Short abstract

Choosing the appropriate scale of analysis is a well-known problem in regional studies. Changing the level of quantitative spatial studies is not trivial and can have many substantial effects on the indicators, their representation and their interpretation, especially if the original data is spatially diverse and at a fine granularity. One cannot re-group spatial data points automatically without consequences. As geographical datasets are becoming increasingly available, and with a finer resolution, we want to provide some pointers and use-cases scenarios from a geography standpoint, to present and analyse this issue. What definitions can be used to delineate the idea of a city, an urban area or an agglomeration? At what size, population density, volume of urban activity or surface are we observing a city? We consider this issue especially in a comparative purpose taking the example of analysing networks of European cities.

In this proposal, we provide a comparative analysis of spatial clustering methods or aggregation procedures with an interactive web-based application designed to explain and visualize their effects on different results: volume of urban activity (discrete values) and rankings, on the one hand, spatial configurations, sizes and relational flows (links) of the aggregated values, on the other hand. We will use the themes of demographics and the geography of scientific activity, at the city scale in their spatial and relational dimensions.

Extended abstract

What is a city? On a continuous geographical space, what is the delineation which encloses enough content to allow thinking of it as a city? What if we change the scale of analysis, from regional to international?

This question is at the core of urban geography and regional studies. For a long time, measures were collected and statistics were produced according to political-administrative divisions or dedicated territorial frameworks. This means that they were carried out within, firstly, existing territorial partitions and, secondly, the framework of a discrete approach; the problem with this approach is that such partitions of the geographical space (by definition continuous) into distinct areas are more or less heterogeneous. These divisions were often used by default by analysts from the 19th century until the late 1960s, the early 1970s, and before the so-called spatial turn. This bias is important to take into account when one hopes to compare data spatially and study flows between locations. Various authors in demography, economy and geography, have been able to demonstrate the binding role of such political-administrative divisions in the implementation of geographical or economical models, in particular those concerning the analysis of spatial interactions (Alvanides & al., 2000). Some of them have also proposed partition methods that ignore administrative divisions (for a recent review, see: Van Hamme et al., 2011), or that are directly based on relational data (links or flows), such as methods that maximize cumulated intra zonal interactions for example. In this case, what is important is the choice of the aggregation function, taking into account its effects on the level and the process (Masser & Brown, 1975 ; Hirst, 1977). Similarly, the instability of statistical results in the context of a variable geographical unit (better known as the Modifiable Area Unit Problem -
MAUP problem) is proven (Openshaw, 1977). These problems are acute, particularly at the international and global scale - which is already sensitive to the choice of the mapping projection system: first, administrative areas are not designed to delineate functional areas and, second, they are not easily comparable between countries. As geographical datasets are becoming increasingly available, with a finer scale and local data points, this issue is particularly relevant.

Concretely, the questions that we need to address are:

1) How to consider the geographic space: as a discrete partition where cities are points or areas (depending on the scale) or as a (continuous) surface where cities are defined by a scope with potentially fuzzy boundaries?

2) How to associate local data points into meaningful aggregations, adjusted to the analysis, called clusters or functional regions? The effects of an unadapted clustering method can be quite elusive to the researcher, due to their complexity and subtle variations, particularly spatial ones. Actually, the spatial component of the problem, combined with the different scales of analysis and the exploration of flows can rapidly muddle the situation.

With this contribution, we want to provide an update on the issue, to expose the main methods of spatial clustering and to illustrate the effects of their parametrization on the face of the map. We would also like to explore these variations graphically, by proposing several innovative interactive representations. Indeed, we think that a hands-on approach can be useful to describe the issue, increase its awareness and explore the parameters of several methods and their effects. To illustrate their diversity, these methods will be selected from two families: purely geometrical and weighted. Indeed, several methods of spatial clustering are using only the relative positions and the spatial density of the data points to regroup them. Other methods can take into account weighting and/or spatial parameters such as, 1) contiguity or spatial continuity in the aggregation process, 2) the intensity of a phenomenon (population, scientific production), or 3) the values of networking properties at a global or a local level (as centrality or connectivity) on reticular or flow data. From the first group, DBSCAN (Ester et al., 1996) and its variants are currently being widely used, but we will show that hierarchical classifications (like AGNES, cf. Kaufman and Rousseeuw, 2009) can also be very effective, with a cautious attention to fine-tune their parameters. From the second group, we will include extensions of DBSCAN methods like DBCluc (Zaïane et al., 2002) and DBRS (Wang and Hamilton,, 2005), as well as methods based on graph theory that show some promises, like Autoclust+ (Estivill-Castro and Lee, 2004) and ASCDT+ (Liu et al., 2013). Another useful possibility of these last methods is the consideration of limits, obstacles and spatial friction (or, inversely, of easier connexion between points and regions). Besides these algorithms, we will provide the comparison with pre-defined clusters or delineations, such as administrative divisions or functional spatial territories created precisely to observe the cities of the European space in a comparative manner ("Functional Urban Areas", cf. Guérois et al., 2014, for example, but also "Urban Morphological Zones", from the ESPON projects).

We will use several spatial datasets based on point data (for places) or couples of points (depicting origin-destination flows) with two themes: demography (international migration dataset) and science (relying on a more complex but interesting set about the geography of scientific production). Demography is well suited to study the consequences of the clustering variations, as it presents a well-known geography and can be easily related to city rankings and their evolution in time (Pumain al., 2015). By adding the relational dimension of internal migration, we can examine the effects on flows between cities and regions. The dataset is available with the Eurostat and ESPON usual public providers. The more specific subject of scientific production, which we are exploring for several years with geocoded data from the Web of Science bibliographical database, is especially interesting due to the surprisingly very
recent consideration of this clustering issue in spatial scientometrics (Maisonobe et al., 2018). We aim to demonstrate the harmful effects of dubious clustering decisions, such as the use of administrative divisions to compare the scientific production at a European scale. These two subjects will be examined with several case studies, to provide different levels of complexity for different audiences (pedagogical examples and more in-depth analysis).

The approach we want to implement is intended to be generalizable and reproducible (Giraud and Lambert, 2017). This is why we propose to provide R programs, combined with an RShiny application. This proposal is in accordance with the principle of "muti-cartographic representation" (Zanin et Lambert, 2012), by allowing exploration and visualization of linked graphic and cartographic depictions. The R platform provides a really interesting collection of tools to analyse data in real time (with specialized clustering modules), but also to represent the results on interactive maps and graphs. To expose the spatial component of the clustering problem, an interactive map will show the spatial data points and the limits inherent to several clustering methods. The web application will show interactively the effects of changing the clustering method itself or the variation of its parameters, by redrawing the map in real time. To demonstrate the consequences of the clustering variations on the aggregated end-values, the application will present the tables of aggregated values, with the possibility to filter or query the data (to reduce the set or to explore more finely the results) and several graphical representations: histograms and more comparative graphs such as Sankeys and bubblecharts. The availability of interactive web representations, helped by the development of programming libraries as R modules and D3 JavaScript functions, will permit an interactive exploration of the representations, which will help understanding the reality of the clustering problem and its effects.

Several possibilities of extension exists, especially in the way of interactive flow maps, with the collaboration of the existing gFlowiz research program1, on the one hand, and, on the other, with the exploration of the possibility to combine the data with other subjects, such as transportation data or geographical friction and obstacles.

Bibliography


1 Cf. the website of the project: http://37.187.79.5/gflowiz/


