SORT 41 (1) January-June 2017, 139-158

A quadtree approach based on European geographic grids: reconciling data privacy and accuracy

Raymond Lagonigro, Ramon Oller and Joan Carles Martori*

Abstract

Methods to preserve confidentiality when publishing geographic information conflict with the need to publish accurate data. The goal of this paper is to create a European geographic grid framework to disseminate statistical data over maps. We propose a methodology based on quadtree hierarchical geographic data structures. We create a varying size grid adapted to local area densities. High populated zones are disaggregated in small squares to allow dissemination of accurate data. Alternatively, information on low populated zones is published in big squares to avoid identification of individual data. The methodology has been applied to the 2014 population register data in Catalonia.

MSC: 62-09, 62-04, 62P25.

Keywords: Oficial statistics, confidentiality, disclosure limitation, dissemination, geographic information systems, hierarchical data structures, small area geography.

1. Introduction

One of the main goals of the national statistical agencies is to collect and distribute individual statistical data. Spatial data processing techniques allow gathering accurate geographic information for individuals along with other statistical data. As statistical agencies are totally concerned about confidential use of data, anonymity methods must be applied to ensure privacy.

Dissemination of geographic data gets in conflict with confidentiality, as individual data may be revealed. Classical anonymity techniques avoid the distribution of any piece of information that may lead to individual identification and so, geographic data might be eliminated before disseminating statistical data.

* Data Analysis and Modeling Research Group. Departament d'Economia i Empresa, Universitat de Vic - Universitat Central de Catalunya. Vic (Barcelona), Spain.

Received: June 2016
Accepted: January 2017

Besides, there is a strong demand by the research community for accessing geographic data to perform spatial data analysis. The use of geographic data to publish maps displaying statistical information is very engaging for national statistical agencies.

When publishing geographic information, it is important to develop techniques and methods that preserve confidentiality while at the same time, for spatial data analysis the information has to be reliable. Both goals are obviously conflicting, that is why the development of methodologies to balance them is so valuable.

The Statistical Institute of Catalonia (Idescat) has been using spatial data processing techniques to obtain geographic data for all individuals in the Catalonian 2014 population register data with the aim to provide high-quality and relevant statistical information. Idescat is interested on the automation of techniques to release and visualize the geospatial data as precise as possible without conflicting with confidentiality requirements.

We develop an automatic system to reconcile data privacy and accuracy on the distribution process of geographic social and economical data. The goal of our work is to create a framework, based on European standard grids, to distribute statistical data over on-line maps. The framework pursues the data accuracy at the best possible resolution, without individual information disclosure. Hierarchical geographic data structures are used to build a varying size grid adapted to local area densities.

In addition to the use of a varying size grid, a well known data protection method, our main contribution is the use of European standard based grids and a double threshold system to minimize information loss while maintaining data privacy.

The methodology has been implemented using the R software (R Core Team, 2015) and tested with the Catalonian 2014 population register data. The framework developed in this paper is to be used at the Catalonian Statistical Institute as a method to build and distribute statistical maps and data.

This paper is organized as follows: the second section reviews privacy concepts relying the visualization and delivery of geographical referenced data. The third section overviews the use of standard grids to deliver statistical data. It also introduces our methodology for creating a varying size grid based on the European standards for statistical data collection. The fourth section presents the results obtained applying a varying size grid quadtree methodology to the Catalonian 2014 population register data. Finally, the last section concludes and draws future research lines.

2. Privacy and geographic data

Privacy is an important issue when releasing confidential data about individuals. Many methods have been proposed to preserve privacy when delivering confidential information; Aldeen et al. (2015) present an extensive literature review and classification of privacy preserving on data analysis and publishing; Domingo-Ferrer et al. (2016) also provide a comprehensive overview of privacy models and anonymisation methods.

Concepts of privacy may vary in different countries (Exeter, Rodgers and Sabel, 2013), and the purpose of the research may also lead to different considerations on the level of confidentiality before delivering data to researchers (Fienberg, 1994; VanWey et al., 2005; Vilhuber, 2013). Many health research results, for instance, are worthless if the distributed information has been distorted. In such cases appropriate frameworks must be established between researchers and data custodians to enable information to be disseminated (Exeter et al., 2013).

Geospatial technologies brought on new aspects to be considered when dealing with privacy concepts for publishing georeferenced data (Armstrong and Ruggles, 2005). Two situations may arise when publishing spatial data. Individual geographic coordinates may be confidential by themselves and should be anonymised to avoid the identification of an individual location at a moment in time. This is, for instance, the case of location aware applications, where devices are used to know and handle the user's geographic position with some purpose (Beresford and Stajano, 2003). Location-based services (LBS), where information is sent to a user depending on his position must provide mechanisms for balancing high quality information services and privacy. Such mechanisms usually use obfuscation techniques and intermediation to mask the individual's locations and so ensure privacy and, at the same time, supply enough information to provide a service of satisfactory quality (Sweeney, 2002a; Duckham and Kulik, 2005; Ardagna et al., 2007; Xu and Cai, 2009). On the other hand, geographic coordinates may be published among many other attributes as part of statistical databases. Dissemination of spatial coordinates introduces disclosure risks that may lead to re-identification of individual private information and so should be anonymised (Armstrong and Ruggles, 2005; Curtis et al., 2006, 2011; Cassa et al., 2008; Boulos et al., 2009). For instance, individual addresses may be revealed through reverse geocoding when publishing maps with information on individual locations.

Armstrong et al. (1999) carry out a revision of existing strategies to deal with confidentiality of geocoded individual data. Zandbergen (2014) also compares the effectiveness of the different methodologies in terms of anonymisation and quality for spatial analyses of data being released. When individual data records are to be distributed, geomasking techniques should be applied. Geomasking techniques alter individually the position of each map point location to ensure the actual locations can't be discovered (Armstrong et al., 1999; Zimmerman et al., 2007; Hampton et al., 2010). Different types of random perturbations can be applied on every individual point so there is no coincidence between original and geomasked points. Some systems may also implement variable perturbations depending on the distribution of the risk of re-identification (Young et al., 2009). The perturbation applied directly increases the difficulty of re-identification, but simultaneously reduces the usefulness of the data to perform spatial analyses. In that sense, as many authors have stated (Duncan et al., 2001; Kwan et al., 2004; Reiter, 2012), altering the spatial distribution of individual data may introduce biases on further analyses and may make more difficult to ensure the results of such analyses are not compromised by the alteration method applied. Kwan et al. (2004)

demonstrate how data analyses provide very different results beyond some perturbation thresholds.

