areas and especially in climatology with the goal to better understand the behavior of events such as floods, heat waves or storms. Max-stable processes are a convenient and widely used tool to model such phenomena. In recent years, composite likelihood methods have appeared to bypass the intractability of the multivariate density function of such processes. However, the computational cost of these methods explodes as the number of temporal observations gets large. This is even more noticeable when working with a large number of spatial locations across a study region. To bypass this issue we introduce a symbolic data analysis (SDA) based approach which consists in aggregating data into histograms leading to a reduction of the complexity of the data. A symbolic version of the composite likelihood approach where observations are multivariate histogram-valued is provided and the classical results from Padoan, et al. (2010) are shown to be recovered as a limiting case. The performance of our procedure in terms of inferential and computational efficiency is studied in an extensive simulation study and the impact of coarsening the data and the design of the symbols (histograms) is discussed. Finally, the utility of the method is illustrated through the analysis of fortnightly maximum temperatures at 105 locations across Australia using historical data and simulated data from two climate models.

References

Padoan, S.A., Ribatet, M. and Sisson, S.A. (2010), Likelihood-Based Inference for Max-Stable Processes. Journal of the American Statistical Association 105, 263-277.

Whitaker, T., Beranger, B. and Sisson, S.A. (2018). A composite likelihood based approach for max-stable processes using histogram-valued variables. In preparation.

Beranger, B., Lin, H. and Sisson, S.A. (2018). A general framework for symbolic likelihood functions. In preparation.

++

12h15 – 14h00 ***LUNCH***

++

Second session

LINEAR MODELS

14h00 – 14h30 Z. Wang, T. Huang (School of Economics and Management, Beihang University, China)

Linear Mixed Effects Models for Longitudinal Compositional Data

Abstract: Compositional data, a type of symbolic data that expresses the structure information that parts organize a whole, occurs regularly and is of great practical importance in economics, for instance, the investment, employment and industrial structures. In many cases, measurements of structural economic indicators are taken from individuals (regions or countries) through time, which may probably cause the heterogeneity and the dependency in the population. These compositional data with the feature of longitudinal data are referred to as longitudinal compositional data. We investigate linear mixed effects models for longitudinal compositional data with both dependent and independent variables of compositional structure. Parameters of both fixed effects and random effects are estimated by the EM algorithm, which generalizes the theoretical framework of linear mixed effects models to multivariate compositional data in symbolic data analysis.

References

J. Aitchison (1986): The Statistical Analysis of Compositional Data. Springer Netherlands.

L. Nan, N. Lange & D. Stram (1987): Maximum likelihood computations with repeated measures: application of the EM algorithm. Publications of the American Statistical Association, 82(397), 97-105.

H. Wang, L. Shangguan, J. Wu & R. Guan (2013): Multiple linear regression modeling for compositional data. Neurocomputing, 122(122), 490-500.

J. Chen, X. Zhang & S. Li (2016): Multiple linear regression with compositional response and covariates. Journal of Applied Statistics, 1-16.

14h30 – 15h00 F. De Carvalho (CIn-UFPE, Brazil)

A kernel robust regression model for interval-valued variables

Abstract: The presence of outliers is very common in regression problems and the use of robust regression methods is strongly recommended such that the bad fitted observations does not affect the parameter estimates of the model. Interval-valued variables are becoming common in data analysis problems since this type of data represents either the uncertainty existing in an error measurement or the natural variability present in the data.
Regarding the presence of outliers in interval-valued data sets, few robust regression methods have been proposed in literature. This paper introduces a new robust regression method for interval-valued variables that penalizes the presence of outliers in the midpoints and/or in the ranges of interval-valued observations through the use of exponential-type kernel functions. Thus, the weight given to the midpoint and range of each interval-valued observation is updated at each iteration in order to optimize a suitable objective function. The parameter estimation algorithm converges with a low computational cost. A comparative study between the proposed method against some previous robust regression approaches for interval-valued variables is also considered. The performance of these methods are evaluated based on the bias and mean squared error (MSE) of the parameter estimates for the midpoints and ranges of the intervals, considering synthetic data sets with X-space outliers, Y-space outliers and leverage points, different sample sizes and percentage of outliers in a Monte Carlo framework. The results suggest that the proposed approach presents a competitive performance (or best), in comparison with the previous approaches, on interval-valued outliers scenarios that are comparable to those found in practices. Applications to real
interval-valued data sets corroborates the usefulness of the proposed method.

References

H. H. Bock and E. Diday, editors. Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data. SpringerVerlag, Heidelberg, 2000.

F. A. T. De Carvalho, E. A. Lima Neto, and M. R. P. Ferreira. A robust 819 regression method based on exponential-type kernel functions. Neurocomputing, 234 (2017):58–74.

M. A. O. Domingues, R. M. C. R. de Souza, and F. J. A. Cysneiros. A robust method for linear regression of symbolic interval data. Pattern Recognition Letters, 31(2010):1991–1996.

