14. Confidentiality of spatial data

MAËL-LUC BURON, MAËLLE FONTAINE
INSEE

Abstract

Recent profusion of geocoded sources, often released as grid data, allows many possibilities of analysis for economists, demographists, or sociologists. But it also leads to high disclosure risk, because the number of variables necessary to uniquely identify someone considerably decreases when the intruder knows the geographical location at a detailed level. This issue is even more serious in areas with low population density. Traditionally, Statistical Disclosure Control (SDC) methods do not take spatial features of data into account. This chapter aims at giving suggestions to measure and to deal with disclosure risk, while preserving spatial correlations. Pre-tabular methods seem to be more appropriate to the purpose because they can target the riskiest records, with a measure of risk depending on the local context. But applying only pre-tabular methods cannot lead to a sufficient level of protection, and post-tabular methods can be performed in a second step to guarantee respect of national regulations. In this chapter, we give an overview of the existing literature and we make a focus on a specific SDC method called Targeted Record Swapping (TRS), highlighted by the Eurostat Grant "*Harmonised Protection of Census Data in the ESS*". This method detects the records most exposed to disclosure and swaps them with other similar records belonging to the same geographical neighbourhood. Therefore, individuals with rare attributes continue to be present in the data, but not at their actual location. This ensures that an intruder cannot re-identify them with certainty. We test TRS on French fiscal data for a small region and for several

parameters, and we obtain very little distortion of spatial correlations for variables taken into account in the method or strongly correlated to them.

Prior reading of Chapters 2: "Codifying the neighbourhood structure" and 3 "Spatial autocorrelation indices" is recommended.

Introduction

The analytical richness of spatial data and how they can explain underlying phenomenons has been widely commented in previous chapters. Analytical tools to take advantage of this wealth of information have also been presented.

In the near future, more and more data will be geolocated, leading to a profusion of information available at a very detailed geographical level and giving to economists and analysts many topics to explore. But this profusion also leads to crucial stakes concerning confidentiality of spatial data. The number of characteristics necessary to uniquely identify an observation decreases with the size of the mesh within which information is released, and even more in the context of proliferation of open access geographical visualisation tools. When population density is low in a given area, disclosure risk increases, because the probability to find someone else similar in the neighbourhood is low.

The situation is a conflict between two great principles of public statistics releasing (VanWey et al. 2005). On the one hand, National Statistical Institutes (NSIs) have the vocation of offering as much data as possible with a high level of utility, and on the other hand, they have to manage with strong constraints to guarantee and enforce the confidentiality of information providers. Ensuring confidentiality of spatial data is indeed a particularly difficult task because European and national regulations prohibit NSIs from disseminating any data that could allow an intruder to identify, directly or indirectly, the investigated household or company, and scrupulously that would mean, in most cases, not to release anything. Finally, any data releasing means a non-zero disclosure risk, and the stake is to reduce it to a low and acceptable level. In other words, data protection strategy can be seen as a trade-off between minimising disclosure risk and maximising data utility.

This chapter is not written from the point of view of the user of confidential data; it is written from the point of view of the expert in statistical disclosure control, whose task is to disseminate data satisfying statistical confidentiality regulations, conditionally on some output strategies. Typically, he has a file of individual data (micro-data), and he has to disseminate tabular data for small regional breakdowns or grid data from it, but it is forbidden to disseminate a statistic if it concerns less than a certain threshold of observations. In what follows, a record stands for a household, an individual, or an organisation. We assume the micro-data to be exhaustive: the methods presented are not valid for survey data.

Section 14.1 introduces the disclosure risk: how it can be defined in case of spatial data, and what recommendations can be made to detect high risk observations. Eurostat published in 2007 a Handbook on statistical control (second version in 2010) about standard methods used to deal with confidentiality issues, but spatial data requires adaptations regarding confidentiality treatment. Section 14.2 gives an overview of different methods to deal with disclosure risk for spatial data, including recommendations about risk-utility analysis. Section 14.3 presents the results of some pre-tabular methods tested on a French region, in the context of grid data releasing. For these tests, differencing issues with administrative zoning are not treated, but Section 14.4 will specifically focus on the subject.

14.1 How to evaluate spatial disclosure risk?

14.1.1 General definition of disclosure risk

Maintaining privacy is essential to retaining the trust of providers, with in the background, the fear of falling response rates. A respondent to a survey must be confident that his personal information will be safeguarded. European regulations have been written to coerce confidentiality [1]: according to Article 20, Chapter V of Commission Regulation No 223/2009 on European Statistics: "*Within their respective spheres of competence, the NSIs and other national authorities and the Commission (Eurostat) shall take all necessary regulatory, administrative, technical and organisational measures to ensure the physical and logical protection of confidential data (statistical disclosure control).*". Countries also have their own regulations. Confidentiality constraints usually take the form of adequate thresholds: no information can be disclosed if it concerns less than a given number of observations. The choice of thresholds depends on various parameters: sparsity of the area, risk aversion, sensitivity of variables, future data users. Sometimes, guidelines are available to check if the data file complies with the confidentiality rules (ONS 2006, Insee 2010).

Disclosure occurs when an intruder (also called "data snooper" in some articles) uses released data to learn some information he did not already know. The intruder is not a hacker. He just has released data at his disposal, and he does not attempt to break any security system. A distinction can be made between different disclosure scenarios (Duncan et al. 1986, Lambert 1993, Clifton et al. 2012 [2], Bergeat 2016):

— **identity disclosure** occurs when a direct identifier of a statistical individual (company, household or person) can be found thanks to the released data (for example, it can be easy to identify the company of a sector with the most important turnover);

— **attribute disclosure** occurs when the intruder can reveal some association between a respondent and some of his sensitive variables ("quasi-identifiers"). Identity disclosure always implies attribute disclosure but the opposite is not true. For example, if the intruder knows someone living in an area, and if the released data show that all the inhabitants of this area share a common characteristic, then he can deduce that the individual has the characteristic, even if he cannot deduce his other attributes;

— **inferential disclosure** occurs when an intruder can infer some attribute with high confidence. Generally, this type of disclosure is not taken into account to protect a dataset.

To comply with strict regulations, one approach is to consider different kinds of users. General users will only have access to less information (less variables or larger categories), whereas specific public like researchers will have restricted access to more data in secure centres, if they previously justify their request by some procedure.

A complementary approach is to introduce perturbation in the data, in order to reach an acceptable level of disclosure risk. Applying a statistical disclosure control (SDC) method then consists in reducing data utility in exchange of more protection. Traditionally, SDC methods do not take spatial features into account, and spatial correlations can be very distorted before and after perturbation. The next subsection gives arguments in favour of geographically intelligent SDC strategies.

14.1.2 Specificities of spatial data regarding disclosure risk

Disclosure control experts dealing with spatial data face a paradox. On the one hand, such data need more protection because they permit more identifications, but on the other hand they offer many possibilities of analysis that users do not want to distort too much.

1. With equivalents outside Europe like Australian Privacy Act in 1988
2. Clifton et al. 2012 makes a classification of different disclosure risks and defines first-tier, second-tier et third-tier.

Theoretical considerations

In the handbook of SDC guidelines from Eurostat (Hundepool et al. 2010), three different levels of quasi-identifiers are suggested. Only geographical location is considered in the category of extremely identifying variables. Disclosure risk is indeed higher when considering spatial data, for several reasons.

Firstly, the risk of identity disclosure increases in presence of spatial data because it is easier to mobilise personal knowledge. Indeed, among the characteristics possibly shared with someone (age, gender, etc.), belonging to a neighbourhood leads to a higher probability of personally knowing the person. Moreover, it has recently become possible to identify addresses with the development of web scraping or open access tools like *Google Earth*, that make possible re-engineering (Curtis et al. 2006) or direct identification (Elliot et al. 2014). As a result, population density is a fundamental predictor of disclosure risk: the lower the density, the higher the disclosure risk.

Secondly, the risk of attribute disclosure increases in case of spatial data because of Tobler's first law of geography, which states that "*everything interacts with everything, but two close objects are more likely to do so than two distant objects*". As a result, the degree of dissimilarity of an individual to his neighbours seems to be another good predictor of disclosure risk.

And finally, disclosure risk increases with the differencing issue. When data is disseminated in different non-nested geographies (typically administrative borders and grid of squares), in some cases, someone's attribute can be deduced by subtracting the counting of an area from the counting of another enclosing area. Therefore, anyone proficient with geographic information systems (and subtractions) becomes a potential intruder. This particular question of geographic differencing is the topic of Section 14.4.

Technical considerations

Technically, the dissemination classification (zoning, administrative boundaries, or regular lattice like a grid of squares), is a categorical variable like any other (additional dimension of tabulated data). It is therefore possible, with classical software, to deal with disclosure risk without any geographical consideration, simply considering the mesh as a variable with many modalities. Nevertheless, a geographically intelligent management of disclosure issues will preserve underlying spatial phenomena, but no specific software has been developed yet.