When the goal is map visualization, geomasking techniques do not seem to be the best approach, as they display points on a map that may not coincide with dwellings. In addition, geomasked locations may present paradoxical visual contradictions when being displayed on digital maps because they may fall into clearly uninhabitable zones (rivers, lakes, parks). Those effects may reduce confidence on the whole system. Zandbergen (2014) introduces spatial filters to ensure that masked locations do not fall into some special zones, therefore avoiding those undesired effects of geomasking.

As an alternative, aggregation methods avoid re-identification by grouping individual information into sets of summarized data. Microaggregation techniques pursue data privacy in microdata databases by aggregating individual records into small groups prior to publication (Defays and Anwar, 1998; Mateo Sanz and Domingo Ferrer, 1998). K-anonymity was defined as a basis model to provide guarantees of privacy protection for aggregation methods (Sweeney, 2002b). In k-anonymity models, a release is said to provide k-anonymity if the information for an individual cannot be distinguished from at least k–1 individuals within the same release. k-microaggregation techniques accomplish k-anonymity by creating groups with at least k records to prevent disclosure of individual information.

In the context of geographic data privacy, k-anonymity implies that for any individual in the dataset, his location is indistinguishable from at least k–1 other individuals. In geographic aggregation methods, individual information is aggregated into geographic areas to include and summarize in some way the individual data. Thus, to achieve k-anonymity, geographic aggregation methods must ensure that all the zones created in the aggregation process include at least k individuals (Vu et al., 2012).

Census tracts are a clear example of aggregated spatial data delivering and have been widely used as many information collection processes are related to these units. The area definition criteria is extremely important. Administrative or political boundaries, for instance, may not be a good choice to area definition as they often lead to incompatible statistical geographies (Walford, 2013). Several algorithms have been developed to automate the zone definition process creating uniform areas in terms of some defined variables (Openshaw, 1995, 1977; Martin, 2002; Ralphs and Ang, 2009). Duque et al. (2007) present a review of aggregation techniques to build optimized census regions.

The use of administrative boundaries on data aggregation methods may also lack of stability because they may change over time (Martin et al., 2002). Some methods have been developed to deal with zone matching problems which try to maintain the maximum number of areas without modification and hence maximising the match between different zonal geographies (Martin, 2003; Cockings et al., 2011).

Automated zone definition algorithms create aggregated data from bottom to up, grouping points or small areas to build bigger areas that meet a given criteria usually restricted to administrative boundaries. Grid-based aggregation is a good alternative as it does not rely on administrative areas (Giuliani et al., 2011; Tammilehto-Luode, 2011). Grid-based aggregations perform a top-down disaggregation methodology, in which areas are defined based on an initial rectangular grid (Steinnocher and Kaminger, 2010). Those aggregations offer a stable area system to disseminate data which will not change over time so that successive temporal spatial datasets will be comparable.

The quality of published aggregated information relies on the scale and the boundaries chosen. As stated by Openshaw (1984) in the modifiable area unit problem (MAUP) definition, results of spatial analyses are completely dependent on the spatial units being used to deploy them. This dependence is described by two factors: the scale of the aggregation units and the boundaries in which the area is divided to create those units (Flowerdew et al., 2001).

The modifiable area unit problem has been largely discussed in literature and many works have explored its influence on the results of analyses using aggregated data (Fotheringham and Wong, 1991; Cockings and Martin, 2005; Briant et al., 2010; Burden and Steel, 2013; Walford, 2013; Marceau, 2014). Briant et al. (2010) evaluate the magnitude of biases derived from the choice of the specific zoning system concluding that the size of the units matters much more than their shape. Andersson et al. (2012) results also reinforce the scale effect of the MAUP: they use a square grid zoning system at different resolutions stating that the scale effect is much more important than that of shape on the MAUP.

3. Data and methodology

Idescat has been gathering social and economical data at individual level and aims to publish such data using web maps. The Catalonian 2014 population data register includes 7,566,443 records with information such as sex, age, nationality, study level, along with the geographic coordinates of the individuals. Up to now, Idescat has been publishing such information in an aggregated form using census tracts to keep privacy. With an extension of 32,108 km^2, the Catalonia region is subdivided in 5040 census tracts. The use of administrative zones to gather spatial data makes difficult to perform analyses comparing information from different countries. Grid-based data collection is an alternative for comparable territorial statistics across countries and to perform time series statistical analyses (Tammilehto-Luode et al., 2003). Grid areas are time independent and extend over countries. In addition, in a grid system all the areas have the same form and size and can easily be divided into lower size area grids.

We propose to use a quadtree methodology based on an initial 1km^2 European grid and population data points to create a varying size grid. This methodology results in high populated zones being disaggregated in small squared areas and thus accurate data is disseminated, but at the same time low populated zones being maintained in big areas to avoid identification of individual data.

In the rest of the section we present our methodology to build a grid with much more precision than the census tracts without breaking the confidentiality.

3.1. A Hierarchical European standard grid

The European Forum for Geostatistics within the GEOSTAT project has promoted a common framework for grid based statistics across all the European countries. The GEOSTAT project collected experiences, methods and good practices from national statistical institutes and international bodies to encourage the production of grid data (GEOSTAT 1A, 2011). The project aims to represent various European characteristics in a 1km^2 grid dataset. With the results of the GEOSTAT 1A final report, the GEOSTAT 1B action developed the guidelines for datasets and methods to link the census 2011 based data with grid datasets (GEOSTAT 1B, 2014).

The GEOSTAT 1A project makes the following recommendation for the grid size: "Regarding the recommended size of grid cells, 1km^2 appears to be a good compromise between data availability, data confidentiality and suitability for national to European study areas. The project also recommends introducing intermediate grid sizes based on a two-level data structure (i.e. 250m and 500m as subdivision of the 1km grid)."(GEOSTAT 1A, 2011)

In the definition of the grid dataset, grid cells must be unambiguously identified to ensure a clear linkage between data and its geographic reference. In that sense, the INSPIRE grid coding system is adopted as defined in the INSPIRE Data specifications (INSPIRE, 2010). INSPIRE specifications use the European Terrestrial Reference System 89, Lambert Azimuthal Equal Area (ETRS89-LAEA) projection (Annoni et al., 2001). Each cell of the grid is identified by a code composed by the cell's size and the coordinates of the lower left cell corner in the ETRS89-LAEA system.

There are many techniques to represent hierarchical data structures. Hunter (1978) introduced the quadtree methodology to represent data used in cartographic applications, among others. This methodology is based on the principle of recursive decomposition of space (Samet, 1984, 1988). The region quadtree is based on the successive subdivision of space regions into four equal-sized quadrants, see Figure 1.