R. A. A. Fagundes, R. M. C. R. de Souza, and F. J. A. Cysneiros. Robust regression with application to symbolic interval data. Engineering Applications of Artificial Intelligence, 26 (2013):563–573.

R. A. A. Fagundes, R. M. C. R. de Souza, and Y. M. G. Soares. Quantile regression of interval-valued data. In Proceedings of 23rd International Conference on Pattern Recognition, 2016.

P. J. Huber. Robust Statistics. John Wiley and Sons Inc., New York, 1981.

E. A. Lima Neto and F. A. T. De Carvalho. Centre and range method for fitting a linear regression model to symbolic interval data. Computational Statistics and Data Analysis, 52 (2008):1500–1515.

R. A. Maronna, R. D. Martin, and V. J. Yohai. Robust Statistics: Theory and Methods. John Wiley and Sons Inc., Chichester, 2006.

P. J. Rousseeuw and A. M. Leroy. Robust Regression and Outlier Detection. John Wiley and Sons Inc., New York, 1987.

15h00 – 15h30 H. Wang, T. Huang, S. Wang (School of Economics and Management, Beihang University, China)

Spatial functional Linear Model and Estimation Method

Abstract: The classical functional linear regression model (FLM) and its extensions, which are developed under the assumption that all the individuals are mutually independent, have been well studied and used by many researchers. However, this independent assumption may be violated in practice, especially when we collect data with network structure from scientific disciplines, such as marketing, sociology and spatial economics. Yet relatively few works are available for FLM with network structure. We propose a novel spatial functional linear model (SFLM), incorporating a spatial autoregressive parameter and a spatial weight matrix in FLM to accommodate spatial dependence among individuals. The proposed model is more flexible as it takes advantages of FLM in dealing with high dimensional covariates and spatial autoregressive model (SAR model) in capturing network dependence. We develop an estimation method based on functional principle component analysis (FPCA) and maximum likelihood estimation. The simulation studies show that our method performs as well as FPCA-based method for FLM when there is no network structure; while outperforms the latter when there exists network structure. A real dataset of weather data is also employed to demonstrate the utility of SFLM.

References

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. The Annals of Statistics, 34(5), 2159–2179.

Case, A. C. (1991). Spatial patterns in household demand. Econometrica, 59(4), 953–965.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. Annals of Statistics, 35(1), 70–91.

James, G. M. (2002). Generalized linear models with functional predictors. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 411–432.

Qu, X. and Lee, L. F. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. Journal of Econometrics, 184(2), 209–232.

++

15h30 – 16h00 ***Coffee break***

++

Third session

DIMENSIONALITY REDUCTION

16h00 – 16h30 L. Labiod, M. Nadif (Université Paris-Descartes, France)

Simultaneous learning for clustering and dimensionality reduction

Abstract: Spectral clustering has the advantage of requiring weak assumptions regarding the shapes of clusters. It is therefore effective to a wide variety of data types and similarity functions. However, classical spectral methods such as Ratio Cut [1] and Normalized Cut [2, 3] use generally k-means to perform the clustering on the relaxed continuous spectral vectors in order to obtain the final clusters. The disadvantage of this approach is that it consists in optimizing two different objective functions. Hence, spectral low-dimensional embedding and clustering are successively, and not simultaneously, used. This leads, that certain obtained continuous low-dimensional embedding can deviate far from a good clustering solution. To overcome the disadvantages of such an approach in two separate steps, in this talk, we aim to combine simultaneously the spectral clustering and the dimensionality reduction. We propose a novel framework, referred to as Joint Spectral Dimensionality Reduction and Clustering (JSDRC) allowing to alternate both tasks iteratively.

References

[1] L.W. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. IEEE Transactions on CAD, 11(9):1074–1085, 1992.

[2] J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 22 (8):888–905, 2000.

[3] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In Advances in Neural Information Processing Systems (NIPS), pages 849–856, 2001.

16h30 – 17h00 O. Rodriguez (Costa-Rica University)

Optimized dimensionality reduction methods and symbolic principal component for interval-valued variables

Abstract: In the last two decades, principal component analysis was adapted for symbolic data, first in the context of interval-valued data. A number of approaches have been proposed. In Diday E. (1997) and Billard L. (2011), the authors proposed the centers method and the vertices methods to extend the well-known principal components analysis method to a particular kind of symbolic objects characterized by multi–valued variables of interval type. Two methods were proposed, a vertices method which uses all the vertices of the observation's hypercube, and a centers method which uses the centroid values.

This paper aims to improve the centers method applying an optimization algorithms in which instead of projecting the centroid value we look for the best point to project in supplementary the vertices. The best point in the sense that it minimizes the distance of the supplementary individuals to that point or the point that generates an principal components analysis with the best inertia in the first components and then from this projecting the vertices as supplementary elements. We obtain interval–valued symbolic principal components which recapture better the internal variation of the observations or maximizes the correlation measures between these principal components and the random variables and/or the observations themselves.