In practical terms, dealing with spatial data adds a layer of complexity in the disclosure control process because it requires much computing power. On micro-data, some SDC methods involve specifying the neighbourhood structure with a weight matrix (so-called "W matrix"), whose size can easily become unmanageable for classic computers. On tabular data as well, detecting the risky observations by differencing sometimes requires to combine many dimensions (NP-hard issue).

A growing preoccupation

Last but not least, the specificity of spatial data is that they have become more and more numerous and popular, especially in the form of grid data. Increasing geolocation of data by NSIs make it possible to disseminate increasing amount of grid data (at national or global level [3]).

Grid data has many advantages. It brings a satisfactory answer to the need of having a better representation of socio-economic realities and getting rid of administrative zoning, that do not reflect socio-economic or natural realities (Clarke 1995, Deichmann et al. 2001). It gives a better description of sparse areas, like in Finland and Sweden (Tammilehto-Luode 2011). Since the

3. In the 1990s, the project "*Gridded Population of the Word*" began to apply these principles to global geography. It has been followed by a continuous improvement of the resolution (Deichmann et al. 2001). In the beginning of 2010, the Geostat project was launched in cooperation between Eurostat and the European Forum for Geography and Statistics (EFGS). The first part of the Geostat project dealt more specifically with grid data (Backer et al. 2011), and the second part aimed at fostering a better integration of statistics and geospatial information in order for the statistical community to provide more qualified descriptions and analyses of society and environment (Haldorson et al. 2017).

squares have always has the same size, grid data ensures comparability over territories and time. If needed, squares can be reassembled to form customisable study areas. Grid data also constitute a good source for auxiliary data or for local sample. Finally, it is easy to integrate data of different nature to grid data, with possible use cases in many disciplines: meteorology, environment, health, telecommunications, marketing, etc.

The next section presents how, in this context of profusion, geographical concerns can be introduced into the choice of SDC methods to keep maximum data utility for geographical analysis.

14.1.3 Recommendations to measure disclosure risk of spatial data

Quantitatively evaluating disclosure risk is a crucial step for SDC experts. With non spatial data, disclosure risk metrics have been developed and discussed (Willenborg et al. 2012, Duncan et al. 2001, Doyle et al. 2001). They are often based on a decision-theoretic characterization of the intruder (Lambert 1993, Duncan et al. 2001). To describe the final micro dataset, *k*-anonymity and *l*-diversity are common concepts. A dataset satisfies *k*-anonymity if for each combination of values of quasi-identifiers there are at least *k* observations, whereas the dataset satisfies *l*-diversity when for each combination of quasi-identifiers there are at least *l* "well represented" values for sensitive attributes. The *l*-diversity model extends the concept of *k*-anonymity with intra-group diversity for sensitive attributes in order to prevent group disclosure through homogeneity.

With spatial data, individual measures of risk can be calculated to take into account the fact that a record is risky conditionally to a geographical level or to the personal knowledge of the intruder. But the task is not easy because there is no consensual binary measure of risk.

Whether the data is spatial or non-spatial, one approach is to build the tabular data just as if it were disseminated without any constraint, and to flag risky cells as cells that do not satisfy the constraints (minimum cell sizes, dominance rule (also called (n,k) rule), p% rule [4]). Then risky records are all the records inside risky cells. For grid data or small mesh data, risky areas can be flagged with these same rules, considering the mesh or the square like a dimension like another of the tabular data.

Another approach, appropriate for pre-tabular methods, is to work directly on the micro-data. Each observation is associated to a probability of being reindentified by an intruder. The underlying idea is that an observation is risky if it is not surrounded by similar observations. Conditionally to a list of quasi-identifiers, a score evaluates, for each record, the likelihood to find someone else sharing the same characteristics in the neighbourhood. An individual alone in an empty area will always be considered as risky, but an elderly man located in an area with mainly young people will be risky as well.

Ideally, such a score requires choosing a definition of distance or neighbourhood between two records (euclidean distance, number of households in a circle, queen or rook neighbourhood [5]), and to build a *n*n* matrix on exhaustive data [6]. For populous areas, this computation quickly encounters computing power issues. To solve this, an alternative is to base the risk measure on:

— frequency counts of sensitive variables (see also the "special uniques" algorithm developed in Elliot et al. 2005);

— a simpler definition of the neighbourhood: belonging to a same area at a superior hierarchical level. That supposes to have a nested system of geographical levels [7]. In this case, spatial location of the records is not directly used.

4. All these rules are well-known in disclosure control literature and will not be developed here.
5. See Chapter 2.
6. Where *n* is the number of records in the micro-data
7. This hierarchical system can be the final releasing support, or can be drawn specifically.

Two examples to target the SDC method to the riskiest records are presented below.

Box 14.1.1 In Shlomo et al. 2010, a score is calculated for each record as follows. M key variables (quasi-identifiers, all categorical) are selected, each having k_m categories ($m = 1, ..., M$). We are in a hierarchical system of geographical levels (for example nested NUTS partition, or grid of squares of different sizes). For each geographic level l with G modalities ($g = 1, ..., G$, for example G squares), the univariate frequency count is denoted $N_k^{g,m}$ ($k = 1, ..., k_m$). The table of $N_k^{g,m}$ below has $G * \sum_{m=1}^{M} k_m$ cells.

g	Mod. A1	Mod. A2	Mod. A3	Mod. B1	Mod. B2
1	5	4	1	7	3
2	4	3	3	9	1
...					
G	5	0	5	6	4

For each level of geography l (for example for the square level), Shlomo et al. 2010 calculate for every record i (having modalities $(k_1^i, ..., k_M^i$ and belonging to the mesh g^i) a score equal to the average of the reciprocal counts:

$$R_i^l = \frac{\sum_{m=1}^{M} 1/N_{k_m^i}^{g_i,m}}{M} \tag{14.1}$$

In the example above, an individual i in region g=1 taking modalities ($A1$, $B1$) will have a risk equal to $(1/5 + 1/7)/2 \simeq 0.17$. Then thresholds T^l are set for each level of geography l, and scores above thresholds flag risky records. Thresholds generally correspond to quantiles; they are set by the expert who decides which proportion of the population must be considered as risky, with the problem of being data-specific [a].

Hungary Census grid-based Statistics use the same approach to target risky individuals but introduce multivariate distributions: in Nagy 2015, flag values are calculated for every possible combination of 3 chosen attributes (including the grid square), and the n riskiest records will be the n first, after sorting the micro-data by decreasing sum of flag values. The number of cells in the frequency counts table is then $G * \prod_{m=1}^{M} k_m$ cells (sparse table). If M is high (or if most of k_m are high) then computation power issues can be encountered. A solution can be to create *ad hoc* quasi-identifiers crossing relevant variables, or to add *a posteriori* to the at-risk sample records with very rare combination of modalities (like widows under 20 years old).

a. A threshold can be relevant for some size of mesh but irrelevant for another. For example, 10% of risky squares does not mean the same thing if the mesh is 10 meters or 10 kilometres.

In all cases, high risk households are generally defined as any household having at least one high risk record.

14.2 How to deal with disclosure risk?

Now risky records have been identified, perturbation must be added to the data, in order to make the global level of risk acceptable. Section 14.2.1 gives generalities about disclosure control techniques, and Section 14.2.2 gives an overview of SDC methods taking specificities of spatial data into account. To finish, Section 14.2.3 suggests tools and metrics to assess the effectiveness of the chosen method.

14.2.1 Pre-tabular or post-tabular SDC methods?

Traditionally, in disclosure control literature, a distinction is made between post-tabular methods, applied on tables (hypercubes) and pre-tabular methods applied on micro-data. Concerning census data, in practical terms, most countries adopt post-tabular methods, for example aggregating cells until sufficient thresholds are reached. These methods have to be applied several times, and this becomes very cumbersome when different geographies are used or when consistency is required between different linked tables. Moreover, post-tabular methods can distort relationships between variables (Kamlet et al. 1985) and spatial correlations.

Pre-tabular methods appear to be a very attractive solution [8]. Firstly, they only have to be applied once, because if micro-data are safe, so all possible aggregations from them will be safe too, and consistency is preserved. Secondly, they are more customisable [9] and they allow a great flexibility of statistical products, both with grid data or hypercubes (they also permit tailored data for users). Another advantage is that some pre-tabular methods (like record swapping) can be unbiased, whereas most post-tabular methods involve suppressing cells and then introducing bias in the estimation of parameters, or turning some parameters not estimable. Nevertheless, in practice, a single table from which every table could be safely extracted, is not realistic, because for a given level of risk, the SDC expert will have to alter too many records (Young et al. 2009), which is not reasonable for an NSI. Moreover, pre-tabular methods can let the users believe that nothing is being done to ensure confidentiality (Longhurst et al. 2007, Shlomo 2007), because applied alone, they can lead to releasing small cells for sensitive variables.

