International Workshop on Spatial Data and Map Quality, January 2015

Reshaping socio-spatial representativeness from probabilistic survey data: a case study from Marseille

Frédéric Audard[1] *& Samuel Carpentier*[1]

[1] Aix-Marseille Université, CNRS, ESPACE UMR 7300, Aix-en-Provence, F-13621, France
frédéric.audard@univ-amu.fr; samuel.carpentier@univ-amu.fr

The emergence of big data, as well as open data, has led the scientific community to question data collection in different ways: construction, acquisition and ownership. Traditionally, social sciences have been concerned by the opportunities given by data collection or access to existing datasets. The recent changes related to automatic data collection and sharing makes attitudes and expectations of researchers evolve. Social sciences are thus expected to change their current approaches and methodologies regarding data. Geography is a discipline largely concerned with data collection and analysis and is then at the core of this dramatic change. The wide and fast spread of numerous geolocation tools (based on GPS, GSM, WiFi or IP address) leads to the surge of georeferenced data (Audard, Carpentier, Oliveau, 2014).

If an enthusiastic posture currently dominate the appreciation of open/big data in terms of scientific perspectives (Marx, 2013), some cautions emerges regarding their use. These data are generally not built for the purpose of social sciences (Pumain, 2014) and are thus to be used with care for scientific work (Terrier, 2011).

Beyond those general considerations, one major fact related to their use in geography is the question of the spatial, social, demographic and economic representativeness of such « second hand » datasets built for their own purpose and objectives. In the case of big data, the illusion of completeness tends to obscure the question of representativeness. The huge number of observations does not mean that there is no selection bias; even if a big sample size produce good confidence intervals. That's the main question of this paper: how to reshape the socio-spatial representativeness from data that are not built with regard to the question of spatial representativeness?

The equation (1) describe the confidence interval related to our sample. This interval increases dramatically regarding different control variables dedicated to ensure the sample representativeness with respect to demographic and socio-economic variables (2). However, for big samples those confidence intervals remain acceptable.

(1) $$Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$Z_{\alpha/2}$ is the confident interval, with α = confidence level, σ = standard deviation and n = sample size

(2) $$Z_{\alpha/2} \times p \times \frac{\sigma}{\sqrt{n}}$$

With p = number of control variables

The problem arises when we want to ensure both social and spatial representativeness. The study of geographical phenomenon, occurring in defined spatial units, implies to take account of the related socio-spatial structure. For example, mobility behaviors are known as varying depending on gender, social status or income. Each spatial unit has its own characteristic regarding those variables, it is then necessary to ensure specific representativeness of every spatial unit. Thus, if such representativeness has not been taken into account in the sampling procedure, the confidence interval is modified (3) and can be problematic.

(3) $$Z_{\alpha/2} \times p \times i \times \frac{\sigma}{\sqrt{n}}$$

With i = number of spatial unit of the study area

In order to tackle this issue, this paper suggest a methodology built on the measurement of confidence intervals allowing to reshape socio-spatial representativeness. The principle of this method is to identify the most underestimated sub-population in each spatial unit comparing to a theoretical expected frequency. Then, these sub-population are re-sampled in order to fit the theoretical frequency (under the constrain of a minimum acceptable frequency for each sub-population).

For each couple of control variables q and r, we can calculate a corrected sample of the sub-population concerned by the control variable *q* within the spatial unit *k* (noted $n'_{q.k}$,) with:

$$n'_{q.k} = \begin{cases} n_{q.k} & \text{if for each r we have: } \frac{\widetilde{n_{q.k}}}{N} - \frac{n_{q.k}}{N} > \frac{\widetilde{n_{r.k}}}{N} - \frac{n_{r.k}}{N} \\ \frac{n_{q.k}\left(\widetilde{n_{r.k}}/N\right)}{\widetilde{n_{q.k}}/N} & \text{else} \end{cases}$$

With $\widetilde{n_{q.k}}$: theoretical frequency of sub-population concerned with the control variable q within spatial unit k
$n_{q.k}$: observed frequency of sub-population concerned with the control variable q within spatial unit k

Finally, we are developing a control indicator. Indeed, the gain corresponding to the best socio-spatial representation should not be lost by an excessive cut of the sub-population that would increase the confidence interval instead of reducing it. We then have to checked independently for each spatial unit *k* that:

$$Z_{\alpha/2} \times p \times \frac{\sigma}{\sqrt{n_{q.k}}} < Z_{\alpha/2} \times p \times \frac{\sigma}{\sqrt{n'_{q.k}}}$$

The case study focuses on the analysis of leisure mobility on the Marseille's coast in France. The data comes from the Household Travel Survey of the disctrict of Bouches-du-Rhône from 2010 (n=13600 households).

References

Audard, F, Carpentier, S., and S. Oliveau (2014), "Les big data sont-elles l'avenir de la géographie [théorique et quantitative]?" Proceeding of the conference Géopoint. http://www.groupe-dupont.org/ColloqueGeopoint/Geopoint14/Documents/GP14_PropositionsDebat_Web/GP14-A2-1-Audard-Carpentier-Oliveau.pdf

Pumain, D., (2014), "Editorial: Observation, observation, observation", *Cybergeo-European Journal of Geography*, http://cybergeo.revues.org/26248

Terrier, C., (2011), "La valeur des données géographiques", *L'espace Géographique*, 40(2):103-108.

The authors thanks the support of the "OHM Littoral méditerranéen" from CNRS