Statistical disclosure control on visualising geocoded population data using quadtrees

Eduard Suñé (esl@idescat.cat)[1], Cristina Rovira (crovira@idescat.cat) [1],
Daniel Ibáñez (dibanez@idescat.cat) [1], Mireia Farré (mfarre@idescat.cat) [1]

Keywords: geolocation, quadtrees, European standard grid, Monte Carlo methods

1. INTRODUCTION

The Statistical Institute of Catalonia (Idescat) is updating its statistical methods and information systems following the ESS Vision 2020 guidelines of the European Statistical System [1]. This vision foresees an integrated system of administrative and statistical information in order to increase efficiency. To achieve such an aim, it will be compulsory (essential) to integrate administrative and statistical information.

To make possible this integration of information, one of the main keys is the Territory Statistical Register (RET in its Catalan acronym), which has to allow for the geocoding of the available geographic information in our statistical system. The aim of the RET is to have all postal addresses standardised and geolocated, including residences and other premises (commercial, industrial...).

Our aim is to establish a future-proof strategy of dissemination and communication that will satisfy user needs.

One of the objectives is the dissemination of geocoded information without this allowing for the disclosure of confidential data on the statistical subjects, people or companies. The strategy followed consists of aggregating a layer of points into a layer of polygons, using a quadtree [2] structure based in the European standard grid [3].

Finally, we have estimated, using Monte Carlo methods, the errors in the calculation of the population when different-sized quadtrees are used.

2. METHODS

Our method consists of the development of quadtrees for the diffusion of geocoded population data. The proposal is to perform the subdivision of the quadtrees in order to optimize their resolution in function of the population density.

In this type of structure, the space can be divided recursively, attaining more detailed levels depending on the population density, so that you can strike a balance between accuracy and the risk of disclosure. The division process is governed by a threshold value, below which the split ends.

The opposite process, aggregation from a maximum resolution, leads to the same results

[1] Institut d'Estadística de Catalunya

as the division of space, with the advantage that the division is more efficient from the point of computation.

In any case, if the chosen method is division, it must be started from a minimum resolution and finish in the place where you want to achieve the subdivision, according to some arbitrary threshold. Similarly, if you choose an aggregation method, it must be started from a maximum resolution and spatially add up to a reasonable minimum resolution while the total population lies below the threshold.

The method chosen for the construction of the quadtree is aggregation, based on a maximum resolution associated with a grid compatible with the European standard grid, 125m, where a threshold of 17 inhabitants was used, reaching a maximum resolution of 250 m.

However, in areas where the population density rapidly decreases, undesirable aggregations may appear as a consequence of the algorithm used in the quadtree generation. If the method chosen is the subdivision of space, in these type of areas mentioned above, the process of division cannot be performed to higher resolutions. This effect can lead to significant errors in the calculation of populations in areas that intersect with them.

The solution used in areas with significant changes in population density is the optimization of the quadtrees, by translating some of the population between common elements in the hierarchy of the quadtree. In areas with high degrees of variance, this solution prevents the excessive loss of information derived from aggregation method and preserves statistical information.

The criteria used to decide upon these translations is as follows: if the absolute error calculated when the situation after aggregation is greater in the previous situation than the absolute error calculated before and after the translations, then the disturbance has to be performed.

The elements being moved are distributed in proportion to the total number of potential donors in the same-level quadtree hierarchy and the choice is made randomly in order to minimize distortions in the distributions of the characteristics of the population.

Finally, the distribution of relative errors in calculating population has been estimated using Monte Carlo methods, in the three following situations:

- Maximum resolution of quadtree: 125m; Minimum resolution of quadtree: 125m; threshold: 17
- Maximum resolution of quadtree: 125m; Minimum resolution of quadtree: 250m; threshold: 17
- Maximum resolution of quadtree: 125m; Minimum resolution of quadtree: 250m; threshold: 17, with translations

These quadtrees have been constructed with data from the population register of Catalonia (1 January, 2014) in order to estimate the error that users would make in calculating populations in areas of their interest and the relation of errors with the parameters defining the quadtree.

For these experiments, two sets of geometries with random positions have been constructed:

- A set of 50,000 squares, with sides of 500m, 250m, 125m, 62m and 32m
- A set of 50,000 polygons, randomly created

Each element has been calculated for the relative error in the estimation of population X by comparing the initial layer of points in each quadtree Q:

$$\varepsilon_x = \frac{|p'_x - p_x|}{p_x}$$

where

$$p'_x = \sum_i^n p_i \frac{\text{AREA}(Q_i \cap X)}{\text{AREA}(Q_i)}$$

assuming that each element of the quadtree population is evenly distributed within the space.

3. RESULTS

The results of the solution proposed for dealing with the "border effect" can be seen in Table 1.

Table 1. Frequencies and population by size of the grid.

	With translation		Without translation		% (population)	
Size of the grid	**F**	**Population**	**F**	**Population**	**With translation**	**Without translation**
250 m	20,263	266,639	26,420	1,208,272	3.52	15.97
125m	51,744	7,299,825	31,.769	6,.358,.192	96.48	84.03
Total	72,007	7,566,464	58,189	7,566,464	100	100

As shown above, the untranslated results imply that nearly 16% of the population is positioned within a square of 250m. In the case of the translation, this percentage is only 3.52%.

It should be also noted that with this method the population actually translated is 64, 056 inhabitants, which is only 0.85% of the total population. It is, therefore, a data disturbance of a very limited scope.

The results of the error estimates can be seen in the following box plot:

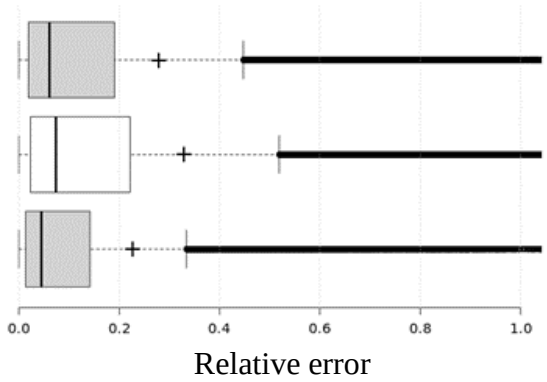

Figure 1. Box-plot for all cases of Monte Carlo simulations

As it can be seen, the worst results are achieved in the case of quadtree with resolutions of 125 and 250 m. The best results correspond to the quadtree of maximum resolution of 125m and minimum resolution of 125m. The quadtree with translations is neither the best nor the worst. We estimate that the median relative error of the estimation of population for the 125m and 250m quadtrees with translations is 5.3%. Our experiments also show that both the median and dispersion parameters of the relative errors increase as the surface area of the quadtree where the calculations are performed decreases.

4. CONCLUSIONS

This work aims to propose a solution for visualizing geocoded data while preserving confidentiality in areas with significant changes in population density. The solution proposed is the optimization of quadtrees by translating some individuals of the population between elements in the same hierarchy of the quadtree. In areas with dramatic changes in the distribution of population density, this solution prevents the excessive loss of information derived from the aggregation method and preserves statistical information. It also guarantees against the risk of statistical disclosure in the visualization of geocoded data without distorting the original distribution.

REFERENCES

[1] http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020

[2] Using quadtree representations in building stock visualization and analysis. Erdkunde. Vol. 67. Nº 2. 2013

[3] EEA reference grid - European Environment Agency.