A classic trade-off is to implement basic protection in the micro-data file, and then to add protection to tables when needed (Massell et al. 2006, Hettiarachchi 2013). Post-tabular methods are indeed applied under some conditions for output products (thresholds, (n,k) rule, $p\%$ rule, etc.). For example, after perturbation on micro-data, cells that are still unique regarding a given variable will be suppressed.

To take spatial features into account, pre-tabular methods seem to be more appropriate because it is possible to use the geographical information directly to target the riskiest records for the perturbation.

14.2.2 Overview of SDC methods taking geography into account

Traditional SDC methods are already the purpose of a dedicated Eurostat handbook (Hundepool et al. 2010, Hundepool et al. 2012) and are therefore not detailed in this chapter. However, we here describe and give references of methods with more explicit consideration of geographic information.

Local imputation

Markkula 1999 is one of the first articles that takes the fact of having geographical data in the choice of the SDC method into account. His method, local restricted imputation (LRI), was co-developped by Statistics Finland and the University of Jyvaskyla, and has been tested on census data by Statistics Finland. The method includes three phases:

1. definition of the setting: minimum size of an area, and spatial configuration (3 nested levels called microdata level, macrodata level and stencil level);
2. identification of risky areas, with the number of individuals under a safety threshold;
3. imputation of the new values for the risky areas. Two techniques are considered: *(i)* imputation by the mean of all risky areas belonging to the higher hierarchical level and *(ii)*

8. The purpose of pre-tabular methods is not to release micro-data themselves, but to provide a common base to build tabular data or grid data.

9. In general, NSIs do not reveal to the users the setting of parameters of the SDC method (rate of swapped observations, PRAM matrices, parameter of a distribution law, etc.), in order to minimise the risk of retro-engineering by the intruder (Shlomo et al. 2010, Zimmerman et al. 2008).

imputation by random permutation: value in the risky area is replaced by a value from another risky area drawn randomly in the surrounding area.

The LRI method mainly aims at preserving spatial correlations, while restoring as much information as possible about the data. Then it can be appropriate for grid-based data (Tammilehto-Luode 2011). The advantage is that it is simple to understand and offers consistency of results (totals in the higher hierarchical level are preserved), but the method lacks documentation to be precisely reproduced.

Geographical aggregation

Most of the time, the data protection rule takes the shape of a threshold below which data cannot be disseminated. In the context of grid data, a strategy consists in aggregating contiguous grid cells into bigger polygons (*e.g.* rectangles or bigger squares), so that each polygon respects the threshold. These methods consist in finding the optimal grid where information is released at the most detailed level it can be released. It stands somewhere between SDC methods and data visualisation, because it consists in creating maps of varying resolutions, according to some statistic criteria. New polygons can be obtained by bottom-up aggregation (grouping cells until the threshold is reached), or by disaggregation (starting with a large group of grid cells and cutting it until no sub-cutting is possible any longer).

These methods have interesting properties: for additive data, additivity is preserved, and for average values, the method is perfectly equivalent to using imputation by the mean of the polygon for all squares within a same polygon. This method does not introduce "false zeros" and respects, by construction, the threshold rule.

On the other hand, geographical aggregation does not solve other issues like easy re-identification of extreme values or rare combinations. It also sometimes leads to polygons that do not correspond to any geographical reality: for example, an island can be grouped with the closest mainland. And finally, it also leads to differencing issues with other levels of releasing (co-existence of geometric and administrative boundaries).

Two variations of this principle of geographical aggregation are presented below. The former was developed by INSEE (French National Institute of Statistics) for fiscal data releasing in 2013, and the latter for building stock visualisation in Germany (Behnisch et al. 2013).

Box 14.2.1 — Exemple 1: a grid of rectangles. In 2013, to release fiscal data at the level of 200-meter squares - grid-based level in France, INSEE had to respect a regulatory threshold: no fiscal statistic can be released if it does not refer to a minimum of 11 fiscal households. To do this, INSEE used a disaggregation algorithm. The metropolitan territory is first divided into 36 large squares of similar sizes. Each large square is cut horizontally or vertically to form two smaller rectangles. The resulting rectangles are then cut horizontally or vertically, and so on. Horizontal and vertical cuttings always pass through the centre of gravity weighted by the population.

The choice at each stage between horizontal cutting, vertical cutting, or absence of cutting is arbitrated as follows (see Figure 14.1):

— if the horizontal and vertical splits each produce at least one rectangle of less than 11 households then the division is not carried out, in order to respect the constraint of statistical confidentiality

— if only one of the two splits produces a division into two rectangles of 11 or more households each, this split is carried out

— if the two splits each produce two rectangles of more than 11 households, the chosen split is the one that produces two rectangles within which the inhabited squares are least dispersed. The dispersion of a rectangle is measured by the sum of the squared distances

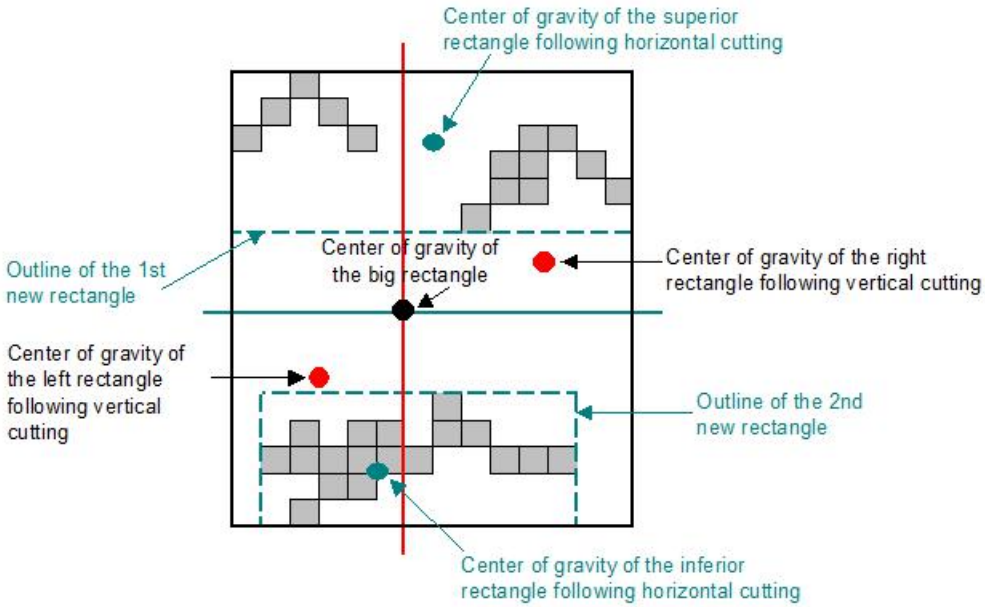

Figure 14.1 – Example of trade-off between horizontal and vertical cuttings

between its centre of gravity and its population-weighted population squares, and the chosen chopping minimizes the sum of the dispersions of the two resulting rectangles.

The grid of rectangles reflects well the spatial irregularity of data. However, it is designed conditionally to a version of data set. The grid is not stable for different sources or for different versions of the same source.

Box 14.2.2 — Exemple 2: Quadtree method. The quadtree method is another geographical aggregation algorithm that allows multiple resolutions of data in one visualization. It was conducted by the Leibniz Institute of Ecological Urban and Regional Development (Behnisch et al. 2013) for building stock visualization in Germany. The algorithm begins with the highest resolution grid (*e.g.* 250m * 250m) and if a grid cell contains a number of units under the threshold, it is aggregated with its neighbours to form a cell of bigger level (500m * 500m). The algorithm stops when all cells are above the threshold (see Figure 14.2).

The quadtree approach offers a consistent grid for different sources, but also for different versions of the same source over the time. It means that it is possible to find a level where different sources can be combined for analysis purposes. The drawback is that it masks some cells above the threshold (in bold on Figure 14.2), and it does not totally solve the MAUP (Modifiable Areal Unit Problem), since data are aggregated into a grid defined in a deterministic way.

