Uncertain clustering of social specialization in metropolitan areas

Giovanni Fusco[1*], Cristina Cao[1]

[1] UMR ESPACE, CNRS / University of Nice Sophia-Antipolis, France

*Corresponding author: giovanni.fusco@unice.fr

Abstract: Instead of defining *a priori* target populations, sociodemographic variables and sector of residence are combined in order to identify geographically meaningful clusters of households on the French Riviera. A Bayesian classifier produces uncertainty-based clusters whose uncertain knowledge is represented through an interactive geo-dataviz solution. Cluster characteristics, sociodemographic content of geographic sectors, socio-spatial contrasts between neighboring sectors and proximity in variable space are explored taking into consideration the uncertain character of the available knowledge.

Keywords: Social Specialization, Uncertainty, Clustering, Bayesian Networks, Visualization, French Riviera

I INTRODUCTION

Social specialization of residential space within an urban area is the concentration of households according to some characteristics like social status, demography, ethnicity, etc. in different urban subspaces. The metropolization process is often associated to increased social specialization in large urban areas (Lacour and Gaschet 2008). Taken to the extreme, social specialization of residential space can produce residential segregation, creating socioeconomic dividing lines within the metropolitan space and undermining social and territorial cohesion (Massey 1985). In France, for example, important policies are carried out in order to assure that municipalities within metropolitan areas have a minimum social mix in terms of household income (Blanc 2010). Nevertheless, knowledge of social specialization of space in metropolitan areas is still incomplete and general assumptions have to be confronted to empirical data in case studies. Understanding the logics of social specialization is of course crucial in order to define policies like urban planning or housing. More specifically, which groups of households are most opposed in residential space? And which socio-demographic factors contribute most to these oppositions in space? How segregated are they in space? If we recognize that the answers to these questions are affected by high levels of uncertainty, which uncertainty-based methods could be used in order to describe our uncertain knowledge of social specialization? Finally, how could we represent and communicate most effectively this uncertain geographic knowledge?

In order to answer these questions, a new clustering approach is proposed integrating both sociodemographic descriptors of households and geographic distribution of their place of residence. The case study of the analysis is the metropolitan area of the French Riviera, a coastal conurbation of more than 1 million inhabitants stretching over 60 km west of the French-Italian border, and including the coastal cities of Nice (348 000 inhabitants in 2013),

Antibes (76 000 inhabitants), Cannes (74 000 inhabitants) and the Principality of Monaco (38 000 habitants), which is an independent city-state within French territory. These cities form nowadays an urban continuum and extend their influence over their alpine hinterland, which absorbs a growing part of the metropolitan population. Being a traditional residential destination for affluent retirees, a more recent hub of high-tech development and the necessary home for large populations of low-skilled workers of the tourist and residential economy, the French Riviera is particularly concerned by social specialization of residential space (Centi 1993, Billard and Madoré 2009, Fusco and Scarella 2011).

II METHODOLOGY: BAYESIAN CLUSTERING WITH SPATIAL CONSTRAINT AND INTERACTIVE GEO-DATAVIZ

Analyses of social specialization normally start by identifying target populations, whose segregation indicators are later calculated (Apparicio 2000). In our research, we opted for a bottom-up uncertainty-based approach: clusters of households were identified through data mining of selected sociodemographic variables within the 2008 Household Mobility Survey (which unfortunately does not cover the Principality of Monaco) combined with place of residence. Our starting hypothesis is that wealth differences are not the only factor beyond social specialization. A sample of 7539 households, representative of the population of the 94 sectors of the metropolitan area, has been analyzed through 16 variables describing social status, household structure, household demography and place in the workforce. Weights are attributed to variables in order to give approximately the same total weight to the four different dimensions of the analysis (Table 1), social status being slightly overweighted as its three variables cover more diverse issues. The *a priori* clustering of variables in four groups has been validated through Bayesian clustering algorithms based on mutual information among variables, given the empirical data. A 10-fold cross-validation procedure yields an average fit score of over 90% for this variable grouping.

