Grid data and Confidentiality

Keywords: *disclosure, grid data, privacy, census data, graph*

1 Introduction

Grid data consists in creating a grid of cells on which the geolocalized information is aggregated. The advantage of such a grid is that it provides finely localized information allowing a user to analyze *a posteriori*, the zoning that would be specific to him by grouping cells. An important issue with this kind of data is the confidentiality; the smaller the cell, the greater the risk of statistical disclosure.

A distinction is generally made between primary statistical secrecy, which relates to information directly made available to users, and secondary statistical secrecy, which relates to information that a user could infer indirectly from all the data released. The desire to preserve these two types of secrecy leads us to explore two methods. The first one based on the quadtree algorithm enables the respect of the primary secrecy by construction, the second makes it possible to identify areas at risk with regard to the secondary secrecy.

2 Methods

We introduce here a method of geographical aggregation based on the "quadtree" approach [1] applied on grid data, allowing to preserve confidentiality while detailing at the finest level. The new contours introduced with this first method, combined with other administrative contour can generate statistical disclosure through geographical differentiation, which leads us to study a graph-based method in order to detect these problems. This method described in the article "Detecting geographical differencing problems in the context of spatial data dissemination" [2] was imagined by Vianney Costemalle.

2.1 Geographical aggregation

The threshold rule is the most obvious way to respect confidentiality and consists in not releasing statistics for cells below a chosen number of observations. In the case of grid data, one strategy may be to assemble contiguous cells into larger polygons (e.g. larger rectangles and cells) so that each polygon meets the threshold.

The new polygons can be obtained by aggregation, *i.e* by grouping the cells until the threshold is reached, or by disaggregation, starting from the largest cell and splitting it until it is no longer possible to cut it without going below the threshold. These methods have the advantage of preserving the additivity but also allow to avoid the creation of "false zeros". In addition the threshold rule is respected by construction. The areas first obtained by disaggregation are called the "natural" areas and they can be of different sizes depending on when we stop splitting.

However, it is possible to obtain more accurate information if we continue to divide the resulting cells in smaller cells and if we decide to hide the information from the cells that are below the threshold. Obviously, it is not sufficient to hide only the cells below the threshold since the information can be found by geographical differentiation comparing the finer level of detail to a coarser level of detail. Therefore, we have to hide the information contained in another cell at the same level. This process requires disseminating information on several grids ("composite" grid) corresponding to different levels of detail, which is equivalent to disseminating information on the same grid with cells having different shapes and sizes. This can be difficult to understand for users.

Instead of hiding this information it is also possible to use a distribution key to disseminate information at the finest level: all cells below the threshold that are close to each other are grouped together, which defines a group G. Then we observe the total of the variable we want to disseminate on G and we distribute it between the different cells of G according to a distribution key built from the share of the group G population in each cell.

Figure 1 provides an illustration of the three detailed possibilities.

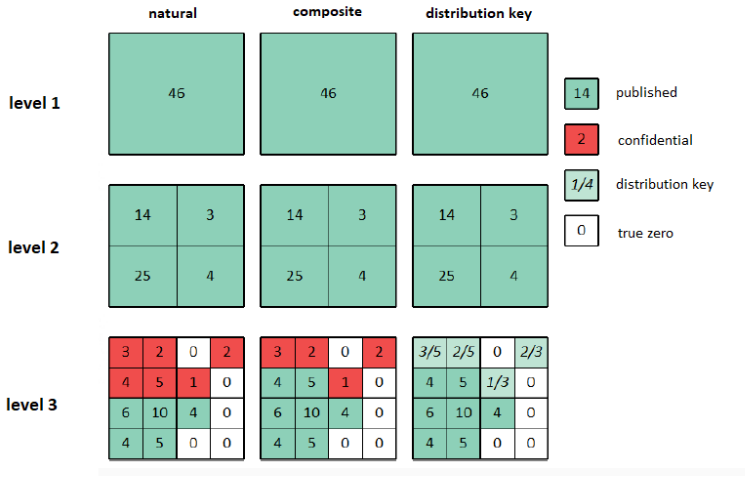

Figure 1: Different aggregation method for grid data with threshold 3.
Note: The disaggregation method with distribution key allow to release data at the finest level preserving confidentiality.