Besides, the reader may use all the methods presented herein and verify the results using the RSDA package written in R language, that can be downloaded and installed directly from CRAN , see Rodrıguez, O. (2017).

References

Billard, L. & Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley & Sons Ltd, United Kingdom.

Cazes P., Chouakria A., Diday E. et Schektman Y. (1997). Extension de l'analyse en composantes principales a des données de type intervalle. Rev. Statistique Appliquée, Vol. XLV 3 , 5–24.

Hastie,T. (1984). Principal Curves and Surface. Ph.D Thesis Stanford University.

Hastie,T. & Stuetzle, W. (1989). Principal Curves. Journal of the American Statistical Association, Vol. 84 406, 502–516.

Rodriguez, O. with contributions from Olger Calderon and Roberto Zuniga (2017). RSDA - R to Symbolic Data Analysis. R package version 2.0. http://CRAN.R-project.org/package=RSDA.

17h00 – 17h30 A. Irpino, (Dept. of Mathematics and Physics, University of Campania L. Vanvitelli, Caserta, Italy), J. Arroyo Gallardo (Universidad Complutense de Madrid, Spain)

Dimension reduction technique for histogram variables: an application on a Human Activity Recognition dataset

Abstract: Distributional Data Analysis (DDA) allows the description of statistical units by empirical distribution of values observed for numerical variables. In this talk, we will show how to apply Multiple Factor Analysis (MFA) to observations described by distributional variables. In particular, we will present some new visualization tools called "Spanish-fan" plots. The MFA is implemented in the R package named "HistDAWass" that is freely available from CRAN.

We will illustrate the use of MFA as an exploratory analysis tool. The MFA analyzes a sensor data set that describes 19 physical activities performed by 8 people. In this context, we will temporally aggregate the windows of sensor data using distributions instead if the standard approach of computing statistical measures describing the window data. Given a particular activity, MFA will serve us to discover its main sources of variability in terms of sensor distributional variables and to analyze inter-subject and intra-subject variability when performing that activity.

References

Altun, K., Barshan, B. and Tunçel, O. (2010) Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recognition, 43(10) pp 3605-3620.

Bock H.-H., Diday E. (2000) Analysis of symbolic data, exploratory methods for extracting statistical information from complex data. Studies in Classification, Data Analysis and Knowledge Organisation. Springer-Verlag, Berlin.

Verde R., Irpino A., Balzanella A. (2016). Dimension Reduction Techniques for Distributional Symbolic Data. IEEE Transactions on Cybernetics, vol. 46, pp. 344-355, ISSN: 2168-2267, doi: 10.1109/TCYB.2015.2389653

Verde R., Irpino A. (to appear in 2018), Multiple Factor Analysis of Distributional Data, Statistica Applicata - Italian Journal of Applied Statistics Vol. 29 (2-3).

17h30 – 18h00 W. Litwin (LAMSADE, Univ. Paris-Dauphine, France), S. Jojoda (G. Mason Univ., USA), Th. Schwarz (Marquette Univ., USA)

Trusted cloud SQL DBS with On-the-fly AES Decryption/ Encryption for Big SQL Data Bases

Abstract: During latest decades, a number of encryption tools for Big SQL DBs, especially in a cloud, have been proposed. Unfortunately, none of them supports aggregate value expressions in queries that DB data analysis usually needs. We have proposed a new software-only tool aiming on that goal: Trusted Cloud Database System. It securely manages client-side encrypted cloud DBs. Queries may include encryption keys. The DBS decrypts/encrypts the data on-the-fly at the cloud. Plaintext is only in protected run-time variables. Stored data are by default probabilistically encrypted through AES. Any SQL queries are feasible, with negligible processing overhead and practical storage overhead. This is major advance over all the current alternative research proposals. Including those proposing the hardware add-on to the cloud node, usually called Trusted Processing Module. We detail capabilities of a trusted DBS. We adapt SQL to client-side key management. Queries may remain usually almost as non-procedural as now. A prototype implementation appears easy.

References

Holland, David A., Ada T. Lim, and Margo I. Seltzer. 2005. An architecture a day keeps the hacker away. 2004 Workshop on Architectural Support for Security and Anti-Virus. Boston, MA. Special issue.

Jajodia, S. Litwin, W. Schwarz, Th. Numerical SQL Value Expressions over Encrypted Cloud Databases. 8th Intl. Conf. on Data Management in Cloud, Grid and P2P Systems (Globe 2015). In DEXA 2015. Springer, 2015.

Jajodia, S. Litwin, W. Schwarz, Th. On-the fly AES256 Decryption/Encryption for Cloud SQL Databases. Position Paper. BDMICS 2016, Porto (Sept. 2016), 5p, IEEE, publ., to app. (a) Extended Preliminary Version: LAMSADE Res. Rep. June 2015, 13p.