Thematic Area	Indicator	Weight
Place in the Workforce	Occupation of the person of reference	1
	Number of working active people	1
	Number of unemployed people	1
Total Weight 4	Number of inactive people	1
Social Status	Maximum profession and socio-professional category among the spouses	2
	Maximum qualification among the spouses	2
Total Weight 6	Occupancy status of the dwelling	2
Household Composition	Presence of spouse in the household	1
	Number of household members	1
	Number of children	1
Total Weight 4	Number of other members	1
Household Demography	Number of minors (0-17 years)	0.8
	Number of young adults (18-29 years)	0.8
	Number of adults 30-59 years	0.8
Total Weight 4	Number of seniors (60-75 years)	0.8
	Number of elderly (more than 75 years)	0.8

Table 1 – The 16 indicators used to cluster households on the French Riviera.

Once the residence sector of the household is added as a further variable, different strategies of Bayesian clustering (Korb and Nicholson 2004) have been explored in order to produce an uncertainty-based socio-geographic clustering. More particularly, a naïve Bayesian classifier has been used with the variable weights of Table 1, with different constraints on maximal

number of clusters and minimal cluster content and with different weights for the new variable sector of residence. A minimal cluster content of 4% of the household population has finally been selected to avoid overfitting on the sample. As for the sector variable, weights beyond 3 forces the algorithm to consider it as the leading variable of the clustering: the likelihood of the resulting clustering is maximized by assigning the population of a given sector to one or two clusters only. This is clearly not the goal of our clustering, as we want to find dividing lines in the resident population that take into account place of residence together with and not instead of sociodemographic differences (the poor quality of clustering results based on place of residence only is also witnessed by the low contingency table fit on the 17 variables). The optimal compromise was found with a weight of 2 for the sector variable. In this case, the sector variable is only the fourth strongest variable in terms of mutual information with the clusters and the contingency table fit is 36,4% on all the variables.

This approach is different from previously developed research (Pallez et al. 2015): clustering is uncertain because household assignment to a given cluster is probabilistic and households can have several non-zero probabilities of being assigned to different clusters. Average probabilities in assigning a given household to a cluster range between a maximum of 0.97 (for cluster 11) to a minimum of 0.86 (for cluster 4). Individual households can have much lower probabilities, and sometimes even have similar probabilities of being assigned to two different clusters. Passing from the sample to the household population introduces additional uncertainties. Cluster labels are vague, too, in the sense that they are synthetic descriptions combining different variable values which are often (but not constantly) associated. The sociodemographic characteristics of clusters are thus described through Bayesian probabilities. Social specialization of metropolitan sectors with respect to these clusters is evaluated in terms of the classical dissimilarity index (Duncan and Duncan 1955), but different evaluations are proposed for different levels of uncertainty.

One of the main difficulties of the analyses was to convey the uncertainty associated with the results obtained. Several authors have already proposed approaches for graphical representation of uncertain information (MacEachren 1992, MacEachren and Howard 1993, Ehlschlaeger et al. 1997, Cedelnik and Rheingans 2000, Ward 2002). Within our research, we propose an interactive online geo-data visualization solution in order to explore the results of uncertainty-based analyses (Cao and Fusco 2015). Systems of dashboards for interactive visualization seem particularly useful in representing uncertain and complex phenomena like social specialization of space. First attempts of interactive representation of uncertain geographic data were proposed in the 1990s (Ehlschlaeger et al. 1997) even if they were not considered particularly effective in their applications (Evans 1997). Advances of the software interfaces have since been considerable and new applications seem to be both more user-friendly and scientifically sophisticated (Ban and Ahlqvist 2009, Kunz et al. 2011, Fusco et al. 2016). These applications link together interactive representations in the forms of maps, diagrams and text. In our case, different analyses are developed with respect to uncertain knowledge and represented in the geo-dataviz solution.