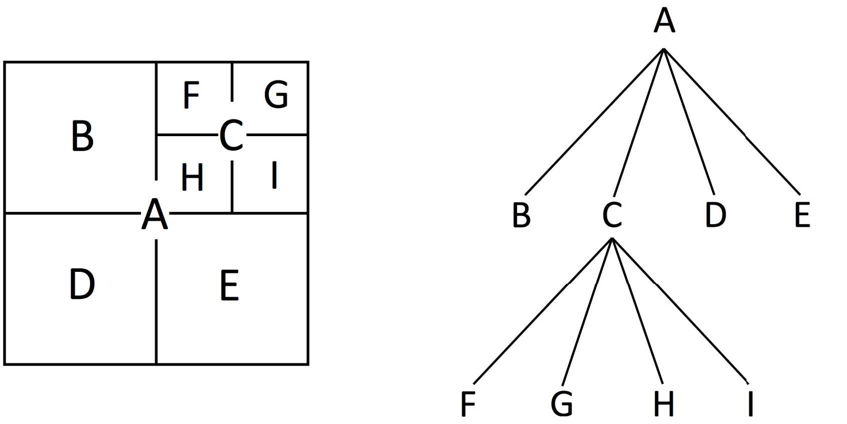


Figure 1: *Example of space subdivisions (left) and the equivalent quadtree representation.*

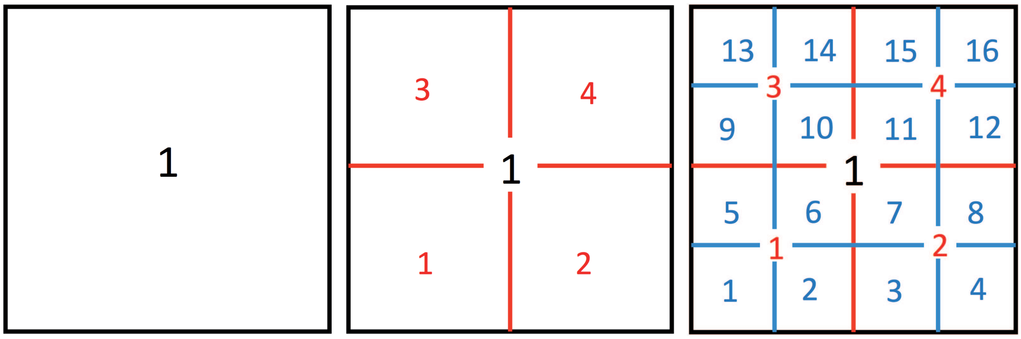


Figure 2: *Quadtree splitting example.*

The 1 km^2 grid can be subdivided into different size levels using a quadtree hierarchical structure in which cells of each size level can be easily linked. Quadtree methodology on spatial structures, can recursively split any squared area into four smaller areas, or, conversely, join four squared areas into the upper level one that they conform (Samet, 1984, 1988).

In a quadtree structure every cell of a grid level links to its containing cell on the upper grid level of larger size. Figure 2 shows an example of how an initial cell is split; each of the four new cells is sequentially numbered starting at the bottom left on left-right and bottom-up direction, and each of them will keep the number of the split cell. The same process may be successively repeated until the desired resolution. Moreover, any cell of the grid will keep its boundaries into a main European standard grid, so data can be easily related to other databases.

Our proposal requires the creation of a series of spatial data structures in order to get the final grid to work with. The initial 1km^2 grid is first intersected with the Catalonian boundaries to get the 32,915 squared grid cells that cover the whole 32,108 km^2 territory. The geographic points from the 2014 population database are then intersected with this grid to discard all cells without population, for matters of efficiency, before dividing the 1 km^2 grid into smaller ones. Once intersected the number of cells of the grid is reduced to 9,145. Using a quadtree structure this grid is divided to generate a new grid with double the resolution (i.e. half the scale) from which cells with no population will be deleted again. This process is repeated until a scale of 62.5 m^2 is reached and a set of 5 grids with 1 km^2, 500 m^2, 250 m^2, 125 m^2, 62.5m^2 squared cells is obtained.

The process to build the structures can be described as follows:

1. Build the initial 1 km^2 grid covering the while territory (in our case the Catalonian boundaries)
2. Set the minimum cell size to 62.5 m^2
3. While the actual grid cell size is bigger than the minimum defined:

   (a) Perform the intersection of each cell with point level information

(b) Aggregate individual information to each corresponding cell
(c) Discard grid cells with no points inside
(d) Subdivide each grid cell into four cells to build a new grid

3.2. A quadtree varying size grid

With the hierarchic spatial data structures created, the varying size grid is dynamically created performing a bottom-up aggregation considering a minimum threshold for the number of individuals in the cell.

Cells with less individuals than the threshold are automatically aggregated to the upper level in a quadtree manner. Given a threshold of k, none of the cell grids have less than k individuals as in a k-anonymity model.

An example of the varying size grid building process can be seen in Figure 3. The numbers in cells represent the number of individuals that meet a given criteria for that area. The grid on the left has several cells with values (in red) under the threshold (in the example threshold value is 10). Those cells are grouped to their upper level cell in the quadtree structure. The process is successively repeated until all cell values in the grid are over the threshold value or until the 1 km^2 cell size is reached in which case the cell is suppressed.

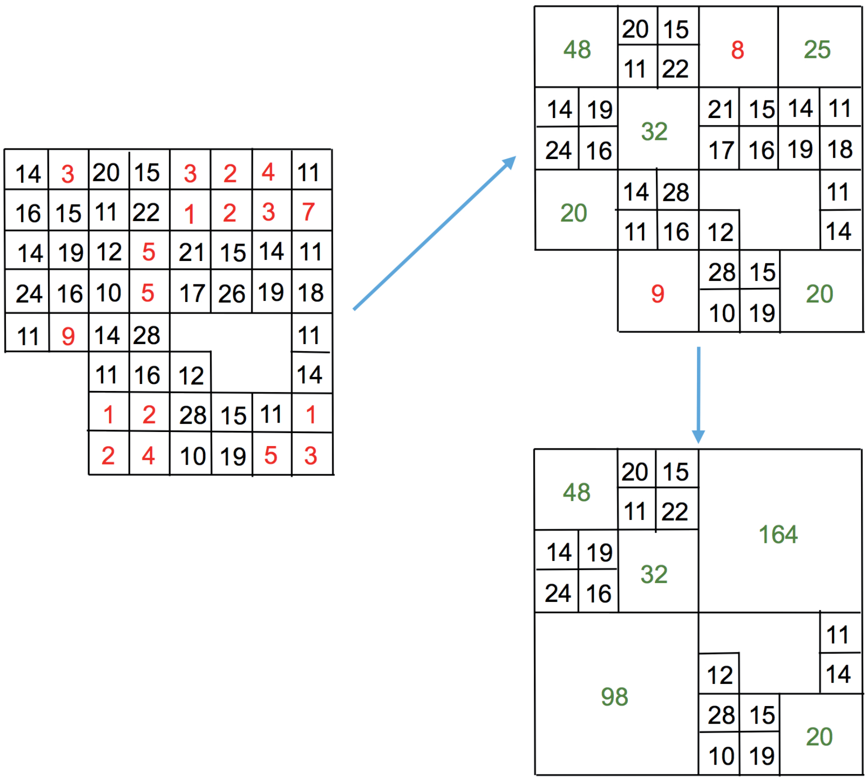


Figure 3: Building the varying size quadtree grid using a threshold value of 10.