Jajodia & al. eds. Moving Target Defense. Advances in Information Security. Vol 1 & 2. Springer, 2011-3.

SiSoftware AES256 Benchmark. 2015. http://www.sisoftware.co.uk/?d=qa&f=cpu_vs_gpu_crypto&l=en&a=

Grant. Hardware AES Showdown - VIA Padlock vs Intel AES-NI vs AMD Hexacore. 2011. http://www.grantmcwilliams.com/tech/technology/387-hardware-aesshowdown-via-padlock-vs-intel-aes-ni-vs-amd-hexacore

==========

Tuesday, January 23rd

First Session

CLUSTERING

9h00 – 9h30 R. Verde, A. Irpino (Naples University, Italy)

Distributional data clustering and visualization for official statistics

Abstract: Distributional Data Analysis (DDA) is a new field of research related to Symbolic Data Analysis (Bock, Diday, 2000). The statistical units, or objects, are described by empirical distribution of values observed for numerical variables. In some cases, the distribution are synthesis or aggregated data, in order to preserve the confidentiality of the information. Many Official Statistics are in form of "histogram data", like the "Financial Characteristics for Housing Units With a Mortgage" data of the ACS American Community Survey 2015, of the UScensus Bureau. In such a case, an interval on proportions, is furnished as a sort of "confidence interval".

Starting from previous techniques, proposed for analyzing DDA (e.g., clustering, principal component analysis), we introduce a further information in the data, given by the intervals of proportions. That leads to consider that each statistical unit is described by a set of distributions for each variable, expressed by the convex combination of the interval values on proportions.

A hierarchical clustering analysis allows to discover classes of objects according to the characteristics of the distribution of the values, taking also into account their variability, expressed by a set of distributions, carried out as convex combination of the bounded distributions. A new visualization tool of the characteristics of the distributions, in a reduced subspace, has been furnished by a Principal Component Analysis (PCA) technique for distributional data (Verde, Irpino, 2016). An extension of PCA to this kind of Official Statistics allows a visualization of the variability of the data related to the several components of DDA.

Finally, DDA presents strong potentiality in the analysis of Big Data, and all that concerns the Volume of values and the Variability.

References

Bock H.-H., Diday E. (2000) Analysis of symbolic data, exploratory methods for extracting statistical information from complex data. Studies in Classification, Data Analysis and Knowledge Organisation. Springer-Verlag, Berlin.

Irpino A., Verde R. (2006) A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Batanjeli V, Bock HH, Ferligoj A, Ziberna A (eds) Data science and classification, IFCS 2006. Springer, Berlin, pp 185–192.

Verde R., Irpino A., Balzanella A. (2016). Dimension Reduction Techniques for Distributional Symbolic Data. IEEE Transactions on Cybernetics, vol. 46, pp. 344-355, ISSN: 2168-2267, doi: 10.1109/TCYB.2015.2389653

9h30 – 10h00 F. De Carvalho (CIn-UFPE, Brazil)

Gaussian Kernel C-Means Clustering Algorithms with Automated Computation of Bandwidth Parameters

Abstract: Conventional Gaussian kernel c-means clustering algorithms are widely used in applications. However, Gaussian kernel functions have an important parameter, the bandwidth parameter that needs to be tuned. Usually this parameter is tuned once and for all, and it is the same for all the variables. In this way, implicitly the variables are equally rescaled and therefore, they have the same importance on the clustering task. This paper presents Gaussian kernel C-Means clustering algorithms with automated computation of bandwidth parameters. In these kernel-based clustering algorithms, the bandwidth parameters change at each iteration of the algorithm, they are different from one variable to another, and they can be different from one cluster to another. Because each variable is rescaled differently according to its own hyper-parameter, these algorithms are able to select the important variables in the clustering process. Examples with data sets of the UCI machine learning repository corroborate the usefulness of the proposed algorithms.

References

A. K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognition Letters, 31 (2010): 651-666.

M. Girolami, Mercer kernel-based clustering in feature space, IEEE Trans. Neural Netw. 13 (2002):780-784.

F. Camastra, A. Verri, A novel kernel method for clustering, IEEE Trans. Neural Netw. 27 (2005): 801-804.

M. Filippone, F. Camastra, F. Masulli, S. Rovetta, A survey of kernel and spectral methods for clustering, Pattern Recognition 41 (2008): 176-190.

B. Caputo, K. Sim, F. Furesjo, and A. Smola. Appearence-based object recognition using svms: which kernel should i use? In Proceedings of NIPS Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer vision, 2002.

E. Diday and G. Govaert, Classification automatique avec distances adaptatives, R.A.I.R.O. Informatique Computer Science, 11 (1977): 329-349.

D. S. Modha and W. S. Spangler, Feature weighting in k-means clustering, Machine Learning, 52 (2003): 217-237.

J.Z. Huang, M.K. Ng, H. Rong, and Z. Li, Automated variable weighting in k-means type clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (2005): 657-668.