III RESULTS FROM THE FRENCH RIVIERA

Results of the Bayesian clustering of households in the French Riviera will be presented and commented through four visualizations, taken from the interactive geo-dataviz solution (Cao and Fusco 2015).

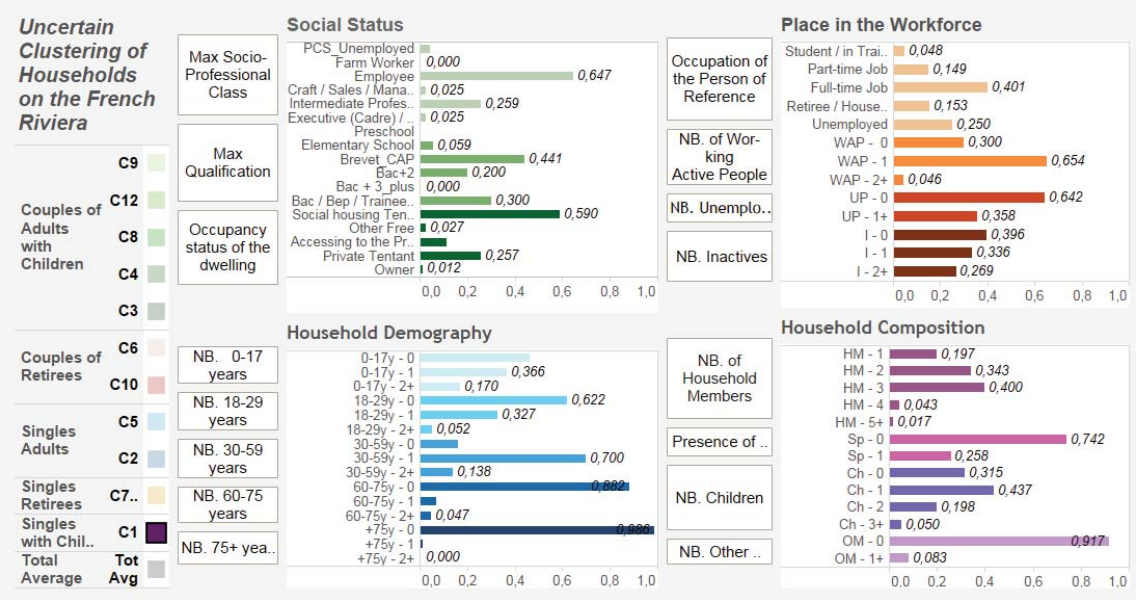

Figure 1: Probabilistic description of household clusters.

Figure 1 gives a probabilistic description of household clusters. The Bayesian classifier identified 12 clusters of households, which were later regrouped in 11 clusters after interpretation : five clusters concern families with children and differ in terms of social status, from highest (Cluster 9, families of skilled executives/professionals, very often owners of their dwelling) to lowest (Cluster 3, families of employees, with lower level of education and more frequently tenant than owner, sometimes social housing tenant); two clusters concern couples of retirees (of different social status); two clusters concern single adults (here social status is at least partially correlated with age); two clusters concern single retirees (they have been regrouped in one cluster only as their sole difference was the different age class 60-75 years or 75 years and more); a last cluster is specific to households of single parents with children, with difficult social situation (low education levels, employee not always with a full-time job and sometimes unemployed, tenant of social housing or in the private sector). The latter is, by no surprise, the most segregated cluster, with a dissimilarity index of 0.75. The exploration of the probabilistic values of the different variables for each cluster is of course richer than this simple summary of main features and lets the user better understand the vague content of the cluster labels and the subtle differences that sometimes exist between two relatively similar clusters.