Targeted record swapping

Swapping in general (sometimes considered as a particular case of Post Randomisation Method (PRAM, Gouweleeuw et al. 1998)) consists in exchanging attributes of two observations. Targeted swapping (as opposed to random swapping) targets the riskiest records of the data for exchanging of attributes. This pre-tabular method is often shown as a good compromise between protection and utility. Swapping offers consistency since one record takes the place of another, so that whatever the variables considered, univariate distributions are preserved, and number of records by cell is not

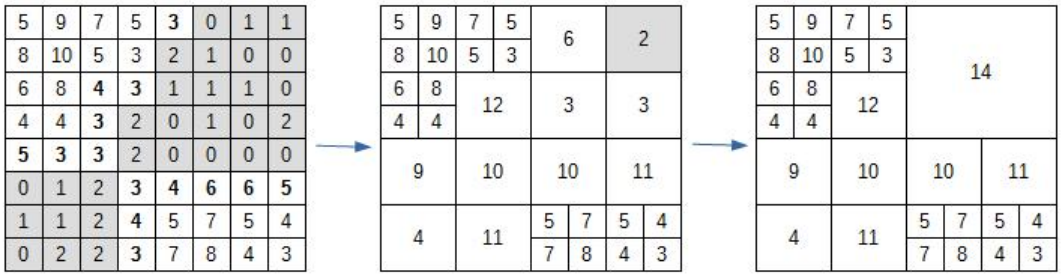

Figure 14.2 – Example of quadtree approach (bottom-up principle) applied to grid data, with a threshold of 3

modified (particularly, swapping does not introduce "wrong zeros" for the global count of records). Inconsistencies only appear when combining variables.

The Office for National Statistics (ONS), the British NSI, had several initiatives about geographically intelligent extensions of swapping for census data releasing (Brown 2003, Shlomo 2007, Young et al. 2009, Shlomo et al. 2010). Targeted record swapping (TRS) is a pre-tabular method that has been tested for some synthetic data from census data in Great Britain. In Japan, Ito et al. 2014 has also tested targeted swapping for the 2005 Census micro-data release. The main addition of TRS is to concentrate the swapping on the observations with the greatest risk of reidentification, defined at the level of a given geography, and to coerce swapped records not to be too distant geographically.

First versions of TRS were developed for hierarchical geographies (Brown 2003, Shlomo 2005), or for grid-based data (Nagy 2015). In these initiatives, two individuals cannot be swapped if they do not belong to the same area at a superior hierarchical level. For a given risky observation, the eligible observations are typically those belonging to the same surrounding area, and among them the match is made on key variables, giving priority to other risky records and eliminating records which have already been swapped.

The local density swapping (LDS) method, described by Young et al. 2009, goes one step further and uses directly the spatial coordinates in the distance function. In LDS, for a given risky observation, the eligible observations are those having the same match variables, and among them a distance function is minimised in order to choose the record to be swapped with. The main idea of LDS is to replace the Euclidean distance by the number of households between the two households to be swapped (*i.e.* located in the circle centred on the original household and with the matching household on the circumference), so as to take population density into account. Priority is given to records which have not already been swapped.

LDS allows a lot of flexibility since it is largely customizable (size of sample, choice of distance, set of matching variables). It also appears to be particularly appropriate to the context of grid data since it takes the geography into account more precisely than the other scenarios. However the drawback of LDS is it is, like every pre-tabular method, not self-sufficient. Moreover the method can leave the impression that nothing has been put in place for disclosure control.

Extensions

Trajectories

Trajectory data can be considered as a specific kind of spatial data. Bi-localised data can be collected by a lot of technologies, but they are highly sensitive, because they say a lot about individual habits (places usually visited, etc.). This is why their de-identification is more delicate.

A trajectory is often associated with a temporal aspect, which it is relevant to take into account in the protection method. Both temporal and spatial aspects of the trajectories can be considered when measuring the distance between two trajectories.

Domingo-Ferrer et al. 2011 present two methods to anonymise trajectory data, named *SwapLocations* and *ReachLocations*. The first one preserves trajectory k-anonymity whereas the second one guarantees location l-diversity.

Shlomo et al. 2013 suggest a protocol to detect and correct trajectory outliers, taking geographical information into account. The authors consider commuting to work statistics: each trajectory is characterized by two geographical positions and one travel time (in minutes). Outliers are defined according to a given mode of transport. In a first step, outliers are detected among the trajectories, using the Mahalanobis distance (based on a multivariate normal distribution). In a second step, outliers are modified. The place of residence is altered, but the workplace is unchanged so as not to introduce inconsistency (perturbation would be easy to detect if a factory is set where there is none in reality). To do that, the authors define a coherence function at individual level, in order to evaluate the plausibility of the trajectory with respect to the multivariate dataset of non-outliers.

Several algorithms have been tested in this article:

— *record swapping*: iterative algorithm where for each subgroup mode of $transport \times sex \times age$, they swap the places of residence while keeping other variables unchanged. At each iteration, consistency is calculated for all possible pairs, and the match is made for records that optimize the consistency. Iterations stop when the general gain of coherence (decreasing at each step) becomes less than a predefined threshold.

— *hot deck*: instead of being exchanged, places of residence for outliers are erased and replaced by imputation from the value taken by a donor having the same characteristics. Selecting the donor can be made by maximising the coherence among all potential donors in a neighbourhood, or minimising the difference in terms of travel time.

Finally, hot deck corrects more outliers than swapping, but swapping minimises the loss of information. In both cases, it is possible that non-outliers become outliers (but fewer cases for swapping).

Geomasking

The term *geomasking* was introduced by Armstrong et al. 1999. It brings together all the methods aiming at altering the geographical position of a point, in order to guarantee more confidentiality for spatial point patterns. One of the most popular geomasking techniques is the donut method, in which every geocoded address is relocated in a random direction, with a distance superior to a minimum and inferior to a maximum.

Geomasking has not been documented much in economics but it is widely used in epidemics or for crime data. In these fields, point patterns have to be released and studied, whereas we assume here that the goal is to release mesh data (on a regular grid or on an administrative zoning), and that micro-data is not the final product but some input we can alter to reach the goal. In the context of census data, moving households are generally excluded, because it could result in obvious inconsistencies (for example it could lead to set a household in the middle of a lake). It also creates "wrong zeros" and does not preserve "true zeros".

14.2.3 How to assess effectiveness of disclosure control?

Spatial measures of utility

Applying an SDC method consists in deteriorating data quality in exchange of more protection and results in a loss of information for users. To arbitrate between different SDC scenarios, measuring utility actually means measuring a disutility or a distortion. According to Willenborg et al. 2012 about impact of SDC techniques on micro-data, there are two kinds of losses of information: an increase of the variance in the estimation of a parameter, or the introduction of a bias (which is

obviously the case for example with the suppression of extreme values).

Different metrics can be used to measure the loss of information (Domingo-Ferrer et al. 2001). In all cases, the sharing of perturbed records has to be checked. In addition:

— for continuous variables, mean square error, average absolute error, average rank change, or comparison of Pearson's coefficient between two variables known as correlated. If the dissemination mainly aims at producing a specific indicator like the unemployment rate, it is also relevant to check if bias is not introduced between the original file and the anonymised file. Other model-based metrics can also be used by computing the confidence interval overlap measure for a given logistic regression (De Wolf 2015);

— for categorical variables, it can be direct comparing frequency counts, entropy-based measures like Hellinger distance (Torra et al. 2013), or comparing contingency tables between two variables known as correlated.

For spatial data, we can add to this list the proportion of impacted geographical units, the absolute average deviation (AAD) of countings of a given attribute before and after SDC, calculated at the level of meshes (or squares), or the Moran or LISA indicators [10], if we know an attribute exhibiting spatial dependency.

R-U Confidentiality Maps

In order to compare different SDC strategies between them or to choose the most appropriate parameters, "risk-utility maps" can be drawn for different risk levels.

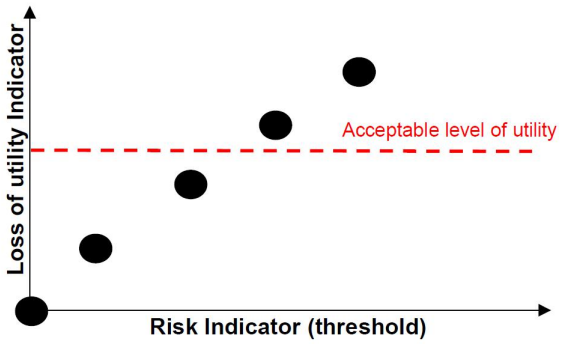

Figure 14.3 – Principle of Risk-Utility Confidentiality Maps

R-U confidentiality maps (Figure 14.3) were first formalised by Duncan et al. 2001 (with an example about additive noise method), but then were used in many papers (Young et al. 2009, Clifton et al. 2012, Gomatam et al. 2005). They constitute a workable tool to frame decision making and to give a synthetic representation of the trade-off between reducing disclosure risk R, expected low, and preserving data utility U, expected high. An R-U confidentiality map is a chart that plots the impact in R and U of changes in the parameters of a disclosure limitation procedure.

14.3 Application for a grid of 1 km² squares

In 2017, the Eurostat grant "*Harmonized Protection of Census Data in the ESS*" aimed at harmonising disclosure control techniques concerning Census in European countries, for hypercubes on the one hand and for grid data on the other hand. For this grant, two complementary methods have been chosen, because they seemed to offer a good compromise between confidentiality and utility loss. Targeted record swapping was selected to alter the micro-data in a first step. Then, grid

10. **See Chapter 3.**

data and hypercubes are built on altered micro-data, and noise is added on cells of hypercubes. This second step is called "Cell-key method" and is inspired from Australian Bureau of Statistics (Fraser et al. 2005).