++

10h00 – 10h30 ***Coffee break***

++

Second Session

NETWORKS

10h30 – 11h00 Vladimir Batagelj (IMFM Ljubljana, UP IAM Koper), N.Kejžar (University of Ljubljana, MF), S. Korenjak-Černe (University of Ljubljana, MF).

Agglomerative clustering with relational constraints of large symbolic data sets

Abstract: Agglomerative clustering algorithms for solving clustering problems with relational constraints were proposed already in eighties (Ferligoj and Batagelj, 1982, 1983). A problem with these algorithms is their scalability. Because they are based on a dissimilarity matrix they can be applied to data sets with up to some ten thousands of units.

In this contribution we discuss two approaches for agglomerative clustering of large symbolic data sets. Both are based on the idea to compute the dissimilarities only between the related (with constraints) units and the assumption that the constraints network is sparse. The first approach is based on the introduction of new dissimilarities between clusters (Bodlaj and Batagelj, 2015, Batagelj et al., 2014), the second on "classical" dissimilarities between cluster representatives (Batagelj, 1988). Both approaches were implemented in R and will be illustrated on some real-life symbolic data sets (Korenjak-Černe et al., 2015).

References

Batagelj, V. (1988). Generalized Ward and Related Clustering Problems. Classification and Related Methods of Data Analysis. H.H. Bock (ed). North-Holland, Amsterdam, p. 67-74.

Batagelj, V., Doreian, P., Ferligoj, A., Kejžar, N. (2014). Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley.

Bodlaj, J., Batagelj, V. (2015). Hierarchical link clustering algorithm in networks. Phys. Rev. E 91,062814 .

Ferligoj, A., Batagelj, V. (1982). Clustering with relational constraint. Psychometrika 47 (4): 413-426.

Ferligoj, A., Batagelj, V. (1983). Some Types of Clustering with Relational Constraints. Psychometrika 48 (4): 541-552.

Korenjak-Černe, S., Kejžar, N., Batagelj, V. (2015). A weighted clustering of population pyramids for the world's countries, 1996, 2001, 2006. Population Studies: A Journal of Demography 69,(1): 105-120.

11h00 – 11h30 M. Malek (EISTI, Cergy-Pontoise, France)

Analysis of complex networks awarded multi-layer

Abstract: We present in this talk a preliminary study that aims to confront and assess private data for a given domain in the context of open data. In our case, open data is represented by a multi-layer complex network while private data correspond to a subset of nodes in the different layer. Our study is carried on two levels: the intra-layer level and the inter-layer one. The induced graph elaborated from the private data is analysed and immersed in the whole network. We revise and compare nodes local measures as centralities, we compare global measures as density and we use also the communities structure in each layer to study the distribution and the connectivity of private nodes compared to the whole layer. We generalise the notion of egocentric network that is defined around a given node and propose an egocentric network around the private data' induced graph. We use this egocentric network in order to evaluate the connectivity strength between the different layers of private data in comparison to the whole network. This work makes part of the actual realisation of a biological application in the context of the Project Blizaar (ANR international).

References

Jean-Philippe Attal, Maria Malek, Marc Zolghadri Parallel and distributed core label propagation with graph coloring, Concurrency and computation : practice and experience, Wiley, 2017.

Djemili S., Marinica C., Malek M., Kotzinos D. Personal Networks of Scientific Collaborators: A Large Scale Experimental Analysis of Their Evolution. Communications in Computer and Information Science, vol 760. Springer, 2017.

M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, Multilayer networks, Journal of Complex Network, vol. 2, no. 3, pp. 203–271, 2014.

Third Session

MULTIBLOCK and TIME SERIES

11h30 – 12h00 V. Cariou (StatSC, Oniris, INRA, 44322, Nantes, France)

Supervised multiblock modelling with P-ComDim. Applications in sensometrics and chemometric.

Abstract: In many areas, the coupling of different kinds of measurements generates a large amount of variables structured into meaningful blocks for the characterization of the same set of samples. To analyze such kind of data, multiblock methods were proposed both in an unsupervised and supervised case. These methods aim at extracting block and global components which highlight the main dimensions that underlie the data. In the supervised case, where a dataset ***Y*** is predicted from *K* other datasets, the multiblock method P-ComDim has recently been proposed. After a presentation of P-ComDim and a comparison with MB-PLS regression, the method will be illustrated on the basis of case studies pertaining to sensometrics and chemometrics.

References

Skov, T., Honoré, A. H., Jensen, H. M., Næs, T. & Engelsen, S. B. Chemometrics in foodomics: handling data structures from multiple analytical platforms. *TrAC Trends in Analytical Chemistry*, 60, 71-79, 2014.

Wangen, L. E., & Kowalski, B. R. A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of chemometrics*, *3*(1), 3-20, 1989.