Figure 2 represents through maps and diagrams the socio-demographic content of metropolitan sectors. Clusters of households are differently distributed in space. Every sector of the metropolitan area has a different probability profile in terms of cluster belonging of its resident population. Clusters can also be of particular importance for a given sector. In this visualization, like in the following ones, sectors can be linked to clusters according to different criteria: modal cluster, most overrepresented cluster (largest positive deviate from metropolitan

average), most underrepresented cluster (largest negative deviate from metropolitan average), most characteristic cluster (highest location quotient compared to the metropolitan average).

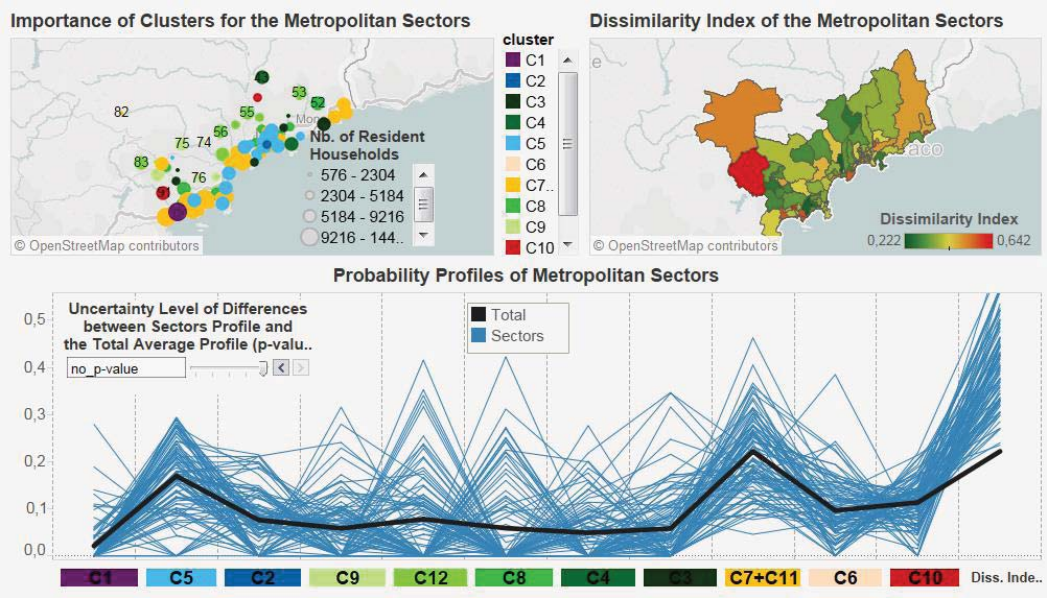

Figure 2: Sociodemographic content of metropolitan sectors.

Finally, the dissimilarity index measures which proportion of a sectors' population should move in order to attain the same probability distribution as the whole metropolitan area. Uncertain knowledge of population content of sectors can be explored in the visualization. Uncertainties derive both by the uncertain cluster assignment of households and by the sample process. Statistical significance of the observed deviates from metropolitan-wide values is evaluated against the null hypothesis that the percentage of a given cluster in the household population of the sectors is given by a binomial distribution whose expected value is the share of the cluster in the whole metropolitan household population. Given the sample size of every sector (always between 70 and 103), the binomial distribution can be approximated by a normal distribution. In the data-viz, the significance level can be set by the user: every time more certainty is required in knowledge, non-significant differences are omitted. The city-center of Cannes, for example, with strong and extremely significant over-representation of single retirees, is not necessarily the sector with the highest dissimilarity index, but becomes such once only the most significant deviations from the metropolitan values are retained. The user can thus interactively identify the deviates characterized by the lowest levels of uncertainty, which contribute significantly to the social specialization or residential space.