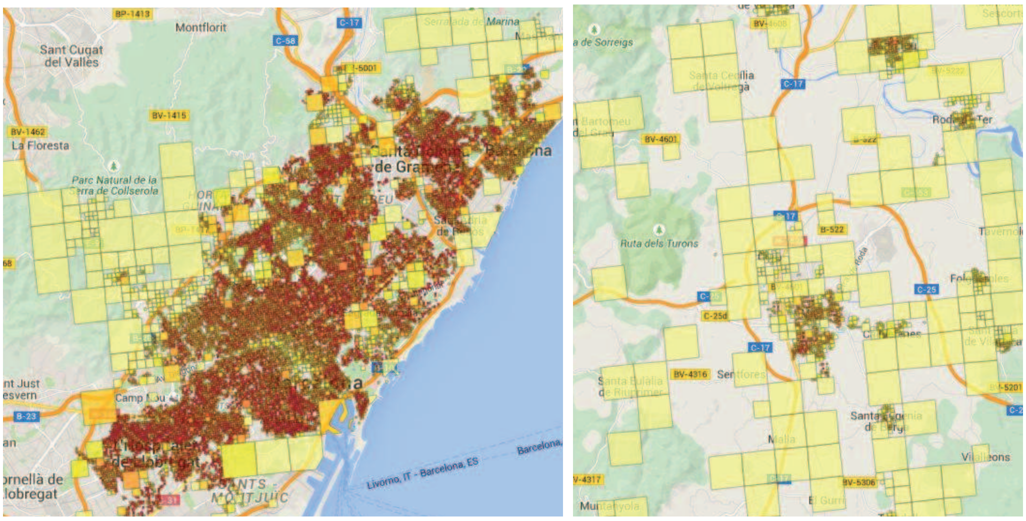


Figure 4: *Varying size quadtree grids examples. Urban region (left) and country side region (right). Color represent cell population density, ranging from yellow (low density) to red (high density).*

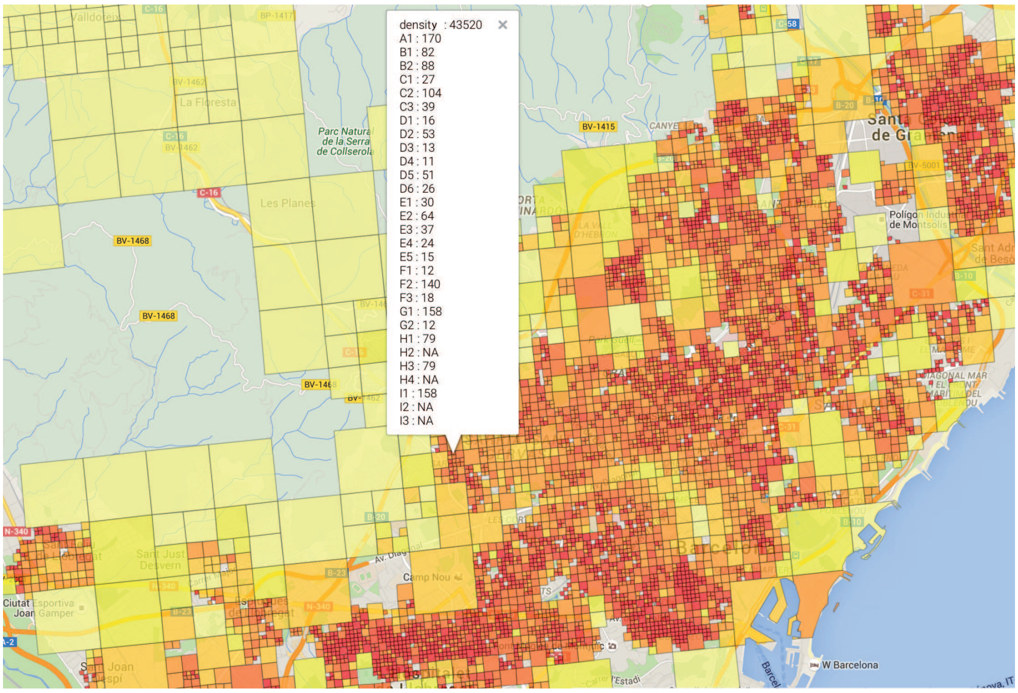
Two example zones of the varying size quadtree grid created using the proposed framework with a threshold of 100 can be seen in Figure 4.

3.3. A threshold for anonymity

The grid created ensures that a minimum number of individuals are clustered in each cell to avoid disclosure, but an anonymization process is still to be applied to the different variables available (i.e. sex, age, nationality). In the aggregation process, all the variables are summarized in each grid cell. Frequency counting was performed on the same categories of the qualitative and grouped quantitative variables being used by Idescat to deliver information for census tracts.

We propose to use a varying size quadtree methodology with a unique population threshold for cell aggregation (an aggregation threshold) and a second lower threshold for attribute anonymity (an anonymity threshold). A unique varying size grid was built using the quadtree structure and the population of each cell was used as the aggregation criteria. The rest of the attributes are qualified as "*Not Available*" when their values do not reach the anonymity threshold.

In Figure 5 an example of the resulting varying size grid created by using this methodology is shown. The example exhibits a zoomed zone around the Barcelona municipality where the different cell sizes is more apparent. A list of variables being published can also be seen; notice that several variables were anonymised to keep privacy as their values were below the anonymity threshold value. The anonymity threshold ensures k-anonymity, as for every cell, attributes whose values do not reach the threshold are not shown.



```
density  : 43520
A1 : 170
B1 : 82
B2 : 88
C1 : 27
C2 : 104
C3 : 39
D1 : 16
D2 : 53
D3 : 13
D4 : 11
D5 : 51
D6 : 26
E1 : 30
E2 : 64
E3 : 37
E4 : 24
E5 : 15
F1 : 12
F2 : 140
F3 : 18
G1 : 158
G2 : 12
H1 : 79
H2 : NA
H3 : 79
H4 : NA
I1 : 158
I2 : NA
I3 : NA
```

Figure 5: *Final varying size quadtree grid obtained with aggregation and anonymity threshold values of 100 and 10 respectively. The list of variables for a cell are shown.*

The use of a unique grid for all the dataset attributes also ensures *reciprocity*. Kalnis et al. (2007) introduced reciprocity as a sufficient property for spatial k-anonymity. Reciprocity requires that a set of individuals are always grouped together for a given k. The unique grid achieves reciprocity as the grouping is the same for any variable or variable crossing.