In this section, we try to assess how targeted record swapping alters spatial correlations, using fiscal data of a small French region. We present the main steps of the method and the results through a risk-utility analysis.

14.3.1 Targeting Record Swapping: details of the method

Implementation choices are taken from an ONS program [11], and adapted in order to stick to French data. The original algorithm is adapted to hierarchical data, structured in 3 different nested levels ($level1 \subseteq level2 \subseteq level3$). The method is designed in four steps detailed below.

Step 1: Targeting the risky records

The first step is to identify the records that need the most to be swapped. An individual can be considered risky or not for a given set of characteristics. Being risky means that there are very few similar records in the same area: a rarity score is computed for each individual as suggested above (average of the reciprocal counts), and individuals with scores above a threshold (a quantile) are flagged as risky. Then, high risk households are defined when there is at least one high-risk individual in the household.

We also associate a geographical level of risk to each individual: if the modality is very rare even at a bigger level (fewer than X individuals sharing the same modality in the area), then he is "unique" for this geographical level. The geographical level of risk of the household is defined as the highest geographical risk among all persons in the household. A risk 2 household may be matched with a more distant household than a risk 1 household.

Step 2: Selection of the sample to swap

The principle of this step is to constitute a sample of risky households, with a size twice smaller than the whole risky population. Then, each household in this sample is associated with another household (preferably risky), so that almost all of the population at risk will be perturbated. We draw a sample stratified by the smallest geographical level, with probability proportional to the arithmetic mean of two indicators (predictors of disclosure):

— a first one increasing with the proportion of high risk households in the geographical unit (this proportion is known throughout the population by construction);
— a second decreasing with the number of households in the geographical unit.

All households have a non-zero chance of being selected for the sample, but high risk households have a much higher probability of being selected. In addition, the sample always contains at least one household per geographical unit. The algorithm also allows to limit the proportion of sampled households in the geographical unit, but results presented below do not use this possibility.

Step 3: Matching

The principle of matching is to find, for each household of the sample, a match outside the sample but with close geographical and / or demographical characteristics, with preference for other risky households. The matching process takes place in different stages and sub-stages. Firstly the focus is made for the records risky for the superior level ($level3$), and finally to the inferior level. For each of these 3 stages, constraints of similarity are less and less strict with the sub-stages.

More precisely, if the current stage is dealing with the hierarchical level l, the principle of each sub-stage is as follows. First, we select a part of the sample. Then, for each household of this sub-sample, we search a "twin" from a "reserve". The match is randomly searched outside the

11. We are grateful to Keith Spicer and Peter Youens for their valuable advice and clarification on the algorithm.

sample. The matching household must have the same profile, be part of a different geographical area (level l), but within the same geographical area at the superior hierarchical level (level $l+1$) [12]. For example, for a household flagged risky in step 1, with a geographical level of risk $level1$, another household will be searched outside the same $level1$ but inside the same $level2$.

There is a preference for households also identified as high-risk. At the end of each stage, the "reserve" is reduced by matched households, so that a household cannot be swapped several times. As long as the sub-stages go on, the constraint on the profile is released, so that at the end all the households in the sample have been matched to another household. This whole method ensures that almost all identified high-risk households are swapped.

Step 4: Swapping

Finally, geographical information is swapped. The method does not introduce "false zeros" in counts of people, but it can for counts of specific variables.

14.3.2 Choice of data and parameters

Data

For this chapter, we chose to apply TRS on exhaustive fiscal data [13]. Unfortunately, tests have only been made on the region Corsica (the smallest French NUTS 2), in order to test many parameters in a reasonable computation time. These tests have been carried out for experimental purposes and should be extended on more populated areas to generalise the conclusions.

For this TRS algorithm, each unit of the 3 needed geographical levels must contain a sufficient number of records. We chose to build *ad hoc* geographical units with INSEE's geographical aggregation algorithm described in section 14.2.2. Squares of 1 km^2 are grouped to constitute rectangles. At the end, we have the following hierarchical structure:

— *level 3*: NUTS 3 (French "départements");
— *level 2*: rectangles containing at least 5000 individuals, intersected with level 3;
— *level 1*: rectangles containing at least 100 individuals, nested in level 2 rectangles (Figure 14.5), and intersected with level 3.

Each level is obtained by disaggregating the previous level, and the most detailed mesh is the 1 km^2 square. If a 1 km^2 square contains at least 100 individuals, then it is not aggregated with neighbour squares to constitute a level 1 unit.

Corsica is made of 2976 1 km^2 squares but only 756 small rectangles (containing at least 100 inhabitants) and 39 big rectangles (containing at least 5,000 inhabitants, see Figure 14.5). Even if rectangles are built to reach a threshold of records (5,000 or 100), all the units do not have the same number of households because some 1 km^2 squares are made of more than 5,000 households in the big cities (Ajaccio or Bastia, Figure 14.4).

In these tests, the geographical level of releasing is not squares but groups of squares ($level1$). If we want then to release counts for quasi-identifiers at a finer level (1 km^2 square), then distribution keys will have to be set, for example randomly in non-empty squares of the $level1$ unit, or proportionally to the number of inhabitants of the square if this quantity is known (not sensitive).

Parameters

Firstly, we chose 4 categorical variables to define disclosure risk - gender, 5 year age-range, place of birth (12 modalities) and place of residence the previous year (7 modalities).

12. More precisely again, the algorithm proceeds in a set of iterations. At each iteration, each household of the sub-sample is randomly assigned another potential household, and it is decided to match it if all the conditions are met. If this is the case, both households come out of the reserve. Otherwise, the first household remains in the sub-sample and the second one remains in the reserve for the next iteration. A sufficiently large number of iterations is selected so that at the end, no more possible matching can be found.

13. The French Census is specific because it is a survey with associated weights, and this chapter does not aim to discuss the weighting issues.

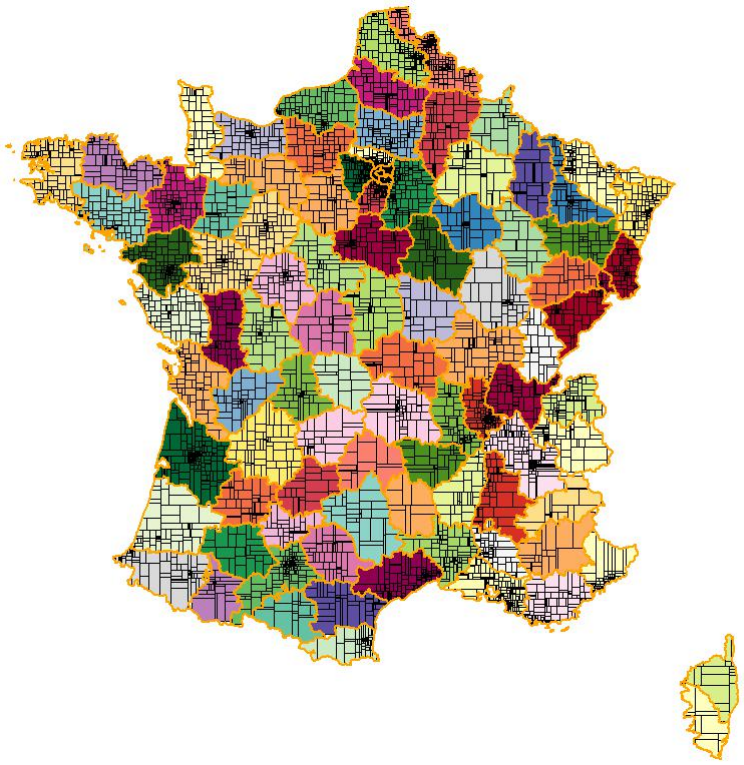

Figure 14.4 – France split into 5,000 individuals rectangles (built at NUTS 3 level)

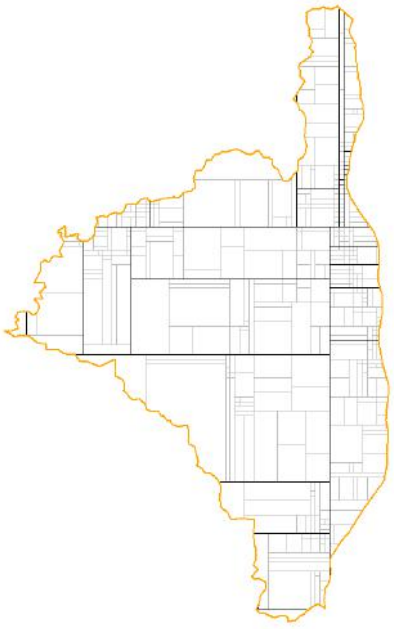

Figure 14.5 – Department 2B (Haute-Corse) split into level 1 and level 2 rectangles

Then, the main parameter is the threshold below which a record will be considered as risky. The sample size and therefore the share of swapped households are derived from it, even though there is no direct formula between the two. By construction, the proportion of swapped households in the population will be slightly higher than this parameter, but of the same order of magnitude. Different parameters (from 1 to 10th percentiles) have been tested, leading to proportions of swapped individuals from 2% to 16% [14].