Socio-spatial contrasts between neighboring metropolitan sectors are represented in Figure 3. The diverse spatial repartition of clusters within the metropolitan area results in socio-demographic differences among contiguous sectors, which can have important impacts on the social functioning of the metropolitan area. The population content of sectors being known as probability distributions, their differences are measured as distribution divergence. More particularly, we use the Jensen-Shannon divergence (Lin 1991), which has the advantage of being symmetrical and defined even when some clusters are absent in a given sector. Socio-demographic distance between contiguous sectors is thus coherently measured with a probabilistic divergence. Once again, the uncertainties in socio-demographic content of sectors

can be used to produce different measures of socio-spatial contrasts, according to significance level of differences from the metropolitan-wide values. The user can thus identify those contrasts which are identified with the lowest uncertainty. The socio-spatial contrasts in the city of Nice (at the center of the metro area) are thus determined with higher levels of certainty than those west of Nice, where some differences are relatively uncertain.

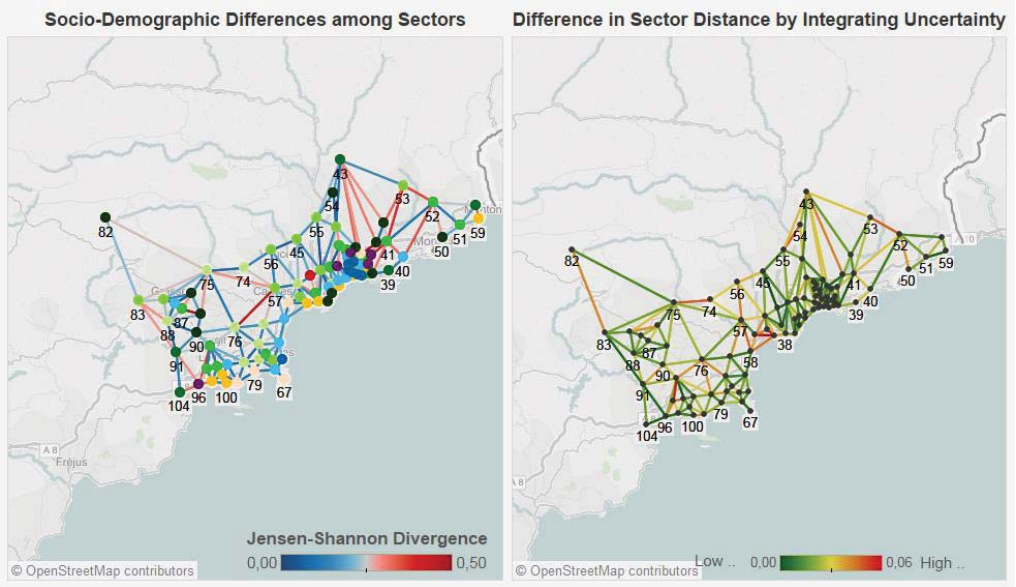

Figure 3: Socio-spatial contrasts between neighbouring metropolitan sectors.