4. Evaluation of the varying size quadtree

To test the methodology explained in the previous section a Web server interface was implemented using the plotGooglemaps R package (Kilibarda, 2015) to project the grid over Google maps (Miller, 2006); the Rook R package (Horner, 2014) was used to dynamically serve the maps in web pages. An anonymity threshold value of 10, as recommended by GEOSTAT 1B (2013), was chosen to avoid disclosure and several aggregation thresholds were assessed. Table 1 shows the results obtained using three different aggregation thresholds. We can observe an obvious inverse relation between the aggregation threshold and the rate of values that do not reach the anonymity threshold and thus are considered *NA* values. A lower aggregation threshold yields a better resolution grid, as more small areas are maintained, but less information will be available for those areas.

Table 1: *Different population aggregation thresholds, with 62.5 m^2 cells rate and NA values rate.*

	62.5 m^2 cells rate	*NA* values rate
Aggregation threshold 30	58.1%	29.4%
Aggregation threshold 50	48.6%	21.8%
Aggregation threshold 100	30.8%	11.9%

From the results shown on Table 1 we propose an aggregation threshold value of 100 so that 30.8% of high resolution cells are maintained on the final grid, with a rate of *Not Available* of only 11.9%.

To reduce information loss, we slightly modify the algorithm so that only 1 km^2 cells with population under the anonymity threshold are suppressed. For those 1 km^2 cells with population between the anonymity and aggregation thresholds (i.e. between 10 and 100), the population number of each cell is disclosed and the rest of the variables are anonymised if needed. In the case of the Catalonian 2014 population register only 0.17% of individuals are removed from the resulting set.

A varying size grid adapts its resolution to the population distribution. The use of a fixed size grid, although may be more stable, as the cells are equally sized, introduces accuracy loss when cells are big, or information loss when cells are small. Regarding

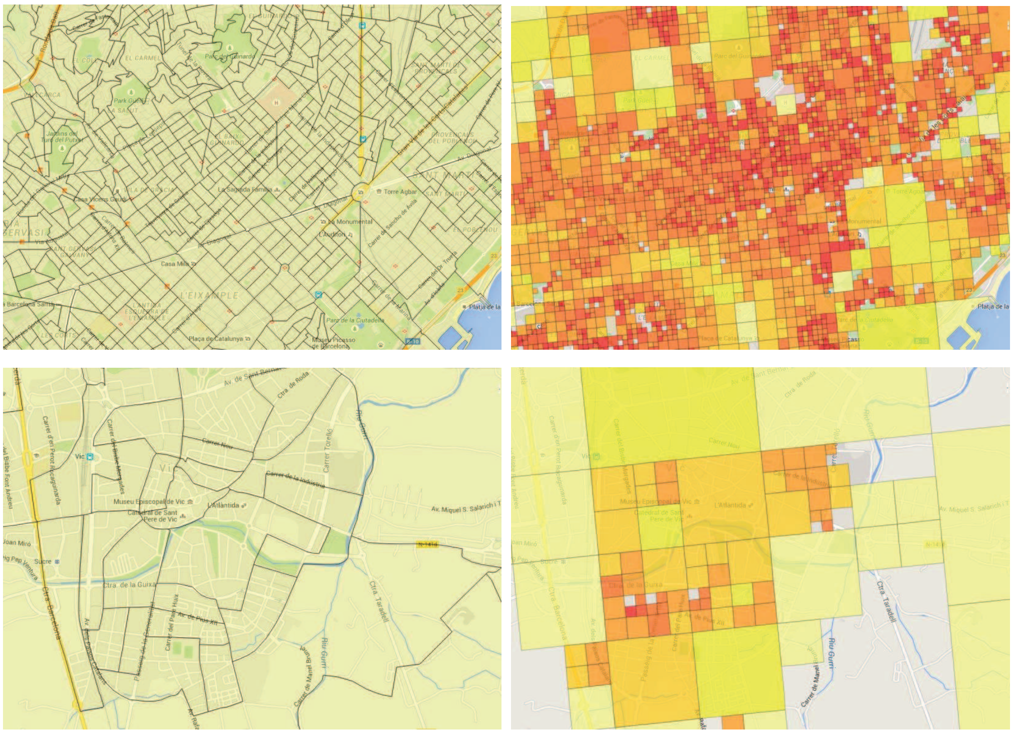


Figure 6: *Census sections (left) compared to varying size quadtree grid (right).*

accuracy, if we compare our varying size grid to a 1 km^2 fixed grid, 45.3% of the cells are subdivided in smaller ones using our methodology. The varying size adapts the cell size to the population distribution delivering information closer to the actual individual point information. In our grid, 30.8% of the cells are high resolution ones (i.e. 62.5 m^2); the grid also maintains 33.1% of 125 m^2, 12.9% of 250m^2 and 10.2% of 500 m^2 cells and only 13.1% are 1 km^2 cells. The total number of *Not Available* values introduced by the disaggregation is only 0.04% higher compared to the 1 km^2 fixed grid. On the other hand, using a 62.5 m^2 fixed grid all the cells with population under the threshold would be deleted to keep privacy, the total population loss would be 45.5% compared to the 0.17% in our methodology.

Hitherto, the Catalonian Institute of Statistics, as most of the public statistical agencies, has been using census tracts as spatial units to collect and distribute statistical data. In Figure 6 those administrative census sections are compared to the quadtree varying grid size for two different zones. Both systems offer varying size areas, but while census sections have administrative boundaries, grid cells keep the boundaries within standard grids. Moreover, the size of the automatically created cells of the grid reach higher level of resolution on high population density areas and therefore, more accurate data are being published.

High resolution spatial zones can also be created using automated zoning systems. AZTool (Martin, 2003; Cockings et al., 2011) was used with the Catalonian 2014 population register to compare the resulting areas with those created with our methodology. AZTool algorithm applies a bottom-up methodology aggregating points into uniform areas in terms of some defined constraints. To get comparable results, a population minimum threshold of 100 was used as the aggregation criteria. Figure 7 shows an example of the resulting areas created using the AZTool algorithm compared to the varying size quadtree grid. The AZTool algorithm also creates a zoning system adapted to population density but irregular polygons are obtained. Our quadtree proposal creates squared cells and, though also generating a varying size grid, due to its hierarchical structure, can be compared at some scale because the boundaries are always within the main grid. As a grid system, it offers more comparable spatial and time serial statistics (Tammilehto-Luode, 2011).