Finally, 3 profiles are defined, from the least detailed to the most precise. Two households will not be swapped if they don't share the same profile. For the following simulations, we chose:

— profile A (most detailed): similar number of persons in each of the 7 gender*age categories [15];
— profile B (intermediate): similar number of persons in each of 5 gender*age categories;
— profile C (less detailed): similar number of persons in the household.

14.3.3 Results

The output of the algorithm is an altered data set containing, for each record, the original area before swapping, and the area after swapping. Counts can then be made with this output.

Results are shown through a risk-utility analysis (see Section 14.2.3). The risk measure is the threshold set as a parameter of targeted record swapping (from 1 to 10%). A high threshold means that a low level of risk [16] is accepted.

To measure utility loss, the following metrics are used [17]:

— share of level 1 units (small rectangles) impacted by swapping (counts are not the same), for two variables - number of males (taken into account in the matching step) and number of people born in France (not directly taken into account);
— absolute average deviation of countings for level 1 units (small rectangles), for the same two variables;
— Moran's I, calculated at the level of the small rectangles, for 4 sensitive variables with different intensities of spatial autocorrelation: number of people born in France, number of children under 5 years old, income, and number of people belonging to a deprived neighbourhood (QPV [18]).

Results of the tests are shown in Table 14.1 and Figure 14.6 (RU-Maps slightly different from suggested previously). Distortion is measured for variables directly taken into account in the method through the matching profile ($V1$, number of males), indirectly taken into account in the method through the rarity score ($V2$, number of people born in France or $V3$, number of children under 5 years old), or not taken into account at all in the method ($V4$, income, or $V5$, number of people in QPV).

We easily see that the higher the acceptable level of risk is (*i.e.* the lower the share of population considered as risky), the lower the distortion in the share of impacted geographical areas and average absolute deviation are.

Even for low values of parameters, the majority of the rectangles are impacted by TRS (for the 1% parameter, the highest level of risk tested, 70% of the counts per small rectangle are changed

14. For another region with more inhabitants, the share of swapped records would be closer to the initial parameter. The reason is that in the case of Corsica, the constraint is more difficult to satisfy and the match is more often found outside the risky population.

15. For some age categories both males and females are grouped.

16. We also considered another risk measure with the 90th percentile of the rarity score defined as above (average of the reciprocal counts of the level 1 unit), but this does not vary enough to make relevant graphs.

17. We do not consider the share of swapped individuals for the risk-utility analysis because with this method, it is by construction highly linked to the threshold parameter.

18. In French, "Quartier Politique de la Ville".

Risk measure (%)									
Threshold parameter	0	1	2	3	4	5	7	8	10
Utility measures (%)									
Share of swapped individuals	0	2	4	5	7	8	11	13	16
Share of impacted level 1 units - V1	0	38	52	56	63	69	73	75	78
Share of impacted level 1 units - V2	0	71	82	85	85	88	90	92	94
AAD (level1) - V1	0.0	0.5	0.8	0.9	1.1	1.2	1.5	1.6	1.7
AAD (level1) - V2	0.0	0.8	1.1	1.3	1.6	1.6	2.1	2.2	2.5
Moran (level 1): V2	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5
Moran (level 1): V3	6.5	6.4	6.5	6.5	6.6	6.7	6.7	6.8	6.4
Moran (level 1): V4	5.5	5.2	5.1	4.6	5.2	6.8	8.2	5.7	6.4
Moran (level 1): V5	7.7	7.7	7.8	7.7	7.8	7.7	7.3	7.2	7.3

V1: number of males
V2: number of people born in France
V3: number of children under 5 years old
V4: average income
V5: number of people in QPV

Table 14.1 – Results of the tests led on Corsica for several parameters

for the variable "number of people born in France"). But the change is reasonable: for the highest level of risk, 0.4% and 0.7% (respectively for V1 and V2) of the absolute changes are under 5% of the count and the AAD are also under 1%. For the lowest level of risk tested (10% parameter), the AAD is 2.5% for the number of people born in France and 1.7% for the number of males.

Concerning spatial correlations, we now focus on the distortion of Moran's indicator (calculated for $level1$ units, before and after TRS). We see that the distortion can be very important (up to 50% of variation of the indicator), that it does not always go in the same direction (the spatial correlation can be increased or decreased with the method), and it is not a monotonic function with level of risk.

Finally, we also see that the utility loss, as regards all the indicators, varies with the variable. If the variable has been taken into account directly in the method (in the definition of the profile: number of males in the tests), then the variable is less distorted than if it has been taken indirectly taken into account (in the identification of high risk people: number of people born in France or number of children under 5 years old in the tests), and *a fortiori* if it has not been taken into account at all (income or belonging to a deprived neighbourhood in the tests).

More specifically about the distortion of spatial correlations: Moran's indicator is unchanged for the variable defining the matching profile (V1), or slightly changed for variables indirectly taken into account (V2 and V3). It is also slightly changed for variables strongly correlated with the matching profile (V5, Figure 14.2). On the opposite, spatial correlations can be very distorted for variables that are not correlated with the matching profile (V4).

The distortion of Moran's I does not particularly increase with the level of risk, but erratic behaviours can appear, due to the randomness of the algorithm during the matching step. Since the method does not consider the income as a variable to preserve, and since this variable is not correlated with another variable that must be preserved, then households with similar incomes can be brought closer or more distant, randomly, from one execution of the method to another.

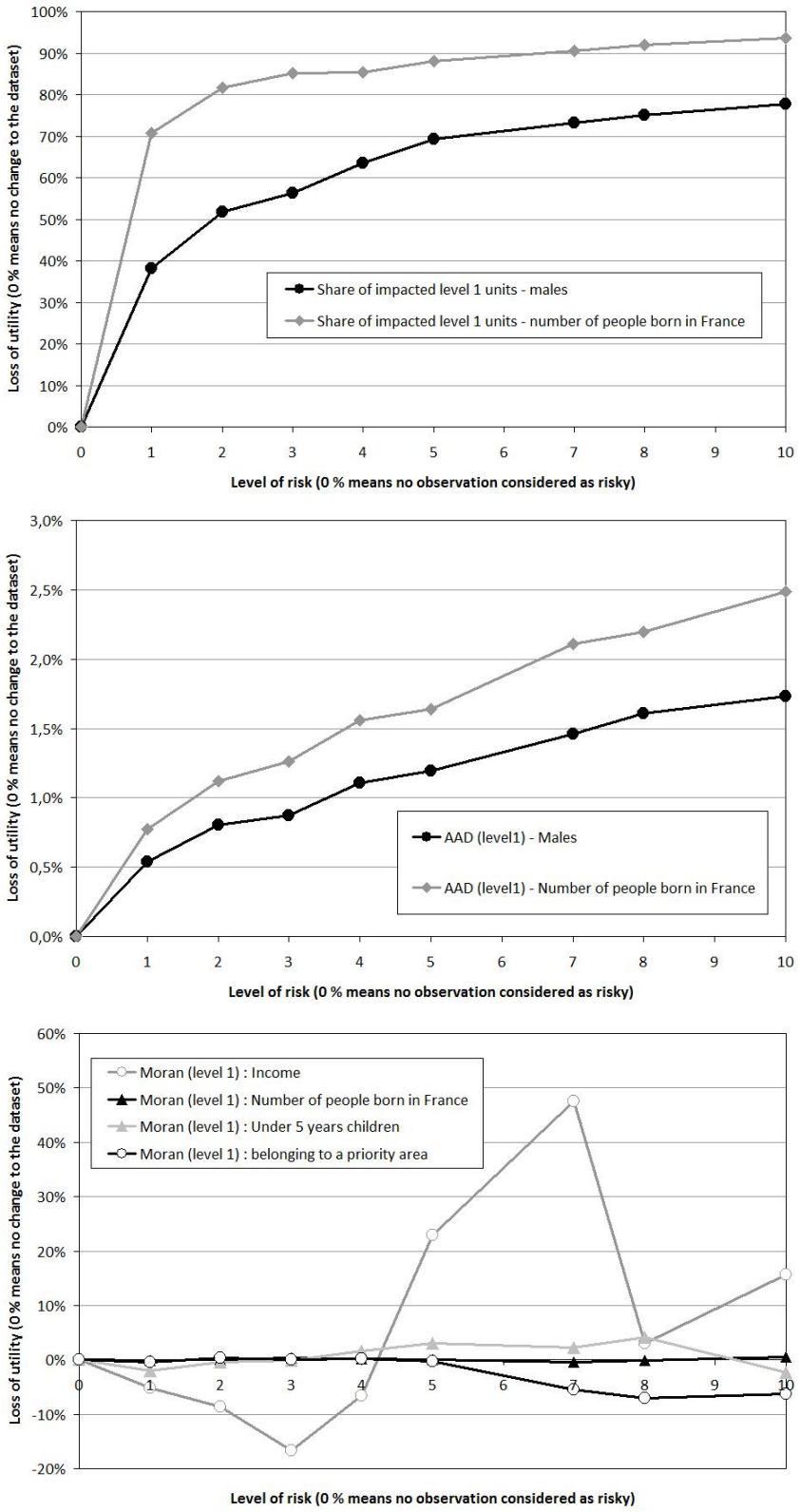

Figure 14.6 – Utility loss as a function of risk level, for 3 utility loss indicators