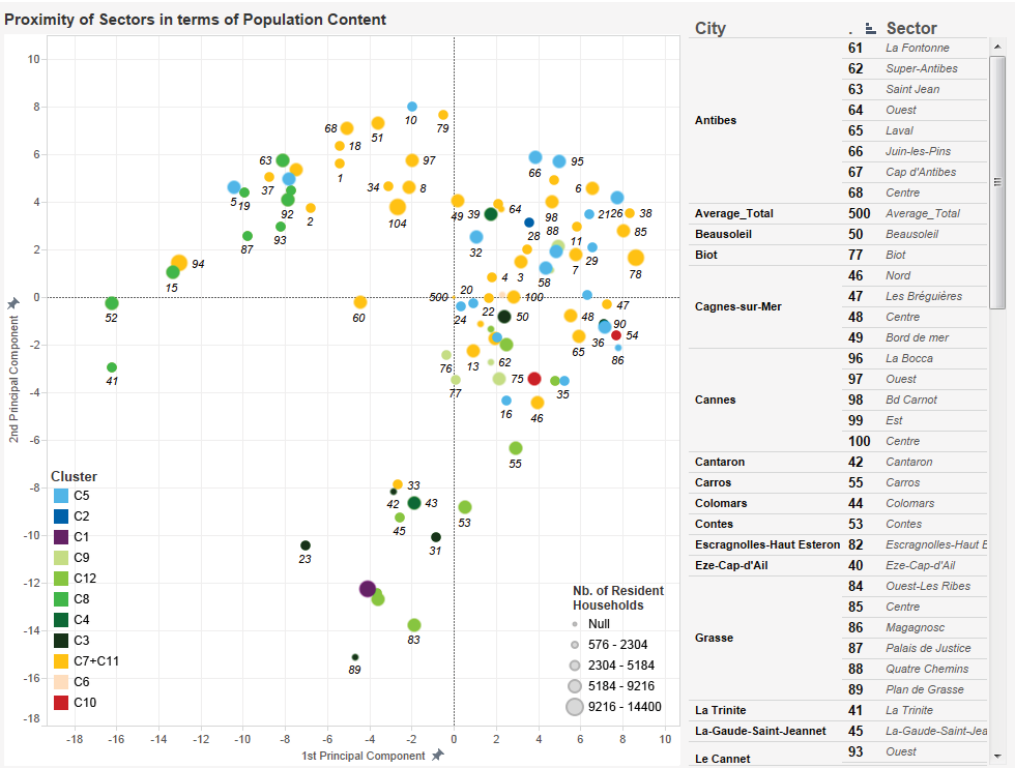

Figure 4: Proximity of metropolitan sectors in variable space.

Figure 4 represents proximity of metropolitan sectors in variable space. Jensen-Shannon divergence values make up a complete distance matrix among the metropolitan sectors. A principal component analysis can be performed in order to visualize the relative proximity of sectors in terms of socio-demographic content, on the main factorial plane. Geographical space proximity can thus be evaluated in conjunction with sociodemographic proximity in a multi-dimensional scaling approach. Once again, reducing the level of uncertainty through minimum significance levels modifies the results of the analysis. Through this analysis we can highlight the extreme diversity of social specialization within the city of Nice, a phenomenon which is not observable in other main cities like Cannes and Antibes. At the same time, the city of Cannes presents a marked opposition between its westernmost sector, strongly marked by the presence of single parents with children, in difficult social situation, and the rest of the city, where all sectors show very high presence of single retirees, and low presence of couples with children. All these results are particularly robust to uncertainty levels and can be considered among the most salient characteristics of our study area.

IV CONCLUSIONS AND PERSPECTIVES

In conclusion, with the example of the French Riviera metropolitan area, our research shows how knowledge of social specialization of residential space is uncertain. In this context, instead of defining *a priori* target populations, soft clustering techniques could be used to identify the most important sociodemographic divides in a given metropolitan area. For the French Riviera, position in the life-cycle as described by age and household composition seems even more important than social status in defining social specialization of space. Results of soft clustering benefit from appropriate geo-data-visualization solutions. In our example, a system of dashboards seems an appropriate way to describe complex phenomena like social specialization of space. Within this solution, knowledge uncertainty can be conveyed by interactive representations, whenever an appropriate calculus (in our case through the use of probabilities) can quantify the most relevant uncertainties.

The present work opens important perspectives in the search for uncertain geographically meaningful clusters. Concerning the analysis of social specialization of space, ethnicity is a crucial aspect which is missing in the French statistical information system and could only be integrated in the study through *ad hoc* surveys. As far as the methodology is concerned, the use of a naïve Bayesian classifier is a first solution and more sophisticated methods should produce better results. Multi-level clustering with hierarchical naïve Bayesian classifiers (Langseth and Nielsen 2006) could for example better exploit the structure of available information, where strong relations exist within groups of variables. The problem of identifying a correct weight for the geographic variable could be eliminated altogether by the use of algorithms of multi-objective optimization. We could thus optimize at the same time clustering likelihood based on the sociodemographic variables and dissimilarity index based on the geographic distribution of the clusters. The exploration of the Pareto front would then be instructive of role of the two criteria in the identification of socio-geographic clusters.