The methodology proposed performs aggregation as an automatic non-disclosure control system, but some aggregations must be examined and treated specially to avoid information distortion. When neighbour zones with highly different population densities are within the same 1 km^2 unit cell, the minimum threshold may lead to an aggregation of many high resolution cells, loosing accuracy and even distorting local densities. For instance, consider the example shown in Figure 8 of an urban zone in a city, with a 1 km^2 cell including a high populated neighbourhood and other sub-zones with few population. When applying the proposed methodology, the two top left cells should be protected for disclosure, but none of the intermediate aggregations reach the minimum level of anonymity so only the 1 km^2 cell can be displayed and thus, population in the cell is shown as if it was equally distributed.

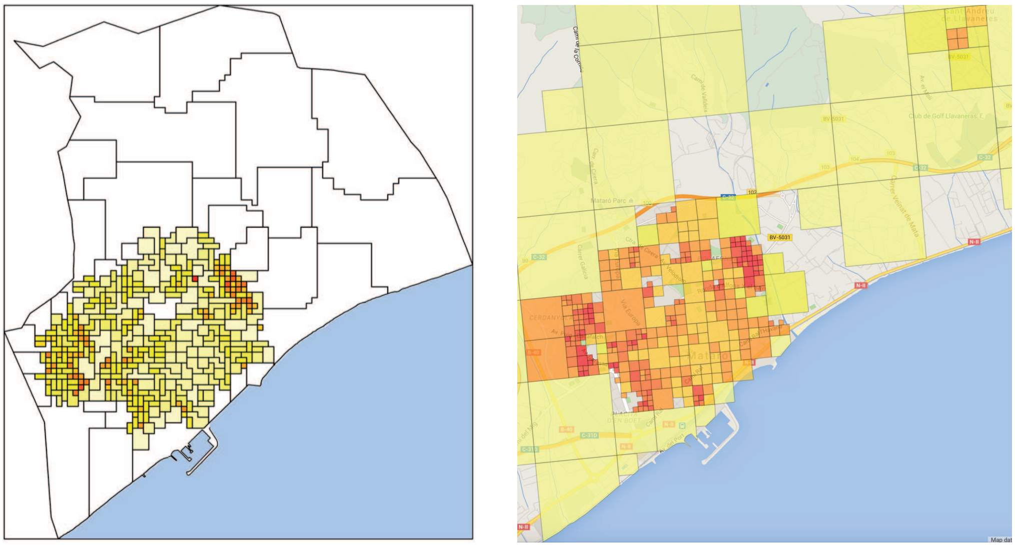


Figure 7: *AZTool zones (left), varying size quadtree grid (right).*

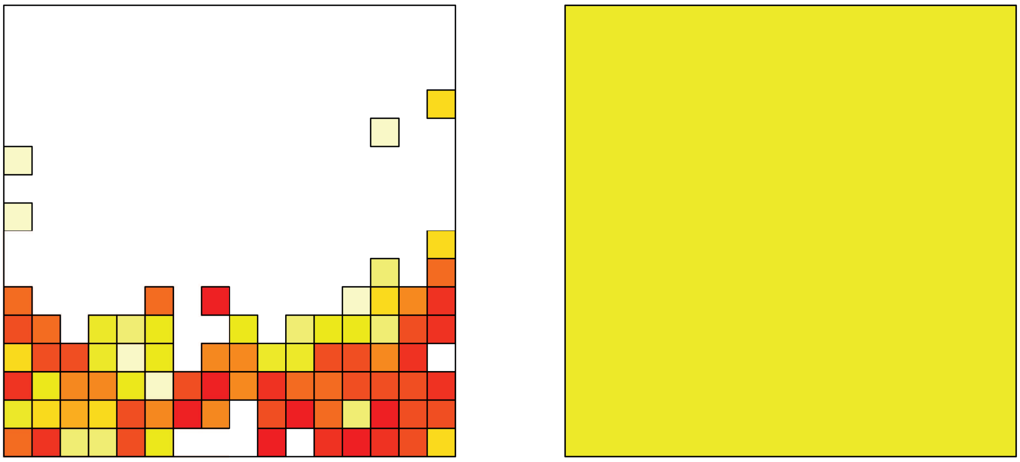


Figure 8: *Undesired aggregation effect example. Original 62.5 m^2 cells (left); resulting 1 km^2 cell being published (right).*

An analysis of the resulting grid was performed, focusing on those extreme aggregations. There are 13% of 1 km^2 cells in the final grid. From those, 14% contain one or more 62.5 m^2 cells with population over the aggregation threshold. Graph in Figure 9 represents the frequency distribution of the number of 62.5 m^2 cells over the aggregation threshold for each of the 1 km^2 cells. There are few extreme cases in which many cells are aggregated.

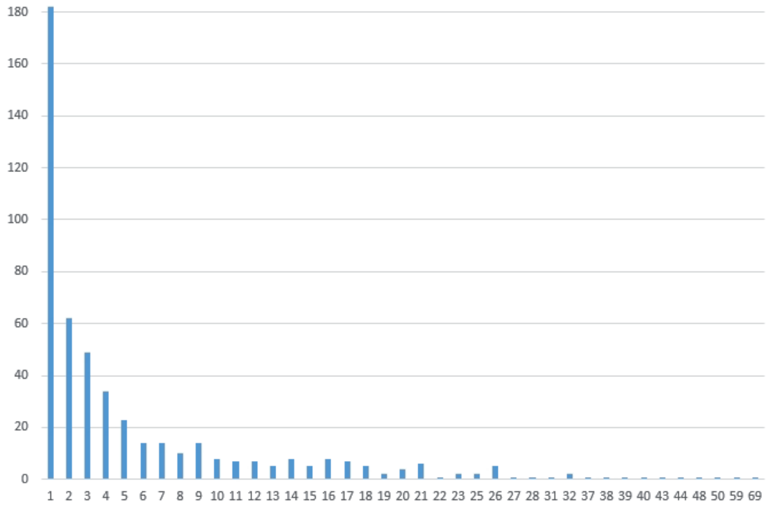


Figure 9: *Number of 62.5 m^2 cells over threshold being aggregated into 1 km^2 cells.*

If we focus on 500 m^2 cells, there are 10% on the final grid of which 25% contain one or more 62.5 m^2 cells with population over the aggregation threshold. Again, the number of extreme cases in which this phenomena is observed is very small.

The quadtree algorithm can be slightly modified to improve those extreme cases where low populated cells lead to aggregation of high populated ones (see Figure 9). The distortions previously discussed result from heterogeneous regions with cells where none of the in-between aggregations reach the threshold and the 1km^2 is then produced hiding cells over the threshold. Removing under-threshold cells from the quadtree process when over-threshold cells are available in the same 1 km^2 leads to a more precise quadtree. The total population of the cells removed from the process can be given within a randomly selected cell, specially denoted for that purpose. Information on that cell could be anonymised in the same way as other cells.