Pearson coefficient	V1	V2	V3	V4	V5	V6
V1 (number of males)	1	1.00	0.97	0.13	0.97	*1.00*
V2 (number of people born in France)	1.00	1	0.96	0.14	0.97	*1.00*
V3 (number of children under 5 years old)	0.97	0.96	1	0.14	0.93	*0.97*
V4 (average income)	0.13	0.14	0.14	1	0.15	0.13
V5 (number of people in QPV)	0.97	0.97	0.93	0.15	1	*0.97*
V6 (total number of people)	*1.00*	*1,00*	*0.97*	*0.13*	*0.97*	*1*

Note: V1, V2, V3 and V5 are totals and are strongly correlated to the total number of persons in the rectangle whereas V4 is an average.

Table 14.2 – Pearson coefficients between variables

14.4 Differencing issues

14.4.1 Definition

Geographic differencing occurs when an intruder can combine data released in various geographies to reconstruct data on a smaller area or deduce the location of an observation. The issue has been presented about Census releasing in many papers (Duke-Williams et al. 1998, ONU 2004), but it occurs for any source releasing.

With nested geographies (*e.g.* regions – departments – towns) the problem is quite simple to solve because the data that can be obtained with subtractions is directly linked to the hierarchy between the various geographies. Thus, once the set of small areas that need to be protected is identified which is called primary secret, SDC software like Tau-Argus can be used to choose the secondary secret. The set of areas that need to be treated so that intruders cannot reconstruct the data of the primary secret. With nested geographies described in a hierarchical tree, the problem is similar to any other variable of interest used in a tabulation (*e.g.* sections – divisions – groups – classes, used in the NACE classification of economic activities).

But the issue of geographic differencing gets more complex when the various geographies used in the release are non-nested (ABS 2015). In that case, there is no hierarchical tree to be used and specific algorithms need to be implemented to identify all the subtractions that an intruder could make between the various areas to get data on smaller areas.

This differencing issue increases when the size of the zone of releasing decreases (blatant in case of small-size grid data). It also increases with the number of various geographies, especially when they are not hierarchical. For example, if NSIs release data on *ad-hoc* zoning in specific partnerships, or if such tailored geographies are constructed by users with web services.

Another example of a differencing issue is when the same phenomenon is observed on various dates. For example, in the case of data about companies released each year, an intruder could compare the various releases to try and find some hidden values. When applying SDC techniques for the latest broadcast, one should therefore take into account what was done for the previous ones and which values were hidden.

14.4.2 Illustration

Figure 14.7 presents examples of possible cases of geographical differencing. Overlapping zones between the circles (A) and rectangles (B) are highlighted in orange. In the first case, zoning B encompasses zoning A. An intruder can reconstruct information about B-A by subtraction and

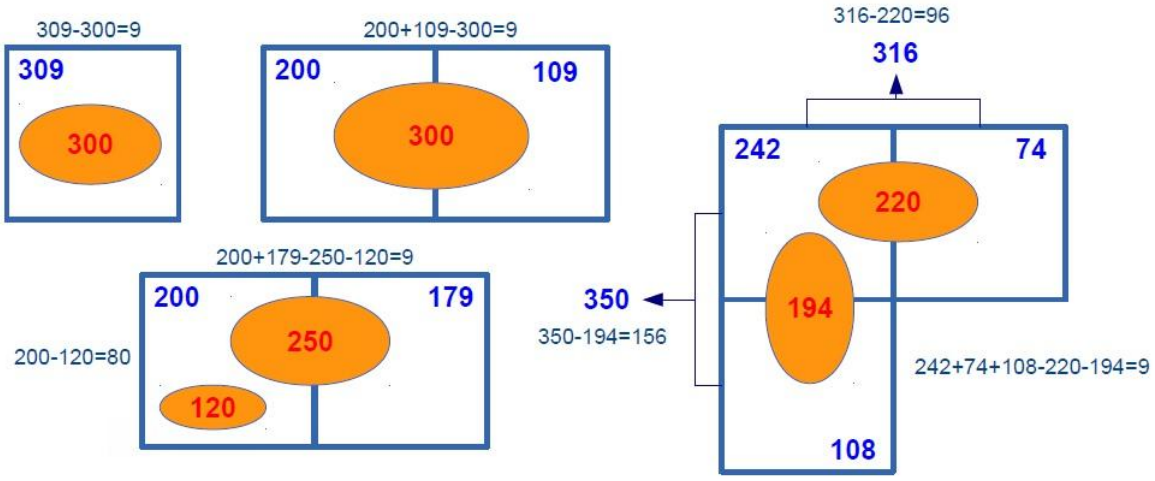

Figure 14.7 – Confidentiality breach by geographic differencing

this may lead to revealing data concerning a small number of individuals. In the second box, the intruder can combine two zones of zoning B to perform the operation (B1 + B2) - A and thus obtain information concerning a non-released zone. The last two cases show that differencing can occur with a combination of any number of the two zonings.

With the frequency counts included in these examples, if information cannot be disseminated if it concerns less than 10 individuals, then there is a breach of confidentiality by geographical differentiation in each of these four examples. In other cases of overlapping, the intruder cannot directly obtain information on a new zone, but by taking into account auxiliary information or the geography surrounding the overlapping area, problems may arise. The geography of the zone needs to be taken into account as it is sometimes impossible for certain areas to contain any observation (lake, highway, etc.). These empty areas cannot be used to protect the data and must be disseminated.

14.4.3 Identifying Risky Areas

The first step in order to deal with the differencing issue is to flag the risky areas. This can be relatively simple with the geographic information system (GIS), but it becomes complicated when the number of non-nested geographies increases because it increases the dimension of the problem to solve and can lead to NP-hard problem.

If the choice is to suppress the information for these risky areas, the method is carried in two steps: primary then secondary suppressions.

The algorithm needs to look for the possible overlaps between the non-nested geographies. As seen in Figure 14.7, problems can arise with a combination of multiple areas of each zoning. A confidentiality criteria needs to be chosen, for example at least 10 individuals in any area.

The algorithm needs to include checks with the totals if one of the non-nested geographies is hierarchical (for example when on the one hand data is released for regions-department-town but on the other hand data is released for a partner with a specific zoning).

It might be important to reduce the information loss and thus to include optimization rules to minimize the number of individuals impacted, or give priority to keep preserved areas if information is more useful there, for example when conceiving public policies for deprived neighbourhoods. This work requires a consequent disk space and a lot of computing power: the algorithm might need a long time to explore all the possible overlaps.

14.4.4 Protection methods

Different types of methods can be used to restore confidentiality when facing differencing issues due to such overlaps.

First, the zoning can be modified: the boundaries can be changed to eliminate areas of overlap, for example by nesting the various geographies and creating a clean hierarchical tree.

Secondly, if boundaries are fixed, various zones can be merged to eliminate overlaps. It reduces the levels of detail but it enables the data provider to release information on these areas.

A third method is to suppress data for specific areas where overlapping occurs. Due to the constraints in the data production system, this option is often chosen when confidentiality problems arise and it leads to a trade-off between the level of details of the release and the number of hidden areas due to differencing issues.

Instead of suppressing data when boundaries and zoning are fixed, data can be perturbed for example by adding or subtracting small numbers to the risky areas in case of frequency counts release. To do this in a consistent way for multiple tables, or for various geographies, ABS has conceived a cell key method that makes use of record keys assigned to each micro-data observation to keep the perturbation consistent between the various geographies (Fraser et al. 2005). The cell key method was adapted by ONS for the census release and the method was also tested for the Eurostat "Harmonized Protection of Census Data in the ESS" Grant.

Conclusion

Reflection about SDC methods goes hand in hand with a strategic reflection of the NSIs on what they want to release *in fine*. Aversion to the dissemination of "false" information and fear of misreading of hurried users must be discussed. Choices also have to be made whenever possible in consultation with potential future users, which is the best way to preserve statistical relationships that will be analysed at the end.