El Ghaziri, A., Cariou, V., Rutledge, D. N. & Qannari, E. M. Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of (K+ 1) datasets. *Journal of Chemometrics*, 30(8), 420-429, 2016.

Cariou, V., Qannari, E. M., Rutledge, D. N. & Vigneau, E. ComDim: from multiblock data analysis to path modeling. *Food Quality And Preference*, https://doi.org/10.1016/j.foodqual.2017.02.012, 2017.

12h00 – 12h30 A. Maharaj (Monash Univ., Melbourne, Australia), P. Teles (University of Porto, Portugal), P. Brito (University of Porto, Portugal)

Clustering of Interval Time Series

Abstract: An interval time series (ITS) $[X_t]$ is a sequence of intervals observed in successive instants in time, where each interval is represented by its lower and upper bounds : $[X_{1,L}; X_{1,U}]$, $[X_{2,L}; X_{2,U}]$,…, $[X_{T,L}; X_{T,U}]$; alternatively, an ITS may also be represented by centres and radius, $[X]_t = X_{t,C}, X_{t,R}$, with $X_{t,C} = (X_{t,L} + X_{t,L})/2$ and $X_{t,R} = (X_{t,U} - X_{t,L})/2$. In this study, we focus on clustering of a set of ITS. In a first approach we compare time-series based on point-to-point comparisons. For each pair of series, we compute the distance between the observed intervals at each time point t =1,…, T, which are then averaged over the full interval time series to obtain a value comparing the two series; a distance matrix consisting of these measures is used as input to hierarchical and dynamical clustering methods. Another approach involves using time domain features and wavelet features of the radius and centres series as variables for the clustering. A further new technique involves fitting space-time models to each of the ITS under consideration and using the parameter estimates of the fitted models as inputs into clustering methods. Finally, clustering may also be performed using distances based on (interval) auto-correlation measures. Simulation studies allow evaluating the performance of the alternative techniques and an application is made to sea level ITS.

References

Arroyo J., Maté C. (2009). Forecasting histogram time series with k-nearest neighbours' methods. International Journal of Forecasting 25(1), 192-207.

De Carvalho F.A.T., Lechevallier Y., Verde R. (2008). Clustering methods in symbolic data analysis. In: Diday E., Noirhomme-Fraiture M. (eds) Symbolic Data Analysis and the SODAS Software, Wiley, Chichester, 182-203.

Dias S., Brito P. (2017). Off_ the beaten track: A new linear model for interval data. European Journal of Operational Research 258(3), 1118-1130.

Genolini C., Falissard B. (2010). Kml: k-means for longitudinal data. Computational Statistics 25, 317-328.

Maharaj, E.A., Teles, P., Brito, P. (2017). Clustering of Interval Time Series. Proc. World Statistics Conference, ISI 2017, Marrakesh, Morocco.

Verde R., Irpino A. (2007). Dynamic clustering of histogram data: Using the right metric. In: Brito P., Bertrand P., Cucumel G., De Carvalho F.A.T. (eds), Selected Contributions in Data Analysis and Classification, Springer, Heidelberg, 123-134.

++

12h30 – 14h00 ***LUNCH***

++

Fourth Session

SOFTWARE AND BIG DATA MANAGEMENT

14h00 – 14h30 O. Rodriguez (Costa-Rica University)

New methods and applications with the R package for Symbolic Data Analysis - RSDA

Abstract: This package aims at executing some models on Symbolic Data Analysis. Symbolic Data Analysis was proposed by the professor E. Diday in 1987 in his paper "Introduction à l' approche symbolique en Analyse des Données" - Premières Journées Symbolique-Numérique, Université Paris IX Dauphine, décembre 1987. A very good reference to symbolic data analysis can be found in "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis" of L. Billard and E. Diday that is the journal American Statistical Association Journal of the American Statistical Association June 2003, Vol. 98. The main purpose of Symbolic Data Analysis is to substitute a set of rows (cases) in a data table for a concept (second order statistical unit). For example, all of the transactions performed by one person (or any object) for a single "transaction" that summarizes all the original ones (Symbolic-Object) so that millions of transactions could be summarized in only one that keeps the customary behavior of the person. This is achieved thanks to the fact that the new transaction will have in its fields, not only numbers (like current transactions), but can also have objects such as intervals, histograms, or rules. This representation of an object as a conjunction of properties fits within a data analytic framework concerning symbolic data and symbolic objects, which has proven useful in dealing with big databases. In RSDA version 2.3, methods like centers interval principal components analysis, histogram principal components analysis, multi-valued correspondence analysis, interval multi-dimensional scaling (INTERSCAL), symbolic hierarchical clustering, CM, CRM, Lasso, Ridge and Elastic Net Linear regression model to interval variables have been implemented. This new version also includes new features to manipulate symbolic data through a new data structure that implements Symbolic Data Frames and methods for converting SODAS and XML SODAS files to RSDA files. This version also include Optimized Center Method and Principal Surfaces to Principal Component Analysis and new plot graphics like radar charts to interval variables. We will also present the package RSDA-PLUS that is a commercial version of RSDA that include some bank applications of Symbolic Data Analysis, for example, to detect money laundering.