2.2 Geographical differentiation

Regardless of the method selected, the use of these new areas may generate a risk of secondary disclosure if it is combined with the same data released according to other non-nested administrative segmentation (see Figure 2). It is therefore necessary to be able to identify areas at risk of disclosure.

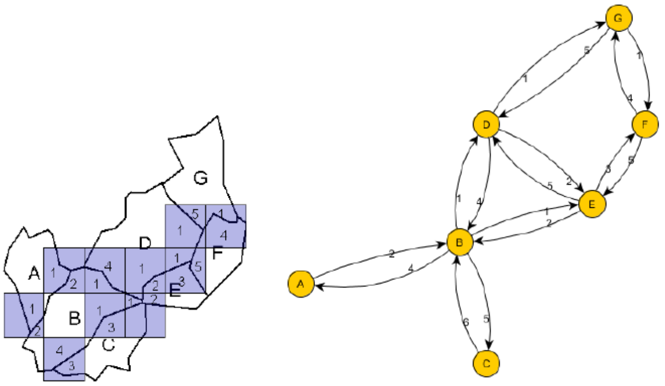

Figure 3: Graph representation of 2 administrative contours.

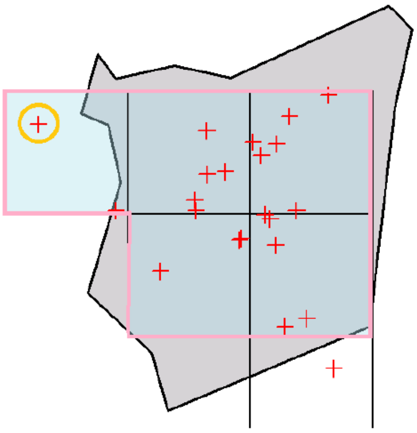

Figure 2: Exemple of geographical differentiation based on grid data.
Note: one can identify the cross surrounded by a yellow circle by geographical differentiation.

Considering two administrative geographies A and B, the strategy here is to use a graph representation in order to represent the areas of one chosen contour (A for example). Each area z^i, $1 \leq i \leq n_A$ of A is thus a vertex of a graph and we say that two vertices z^k ans z^l are connected if an area z_i^B, $1 \leq i \leq n_B$ intersects the two zones. the resulting graph is oriented since we weight the link from z_i^A to z_j^B by the number of observations contained in $z_i^A \cap [\cup_{j=1}^{n_B} z_j^B]$ and vice-versa.

This simplified representation (see diagram in Figure 3) allows us to simply identify the areas at risk since they are necessarily located at the intersections of the contours A and B. From this representation, the goal is then to identify sub-graphs where we can identify geographic differentiation issues. Before searching for these sub-graphs, which can be computationally very expensive, we perform a simplification phase on the initial graph by merging certain vertices. An exhaustive search is then performed on the reduced graph and enables detecting all the areas where disclosure can be achieved by geographical differentiation.

Used with grid data at the "natural" level and municipalities in France with a primary secrecy threshold of 11, this method detects 10 000 households at risk of geographical differentiation.

3 Conclusions

The areas built by disaggregation enables the disseminated data to respect the primary secrecy *a priori* but does not necessarily preserve the secondary secrecy. Since new administrative geographies can be introduced in the future, it is more relevant to control the secondary secret *a posteriori*, which is allowed by the graph-based method.

In the end, the disaggregation method using repartition keys allows a finer detail than the one giving only the natural level and is simpler to interpret than the method of composite disaggregation without distribution keys. This method combined with the graph-based disclosure risk detection method provides a solid process combining confidentiality and accuracy of the released data.

References

[1] Martin Behnisch, Meineln Gotthard, Sebastian Tramsen, and Disselmann. *Using Quadtree representations in building stock visualization and analysis.* Erdkunde, 2013.

[2] Vianney Costemalle. Detecting geographical differencing problems in the context of spatial data dissemination. *Statistical Journal of the IAOS*, pages 559–568, 2019.