References

Apparicio P. (2000). Residential segregation indices : a tool integrated into a geographical information system. *Cybergeo: European Journal of Geography*, 134, http://cybergeo.revues.org/12063

Ban H., Ahlqvist O. (2009). Representing and negotiating uncertain geospatial concepts - Where are the exurban areas?, *Computers, Environment and Urban Systems*, 33(4), pp. 233-246.

Billard G., Madoré F. (2009). Les Hauts du Vaugrenier: un exemple atypique de fermeture résidentielle en France, *Mappemonde*, 1-2009, http://mappemonde.mgm.fr/num21/lieux/lieux09101.html

Blanc M. (2010). The Impact of Social Mix Policies in France, *Housing Studies*, 25(2), pp. 257-272.

Cao C., Fusco G. (2015). *Representing Uncertain Clustering. The case of Social Specialization on the French Riviera*, https://public.tableau.com/profile/fusco#!/vizhome/RepresentingUncertainClustering/Story

Cedilnik, A., Reinghans, P. (2000). Procedural annotation of uncertain information, In *Proceedings of Visualization '00*, IEEE Computer and Society Press, pp. 77-84.

Centi C. (1993). Les enjeux du modèle niçois. L'approche localiste du développement en question. *Revue Economique*, 44(4), pp. 687-712.

Duncan, O., B. Duncan (1955). A methodological analysis of segregation indexes. *American Sociological Review,* 20(2), pp. 210-217.

Ehlschlaeger C., Shortridge A., Goodchild M. (1997). Visualizing Spatial Data Uncertainty Using Animation, *Computers & Geosciences*, 23(4), pp. 387-395.

Fusco G., Cao C., Dubois D., Prade H., Scarella F., Tettamanzi A. (2016). Social Polarization in the Metropolitan Area of Marseille. Modelling Uncertain Knowledge with Probabilistic and Possibilistic Networks, ECTQG 2015 Proceedings, *Plurimondi*, 8 p.

Fusco, G. et F. Scarella (2011). Métropolisation et ségrégation sociospatiale. Les flux des migrations résidentielles en PACA. *L'Espace Géographique*, 40(4), pp. 319-336.

Korb K., Nicholson A. (2004). *Bayesian Artificial Intelligence*, Chapman & Hall / CRC.

Kunz, M., Regamey-Gret, A., Hurni, L. (2011). Visualization of uncertainty in natural hazards assessment using an interactive cartographic information system, *Natural Hazards*, 59(3), pp. 1735-1751.

Lacour, C. et F. Gaschet (2008). *Métropolisation et ségrégation*, Bordeaux: PUB.

Langseth H., Nielsen T. (2006). Classification using Hierarchical Naïve Bayes Models, *Mach Learn*, 63, pp. 135–159.

Lin J. (1991). Divergence Measures Based on the Shannon Entropy, *IEEE Transactions on Information Theory*, 37(1), pp. 145-151.

MacEachren, A. (1992). Visualizing uncertain information. *Cartographic Perspective*, 13, pp. 10-19.

MacEachren, A. M., D. Howard, et al. (1993). Visualizing the health of Chesapeake Bay: An uncertain endeavor, In *GIS/LIS Proceedings*, vol. 1, American Society of Photogrammetry and Remote Sensing/American Congress on Survey and Mapping, Bethesda MD, pp. 449–458.

Massey D. (1985). Ethnic residential segregation: A theoretical synthesis and empirical review. *Sociology and Social Research*, 69(3), pp. 315-350.

Pallez D., Serrurier M., Da Costa Pereira C., Fusco G., Cao C. (2015). Social Specialization of Space: Clustering Households on the French Riviera, *GECCO Companion 15 - Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*, ACM, New York, pp. 1447-1448.

Ward, M.O. (2002). A taxonomy of glyph placement strategies for multidimensional data Visualization, *Information Visualization*, 1(3/4), pp. 194-210.