Figure 10 exhibits the example region previously seen on Figure 8; numbers in cells denote population in each cell; dark coloured cells have less population than the threshold value (a threshold value of 100 is being used). If we look at the two top left cells, they can only be covered in a 1 km^2 aggregation. Removing them from the process, leads to a grid with much more high resolution cells and so more accurate information will be unleashed. The green cell in Figure 10 is randomly selected to denote the total population of cells under the population threshold removed from the aggregation process.

The same process can be applied to consider 500 m^2 or 250 m^2 cells. When the resulting grids lead to loss of information in the sense of grouping a number of high populated cells, the aggregation may be prevented taking low populated cells out of the quadtree generation process. The grid generated is improved, with more high resolution cells being published but, again, avoiding disclosure risk.

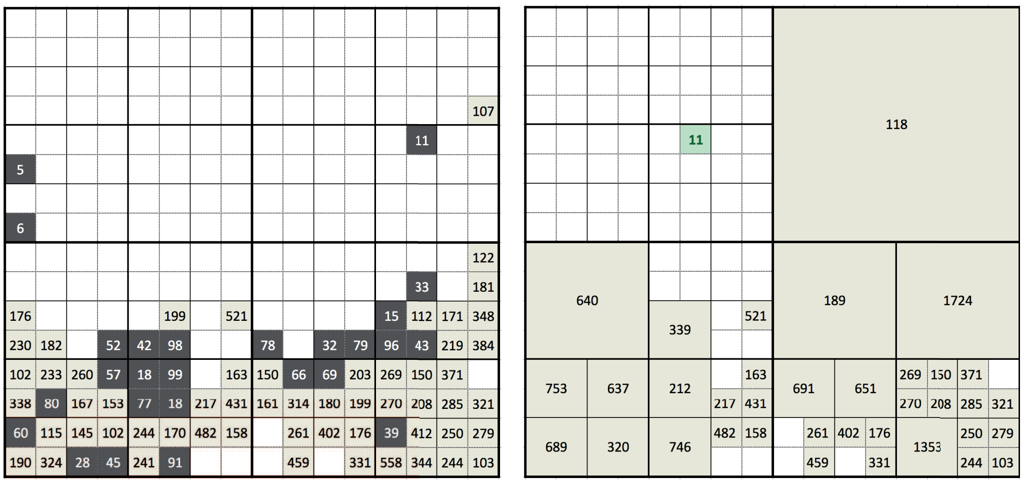


Figure 10: *Original 62.5 m^2 cells with population on each cell (left) and the resulting quadtree produced with the final method proposed (right).*

5. Conclusions

This study sets out a methodology to create a unique varying size grid for the dissemination of statistical geographic data. We use quadtree hierarchical data structure to recursively subdivide space regions from a standard grid.

The methodology proposed implements an automatic privacy control process by adapting the grid cell size based on the population density of the regions represented in the grid. Low resolution cells are produced on low populated zones, while high resolution cells are maintained for high populated zones. Thus, small grid cells are made available when possible, while controlling data disclosure at the same time.

One of the major advantages of grid based statistics is the stability due to the fact that all the areas have the same form and size (Tammilehto-Luode et al., 2003) and they do not change over time. The methodology proposed, indeed, introduces variability on the area size. The varying size grid created completely depends on the dataset being used, here the 2014 population register. The stability of the system, however, is maintained by the reference grid being used. Any grid created using the proposed methodology can be compared at some scale, as the boundaries are always within a main grid. Moreover, as a hierarchical grid system, it can easily be aggregated into a fixed size grid on a lower resolution level.

Up to now, geographic statistical data is restricted to census sections which change over time. The grid proposed will allow publication of more stable and accurate data as the grid reference is fixed, and cell dimensions are much smaller than census sections.

The methodology has been implemented using the R software and applied to the Catalonian 2014 population register data. The resulting grid has been dynamically represented and tested on a web server using Google maps.

Our proposal may also be applied on map visualization of statistical data, as a response to users zooming actions. The different grid levels can be associated to the different scales of zoom thus applying automatic aggregation of information based on the zoom level and the quadtree.

Acknowledgements

This research was partially supported by the Institut d'Estadística de Catalunya (Idescat). The authors gratefully thank Eduard Suñé and Mireia Farré for his valuable suggestions on previous drafts of the paper. We give our gratitude to Rafa Madariaga who has given us his precious critique and proofreading. We thank the Data Analysis and Modelling research group for the fruitful discussions we held together. The work was also supported by grant MTM2015-64465-C2-1-R of the Ministerio de Economía y Competividad (Spain).

References

Aldeen, Y.A. A.S., Salleh, M. and Razzaque, M.A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*,4, 1–36.

Andersson, M., Klaesson, J. and Larsson, J. P. (2012). How local are spatial density externalities? Evidence from square grid data. Technical report, (No. 2012/10). Lund University, CIRCLE-Center for Innovation, Research and Competences in the Learning Economy.

Annoni, A., Luzet, C., Gubler, E. and Ihde, J. (2001). Map projections for Europe. *Report EUR 20120.*

Ardagna, C. A., Cremonini, M., Damiani, E., Di Vimercati, S. D. C. and Samarati, P. (2007). Location privacy protection through obfuscation-based techniques. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 47–60. Springer.

Armstrong, M. P. and Ruggles, A. J. (2005). Geographic information technologies and personal privacy. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 40, 63–73.

Armstrong, M. P., Rushton, G. and Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine*, 18, 497–525.

Beresford, A. R. and Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive computing*, 2, 46–55.

Boulos, M.N.K., Curtis, A. J. and AbdelMalik, P. (2009). Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics*, 8, 46–54.

Briant, A., Combes, P.-P. and Lafourcade, M. (2010). Dots to boxes: do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics*, 67, 287–302.

Burden, S. and Steel, D. (2013). Characteristics of empirical zoning distributions for small area health data. *Working Paper 15-13*, University of Wollongong.

Cassa, C. A., Wieland, S. C. and Mandl, K. D. (2008). Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics*, 7, 45–54.

Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2011). Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales. *Environment and Planning A*, 43, 2399–2418.

Cockings, S. and Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science & Medicine*, 60, 2729–2742.

Curtis, A., Mills, J. W., Agustin, L. and Cockburn, M. (2011). Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems*, 35, 57–64.

Curtis, A. J., Mills, J. W. and Leitner, M. (2006). Spatial confidentiality and GIS: re-engineering mortality locations from published maps about hurricane Katrina. *International Journal of Health Geographics*, 5, 44–66.

Defays, D. and Anwar, M. N. (1998). Masking microdata using micro-aggregation. *Journal Of Official Statistics-Stockholm*, 14, 449–462.