References

Billard, L. & Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons Ltd, United Kingdom.

Rodríguez, O. with contributions from Olger Calderon and Roberto Zuñiga (2017). RSDA - R to Symbolic Data Analysis. R package version 2.3. http://CRAN.R-project.org/package=RSDA.

14h30 – 15h00 C. Biernacki (Lille University, INRIA, France)

MASSICCC: A SaaS Platform for Clustering and Co-Clustering of Mixed Data

Abstract: The ``Big Data'' paradigm involves large and complex data sets where the clustering task plays a central role for data exploration. For this purpose, model-based clustering and model-based co-clustering have demonstrated many theoretical and practical successes in a various number of fields. In this context, user-friendly software are essential for speeding up diffusion of such academic advance inside the applicative world. MASSICCC (massive clustering in cloud computing) is a user-friendly SaaS platform which hosts three software specialized in different clustering tasks and written in C++. This platform allows to manipulate complex data with very light computing tools (as a smartphone), including also some dynamical graphical outputs. However, it offers also the possibility to export the results into a R data format for further more expert tasks. The three embedded software are Mixmod, Mixtcomp and Blockcluster. Mixmod (Lebret et al. 2015) is dedicated to clustering of continuous, categorical and a mixing of continuous and categorical data. Mixtcomp (Biernacki 2015) adds the possibility to cluster totally mixed data (continuous, categorical, count, ordinal, rank, functional), potentially including missing or partially missing (like interval) data. Blockcluster (Bhatia et al. 2017) is dedicated to co-clustering of large data sets composed of different kinds of data like continuous, categorical and count ones.
MASSICCC is freely available at https://massiccc.lille.inria.fr

References

P. Bhatia, S. Iovleff & G. Govaert (2017). Blockcluster: An R Package for Model-Based Co-Clustering. Journal of Statistical Software, 76:9.

C. Biernacki (2015). Model-based clustering with mixed/missing data using the new software MixtComp. 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2015), University of London, UK, 12-14 December.

R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux & G. Govaert (2015). Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. Journal of Statistical Software, 67:6.

15h00 – 15h30 F. Afonso, (SYMBAD – Symbolic Data Lab, Roissy CDG, France)

The SYR software for symbolic data analysis of complex data and its industrial application.

Abstract: We present the main evolutions of the SYR software for Symbolic Data Analysis (SDA) of Complex Data. Its purpose is to merge and aggregate up data from multiple data files of different units and different variables into a single data table describing classes. The data can be heterogeneous (quantitative + qualitative + temporal data + etc.) and multi-source (one or more databases + Open data + sensor readings, etc.). The classes are then described by standard categorical or numerical variables, as well as by interval variables, multi-valued categorical or numerical variables, bar-chart and histogram-valued variables. These new kinds of variables allow keeping the internal variation of each class. It is then possible to find new correlations between data from different databases. The software is presented through industrial, social, demographic and medical applications where it has shown to be useful for analyzing their particularly complex data. Finally, further researches and development of the software are discussed.

References

Afonso, F., Diday, E., Toque, C. (janv. 2018). Data Science par Analyse des Données Symboliques. Technip. 448 pages.

F.Afonso, S. Laaksonen (2015) : Analyzing European Social Survey data using symbolic data methods and SYROKKO software, SDAV 2015, vol. RNTI-E-29, pp.89-100.

Courtois, A., Genest, Y., Vacqué, A., Afonso, F. (2012). In-service inspection of cooling towers, in Nuclear Engineering International, NEI.

Diday, E., Afonso, F., Haddad, R. (2013). The symbolic data analysis paradigm, discriminate discretization and financial application. HDSDA 2013. Revue des Nouvelles Technologies de l'Information, vol. RNTI-E-25, p. 1-14.

Guinot, C., Malvy, D., Schemann, J-F., Afonso, F., Haddad, R., Diday, E. (2015). Strategies evaluation in environmental conditions by symbolic data analysis: application in medicine and epidemiology to trachoma. ADAC (Advances in Data Analysis and Classification). March 2015, vol. 9, Issue 1, p. 107-119.

Nuemi, G., Afonso, F., Roussot, A., Billard, L., Cottenet, J., Combier, E., Diday, E., Quantin, C. (2013). Classification of hospital pathways in the management of cancer: application to lung cancer in the region of burgundy, Cancer Epidemiology journal, Elsevier.