Dealing with spatial data can be seen as an opportunity to refine SDC methods, because density and dissimilarity with the neighbours are a fundamental predictor of disclosure risk. In the actual state of the art, geographical information is taken into account by perturbing the micro-data using local information of the neighbourhood (local imputation, targeted record swapping).

In the future, in conjunction with increasing computing capacities, the geographical coordinates may be used more precisely, for example by measuring local density for each record. But precision improvement must be balanced against the extra-complexity of the SDC method and the inherent additional difficulty to communicate to the users about the protection method.

Tests led on exhaustive fiscal data for one region of France show that for reasonable risk levels, targeted record swapping implies low distortion of spatial correlations, even if these tests would deserve to be continued with other SDC strategies and on bigger regions.

Nevertheless, pre-tabular SDC method is not sufficient by itself, firstly because reaching an acceptable level of global risk in the data set would require perturbing too many records, and secondly because of public perception. Post-tabular methods make more visible the existence of disclosure protection to respect the regulatory threshold. This is why concerning the census, the advice given by Eurostat to SDC experts is to combine pre-tabular methods taking geographical specificities into account, and post-tabular methods.

In any case, whatever the method used, and even if it is the most traditional, it is interesting to measure how much the SDC technique degrades the spatial relationships of some attributes. For this purpose, R-U confidentiality maps drawn for distortion of spatial correlations coefficients make an efficient operational tool.

Although precise parameters used might be kept secret to improve the protection, it is necessary that NSIs and data providers document the method and the choices that were made. Users must be

conscious that the data has been changed or might be incomplete when conducting their analysis. For example, the SDC expert can then communicate to potential users how much Moran's I or LISA are affected, in order to guard users against any misleading use of the protected data.

References - Chapter 14

ABS (2015). « SSF Guidance Material – Protecting Privacy for Geospatially Enabled Statistics: Geographic Differencing ».

Armstrong, Marc P, Gerard Rushton, Dale L Zimmerman, et al. (1999). « Geographically masking health data to preserve confidentiality ». *Statistics in medicine* 18.5, pp. 497–525.

Backer, Lars H et al. (2011). « GEOSTAT 1A: Representing Census data in a European population grid ». *Final Report*.

Behnisch, Martin et al. (2013). « Using Quadtree representations in building stock visualization and analysis ». *Erdkunde*, pp. 151–166.

Bergeat, Maxime (2016). « La gestion de la confidentialité pour les données individuelles ». *Document de travail INSEE M2016/07*.

Brown, D (2003). « Different approaches to disclosure control problems associated with geography ». *Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*.

Clarke, John (1995). « Population and the environment: complex interrelationships. »

Clifton, Kelly and Nebahat Noyan (2012). « Framework for Applying Data Masking and Geo-Perturbation Methods to Household Travel Survey Datasets ». *91st Annual Meeting of Transportation Research Board, Washington, DC*.

Curtis, Andrew J, Jacqueline W Mills, and Michael Leitner (2006). « Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina ». *International Journal of Health Geographics* 5.1, pp. 44–55.

De Wolf, PP (2015). « Public use files of eu-silc and eu-lfs data ». *Joint UNECE-Eurostat work session on statistical data confidentiality, Helsinki, Finland*.

Deichmann, Uwe, Deborah Balk, and Greg Yetman (2001). « Transforming population data for interdisciplinary usages: from census to grid ». *Washington (DC): Center for International Earth Science Information Network* 200.1.

Domingo-Ferrer, Josep, Josep M Mateo-Sanz, and Vicenç Torra (2001). « Comparing SDC methods for microdata on the basis of information loss and disclosure risk ». *Pre-proceedings of ETK-NTTS*. Vol. 2, pp. 807–826.

Domingo-Ferrer, Josep and Rolando Trujillo-Rasua (2011). « Anonymization of trajectory data ».

Doyle, Pat et al. (2001). « Confidentiality, disclosure, and data acces: theory and practical applications for statistical agencies ».

Duke-Williams, Oliver and Philip Rees (1998). « Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure ». *International Journal of Geographical Information Science* 12.6, pp. 579–605.

Duncan, George T, Sallie A Keller-McNulty, and S Lynne Stokes (2001). « Disclosure risk vs. data utility: The RU confidentiality map ». *Chance*. Citeseer.

Duncan, George T and Diane Lambert (1986). « Disclosure-limited data dissemination ». *Journal of the American statistical association* 81.393, pp. 10–18.

Elliot, Mark J et al. (2005). « SUDA: A program for detecting special uniques ». *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, pp. 353–362.

Elliot, Mark and Josep Domingo-Ferrer (2014). « EUL to OGD: A simulated attack on two social survey datasets ». *Privacy in Statistical Databases*. Ed. by Josep Domingo-Ferrer.

Fraser, Bruce and Janice Wooton (2005). « A proposed method for confidentialising tabular output to protect against differencing ». *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, pp. 299–302.

Gomatam, Shanti et al. (2005). « Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers ». *Statistical Science*, pp. 163–177.

Gouweleeuw, JM, Peter Kooiman, and PP De Wolf (1998). « Post randomisation for statistical disclosure control: Theory and implementation ». *Journal of official Statistics* 14.4, pp. 463–478.

Haldorson, Marie et al. (2017). « A Point-based Foundation for Statistics: Final report from the GEOSTAT 2 project ». *Final Report*.

Hettiarachchi, Raja (2013). « Data confidentiality, residual disclosure and risk mitigation ». Working Paper for joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

Hundepool, Anco et al. (2010). « Handbook on statistical disclosure control ». *ESSnet on Statistical Disclosure Control*.

Hundepool, Anco et al. (2012). « Statistical disclosure control ».

Insee (2010). « Guide du secret statistique ». *Documentation INSEE*.

Ito, Shinsuke and Naomi Hoshino (2014). « Data swapping as a more efficient tool to create anonymized census microdata in Japan ». *Privacy in Statistical Databases*, pp. 1–14.

Kamlet, MS, S Klepper, and RG Frank (1985). « Mixing micro and macro data: Statistical issues and implication for data collection and reporting ». *Proceedings of the 1985 Public Health Conference on Records and Statistics*.

Lambert, Diane (1993). « Measures of disclosure risk and harm ». *Journal of Official Statistics* 9.2, pp. 313–331.

Longhurst, Jane et al. (2007). « Statistical disclosure control for the 2011 UK census ». *Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester*, pp. 17–19.

Markkula, Jouni (1999). « Statistical disclosure control of small area statistics using local restricted imputation ». *Bulletin of the International Statistical Institute (52nd Session)*, pp. 267–268.

Massell, Paul, Laura Zayatz, and Jeremy Funk (2006). « Protecting the confidentiality of survey tabular data by adding noise to the underlying microdata: Application to the commodity flow survey ». *Privacy in Statistical Databases*. Springer, pp. 304–317.

Nagy, Beata (2015). « Targeted record swapping on grid-based statistics in Hungary ». *Submission for the 2015 IAOS Prize for Young Statisticians*.

ONS (2006). « Review of the Dissemination of Health Statistics: Confidentiality Guidance ». *Working Paper 5: References and other Guidance*.

ONU (2004). « Manuel des systèmes d'information géographique et de cartographie numérique ». F-79, pp. 118–119.

Shlomo, Natalie (2005). « Assessment of statistical disclosure control methods for the 2001 UK Census ». *Monographs of official statistics*, pp. 141–152.

— (2007). « Statistical disclosure control methods for census frequency tables ». *International Statistical Review* 75.2, pp. 199–217.

Shlomo, Natalie and Jordi Marés (2013). « Comparison of Perturbation Approaches for Spatial Outliers in Microdata ». *the Cathie March Centre for Census and Survey Research*.

Shlomo, Natalie, Caroline Tudor, and Paul Groom (2010). « Data Swapping for Protecting Census Tables. » *Privacy in statistical databases*. Springer, pp. 41–51.

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

Torra, Vicenc and Michael Carlson (2013). « On the Hellinger distance for measuring information loss in microdata ». *Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada, 28-30 October 2013*.

VanWey, Leah K et al. (2005). « Confidentiality and spatially explicit data: Concerns and challenges ». *Proceedings of the National Academy of Sciences* 102.43, pp. 15337–15342.

Willenborg, Leon and Ton De Waal (2012). *Elements of statistical disclosure control*. Vol. 155. Springer Science & Business Media.

Young, Caroline, David Martin, and Chris Skinner (2009). « Geographically intelligent disclosure control for flexible aggregation of census data ». *International Journal of Geographical Information Science* 23.4, pp. 457–482.

Zimmerman, Dale L and Claire Pavlik (2008). « Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data ». *Geographical Analysis* 40.1, pp. 52–76.