Domingo-Ferrer, J., Sánchez, D. and Soria-Comas, J. (2016). Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections. *Synthesis Lectures on Information Security, Privacy & Trust*, 8, 1–136.

Duckham, M. and Kulik, L. (2005). A formal model of obfuscation and negotiation for location privacy. In *International Conference on Pervasive Computing*, pp. 152–170. Springer.

Duncan, G. T., Keller-McNulty, S. A. and Stokes, S. L. (2001). Disclosure risk vs. data utility: the R-U confidentiality map. Technical report, Chance.

Duque, J. C., Ramos, R. and Suriñach, J. (2007). Supervised regionalization methods: a survey. *International Regional Science Review*, 30, 195–220.

Exeter, D. J., Rodgers, S. and Sabel, C. E. (2013). "Whose data is it anyway?" The implications of putting small area-level health and social data online. *Health Policy*, 114, 88–96.

Fienberg, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, 10, 115.

Flowerdew, R., Geddes, A. and Green, M. (2001). Behaviour of regression models under random aggregation. In P. M. A. Nicholas J. Tate (Ed.), *Modelling Scale in Geographical Information Science*, pp. 89–104. Wiley: Chichester, UK.

Fotheringham, A. S. and Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23, 1025–1044.

GEOSTAT 1A (2011). ESSnet project GEOSTAT 1A-representing census data in a European population grid. Technical report, The European Forum for GeoStatistics.

GEOSTAT 1B (2013). ESSnet project GEOSTAT 1B-representing 2011 census data on grid. Technical report, The European Forum for GeoStatistics.

GEOSTAT 1B (2014). ESSnet project GEOSTAT 1B-representing census data in a European population grid. Technical report, The European Forum for GeoStatistics.

Giuliani, G., Ray, N. and Lehmann, A. (2011). Grid-enabled spatial data infrastructure for environmental sciences: challenges and opportunities. *Future Generation Computer Systems*, 27, 292–303.

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I.A., Gesink, D.C., Leone, P.A., Serre, M.L. and Miller, W. C. (2010). Mapping health data: improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, 172, 1062–9.

Horner, J. (2014). *Rook: Rook - a web server interface for R*. R package version 1.1-1.

Hunter, G. M. (1978). *Efficient Computation and Data Structures for Graphics*. Ph. D. thesis, Princeton, NJ, USA.

INSPIRE (2010). INSPIRE Specification on Geographical Grid Systems - Guidelines (D2.8.I.2). Technical report, INSPIRE Infrastructure for Spatial Information in Europe: European Commission.

Kalnis, P., Ghinita, G., Mouratidis, K. and Papadias, D. (2007). Preventing location-based identity inference in anonymous spatial queries. *IEEE transactions on knowledge and data engineering*, 19, 1719–1733.

Kilibarda, M. (2015). *plotGoogleMaps: Plot Spatial or Spatio-Temporal Data Over Google Maps*. R package version 2.2.

Kwan, M.-P., Casas, I. and Schmitz, B. C. (2004). Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39, 15–28.

Marceau, D. (2014). The scale issue in the social and natural sciences. *Canadian Journal of Remote Sensing*, 25, 347–356.

Martin, D. (2002). Geography for the 2001 Census in England and Wales: an overview of the geography system used in the 2001 census, primarily the creation of output geography. *Population Trends Summer*, 108, 7–15.

Martin, D. (2003). Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information*, 17, 181–196.

Martin, D., Dorling, D. and Mitchell, R. (2002). Linking censuses through time: problems and solutions. *Area*, 34, 82–91.

Mateo Sanz, J. M. and Domingo Ferrer, J. (1998). A comparative study of microaggregation methods. *Qüestiió*, 22, 511–526.

Miller, C. C. (2006). A beast in the field: the Google Maps mashup as GIS/2. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 41, 187–199.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers*, 2, 459–472.

Openshaw, S. (1984). The modifiable area unit problem. *Concepts and Techniques in Modern Geography*, 38, 1–41.

Openshaw, S. (1995). Algorithms for reengineering 1991 census geography. *Environment and Planning A*, 27, 425–446.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ralphs, M. and Ang, L. (2009). *Optimised Geographies for Data Reporting: Zone Design Tools for Census Output Geographies*. Number 09 in Statistics New Zealand. Wellington: Statistics New Zealand.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly*, 76, 163–181.

Samet, H. (1984). The quadtree and related hierarchical data structures. *ACM Computing Surveys*, 16, 187–260.

Samet, H. (1988). An Overview of quadtrees, octrees, and related hierarchical data structures. In R. Earnshaw (Ed.), *Theoretical Foundations of Computer Graphics and CAD*, Volume 40 of *NATO ASI Series*, pp. 51–68. Berlin, Heidelberg: Springer Berlin Heidelberg.

Steinnocher, K. and Kaminger, I. (2010). Gridded population-new data sets for an improved disaggregation approach. *European Forum for Geostatistics Workshop*.

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 571–588.

Sweeney, L. (2002b). k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 557–570.

Tammilehto-Luode, M. (2011). Opportunities and challenges of grid-based statistics. *World Statistics Congress of the International Statistical Institute*.

Tammilehto-Luode, M., Ralphs, M. and Backer, L. (2003). Tandem II: towards a common geographical base for statistics across Europe. The final report. *Unpublished. Eurostat. Luxembourg*.

VanWey, L. K., Rindfuss, R.R., Gutmann, M.P., Entwisle, B. and Balk, D. L. (2005). Confidentiality and spatially explicit data: concerns and challenges. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15337–15342.

Vilhuber, L. (2013). Methods for Protecting the Confidentiality of Firm- Level Data: Issues and Solutions.

Vu, K., Zheng, R. and Gao, J. (2012). Efficient algorithms for k-anonymous location privacy in participatory sensing. In *INFOCOM, 2012 Proceedings IEEE*, pp. 2399–2407. IEEE.

Walford, N. (2013). Development and design of a web-based interface to address geographical incompatibility in spatial units. *Environment and Planning A*, 45, 1713–1733.

Xu, T. and Cai, Y. (2009). Feeling-based location privacy protection for location-based services. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09, New York, NY, USA, pp. 348–357. ACM.

Young, C., Martin, D. and Skinner, C. (2009). Geographically intelligent disclosure control for flexible aggregation of census data. *International Journal of Geographical Information Science*, 23, 457–482.

Zandbergen, P. A. (2014). Ensuring confidentiality of Geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in Medicine*, 2014.

Zimmerman, D. L., Armstrong, M.P. and Rushton, G. (2007). Alternative techniques for masking geographic detail to protect privacy. In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*, Chapter 7, pp. 127–138. Boca Raton, Fla.: CRC Press.