++

15h30 – 16h00 ***Coffee break***

++

Fifth Session

APPLICATIONS

16h00 – 16h30 F. Lebaron (ENS, Paris-Saclay Cachan, France)

Classes of living conditions and social classes in Europe

Résumé : Les données européennes EU-SILC constituent une source importante d'informations massives sur les conditions de vie des individus et ménages au sein de l'Union Européenne. Elles permettent notamment de construire des indicateurs de pauvreté non-monétaire («privation matérielle sévère») et d'étudier leurs variations dans l'espace et le temps. Dans cette présentation, on étudiera les données EU-SILC les plus récentes (2016), dans la perspective de construire des classes d'individus relatives à leurs conditions de vie, ainsi qu'à leurs caractéristiques socio-démographiques, et l'on confrontera les résultats de ces diverses classifications aux catégories socio-économiques et aux divisions géographiques entre pays et à l'intérieur de certains pays. Différentes méthodes de classification seront mobilisées dans cette perspective exploratoire et comparative. Il s'agit ainsi d'étudier la solidité relative des différentes catégorisations usuellement employées à propos de l'Europe, en confrontant en particulier les catégorisations géographiques et les catégorisations socio-économiques. L'hypothèse de l'étude est celle de la complémentarité de ces deux types de catégorisations, et, du point de vue méthodologique, celle des différentes techniques de classification utilisées.

References

Pierre Blavier, Les manifestations socio-économiques du chômage de masse et les réaménagements des budgets de ménage pour y faire face. Le cas de la Grande Récession espagnole (2008-2015), thèse de doctorat de l'école d'économie de Paris, sous la direction de Jérôme Bourdieu et Frédéric Lebaron, novembre 2017.

Frédéric Lebaron, Pierre Blavier, "Classes et nations en Europe. Quelle articulation? ", Actes de la recherche en sciences sociales, 219, septembre 2017, p.80-97.

Frédéric Lebaron, "L'espace des conditions de vie des actifs occupés en Europe en 2010-2012", in Coll.,ESEG=European Socio-Economic Groups, Nomenclature socio-économique européenne, Document de travail INSEE, 2016, p. 83-98.

Cédric Hugrée, Etienne Pénissat, Alexis Spire, Les classes sociales en Europe. Tableau des nouvelles inégalités sur le vieux continent, Marseille, Agone, 2017.

16h30 – 17h00 C. Toque (Ministère de la transition écologique et solidaire. Ministère de la cohésion des territoires, France)

Segmentation territoriale des attributions de logements sociaux par l'Analyse des Données Symboliques

Résumé: L'objectif poursuivi était de créer des groupes homogènes d'EPCI (Etablissements publics de coopération intercommunale) en fonction des caractéristiques tant des logements sociaux attribués que des ménages à qui ces logements se voyaient attribués. L'analyse proposée est donc multivariée car elle s'intéresse à la distribution conjointe d'une vingtaine de variables issues des demandes et des attributions du SNE (Système national d'enregistrement des demandes de logements sociaux comptant près de 2 millions de demandes et 500 000 attributions au 1er décembre 2016).

A cet effet, des classifications par nuées dynamiques successives sont exécutées sur des variables de type histogramme calculées au niveau des EPCI (1255 en 2017). Ces variables ne sont ni des moyennes ni des médianes mais des « symboles » restituant au mieux la variation interne des données des communes par EPCI et entre EPCI.

Un partitionnement optimal en 4 classes est retenu pour lequel 3 variables discriminent au mieux les EPCI entre eux. Il s'agit, au moment de la demande, du montant mensuel des aides au logement perçus par le demandeur, du loyer mensuel du demandeur, et du montant maximum de la dépense de logement supportable. Puis, vient la durée d'attribution du logement. Moins discriminantes sont la situation professionnelle du demandeur, les ressources mensuelles du foyer demandeur et la demande du type de logement contrainte à la composition détaillée du foyer (demande « normée »), etc.

Force est de constater que le type de logement attribué est bien moins discriminant entre les classes avec toujours plus de T3 attribués, puis des T4 et des T2. Par ailleurs, il est à noter que le type de logement attribué est plus en adéquation avec celui recherché qu'avec la demande « normée ».

Pour ajouter à l'identification de ces classes, des variables de l'INSEE sont projetées sur le partitionnement en question. La typologie urbaine de 2010 reflète assez bien la segmentation territoriale obtenue. Autrement dit, en classant les EPCI en fonction des caractéristiques des attributions de logements sociaux et des ménages à qui ces logements se voyaient attribués, on aboutit à 4 classes qui reflètent la géographie française (les métropoles, les EPCI en périphérie d'un grand pole, les zones rurales dynamiques et les zones rurales).

References

Afonso, F., Diday, E., Toque, C. (janv. 2018). Data Science par Analyse des Données Symboliques. Technip. 448 pages.

Driant, J.C., Navarre, F., Pistre, P. (déc. 2016). Etude de l'offre du secteur des organismes d'HLM et SEM au regard de la demande de logement social en France métropolitaine. ANCOLS - Etudes et statistiques.

17h00 – 18h00

SUMMARY - OPEN DISCUSSION