UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Confidentiality

26–28 September 2023, Wiesbaden

Spatial SDC experiments and evaluations with multiple countries comparison

Johannes Gussenbauer (Statistics Austria), Julien Jamme (Insee), Edwin de Jonge (CBS), Peter-Paul de Wolf (CBS), Martin Möhler (Destatis)

johannes.gussenbauer@statistik.gv.at, julien.jamme@insee.fr, e.dejonge@cbs.nl, pp.dewolf@cbs.nl, martin.moehler@destatis.de

Abstract

This study utilizes a standardized "census-like" dataset that is structured uniformly across all participating countries to assess disclosure risk based on grid data. We begin by evaluating and comparing the risk using this approach. Next, we apply spatial SDC methods from the R package `sdcSpatial`, including kernel density smoothing and quad tree aggregation. We re-evaluate the disclosure risk using these methods and analyze the resulting utility loss. Our analysis will be conducted across multiple countries, allowing for a comprehensive comparison of the utility loss between them.

1 Introduction

Population grids are by now well-established products of National Statistical Institutes. They map the distribution of population units (persons, families, households, or similar) in geographic space, using as domain a set of regular, small-scale squares, the grid cells. For our purposes it will be useful to view the gridded map as akin to a two-dimensional histogram and each cell as a bin. One advantage over irregular spatial references, like administrative areas, is then that spatial patterns can, in principle, be observed in a comparable manner over multiple countries.
A disadvantage is that grid cells, as small areas, raise the potential for disclosing information on population units, as we may find many cells with only a few inhabitants. To address this problem, methods of Statistical Disclosure Control (SDC) for grid maps have been suggested by Behnisch et al. (2013) and de Jonge and de Wolf (2016), among others. Several such methods have been implemented in the R package `sdcSpatial` by de Jonge and de Wolf (2022), which we employ here.
Our contribution in this paper is two-fold: For once we present the results of experiments with SDC methods for grid data, conducted over multiple contributing countries. We utilize "census-like" population grids from Austria (*Statistics Austria*), France (*Insee*), Germany (*Destatis*) and the Netherlands (*CBS*). Risk measures will be computed before and after SDC. Secondly, we test and compare two metrics for the utility loss resulting from SDC: the Hellinger distance and Kantorovich-Wasserstein distance.
The paper proceeds as follows: In section 2 we outline methods for disclosure control applicable to grid data. In section 3 we consider metrics for measuring the utility loss in protected grids. Section 4 describes the setup and results of our multiple-country analysis. We finish in section 5 with some lessons learned and an outlook on areas that may deserve further investigation.

2 SDC Methods

Geographical grid data can be viewed as table cells, so such data could, in theory, be published as a large table. However most often grid data is plotted on a cartographic map, to show spatial patterns and to make more explicitly use of the geo-referencing character of the data. Because of this dual character of publishing grid data, different SDC methods are available. Some of these methods can be categorized as tabular approaches, where grid data is first represented as table cells, and then the resulting secure table is plotted on the map. One example of such a method is cell-suppression. However, tabular methods often neglect the geographical nature of the grid data, failing to use the spatial neighborhood of a unsafe cell to solve the SDC problem. Spatial statistical disclosure methods try to preserve the geographical utility of grid data. Examples of such methods are the quad tree and smoothing methods. Spatial sdc methods can be applied on a spatial population distribution, variable distribution, probability distribution or mean variable distribution. The experiments in this paper are restricted to the spatial population distribution. In this section we will describe the three methods that can be used to protect the grid data that will be used in our experiments.

2.1 Cell removal

Whenever cells are unsafe to publish (for definition of the risk we used in the experiments, see section 3), the cell is not published. Essentially this is a method that is applied to a tabular representation of the grid data, it suppresses the sensitive cell and its value. When plotted on a map, it means that an unsafe cell does not get a colour corresponding to its value. It is suggested to use a specific colour for ‘not available’ (NA), to distinguish between cells without any observation and cells that are not published.

2.2 Quad tree

The quad tree method implemented in the R package `sdcSpatial` generalizes the method as described by Suñé et al. (2017). The method reduces sensitivity by aggregating sensitive cells with its three neighbours, and does this recursively until no sensitive cells are left or when the specified maximum zoom level has been reached. Given the origin of the raster, grid cells are defined for each level of detail a priori. Each grid cell at a certain level is constructed by combining four grid cells of one level down (more detailed). See figure 1 for an example of a priori defined grid cells at three levels of detail.

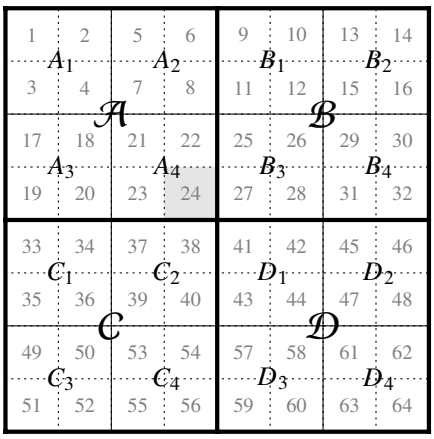

FIGURE 1. Example of three levels of grid cells

Assume that grid cell number 24 at the most detailed level is unsafe (the grayed cell in figure 1). The quad tree method then looks one level up to find a less detailed grid cell that contains the problematic lower level grid cell. In this case it would find grid cell A_4 that consists of the four grid cells $21, \ldots, 24$ at the more detailed level. In case cell A_4 is still unsafe, the method again goes one level up and finds grid cell $\mathcal{A}$ that consists of grid cells $A_1, \ldots, A_4$ of the previous level of detail. This process continues until either a safe grid cell is found or the highest level of the a priori defined grid cells is reached. The most detailed level cells belonging to a safe less detailed cell will share the same adjusted value, which aggregates to the sum of the original cell values. Note that the aggregation process thus depends on the a priori chosen grid cells at the different levels.

2.3 Smoothing

Spatial smoothing uses the spatial structure of grid data so that the values of neighboring cells help to protect sensitive values. In the examples of this paper we are interested in the population density as function of the location. We will denote this density as $f(x, y)$, where (x, y) is a location, i.e., a point in an area $\mathcal{A} \subset \mathbb{R}^2$. The population in region $\mathcal{A}$ can then be seen as a sample from that population distribution, resulting in N observations $(x_i, y_i) \in \mathbb{R}^2$ for $i = 1, \ldots, N$. Each observation is then the location of an individual.
A non-parametric estimator of the population density can be obtained using kernel smoothing (see e.g., Wand and Jones, 1994). The approach used in this paper follows the kernel density smoothing implementation of `sdcSpatial`. That is, the mass of the observed population is spread out over a neighbouring region by means of a Gaussian kernel. The estimate of the population density at point (x, y) is then the sum of the spread out mass at that location:

$$\hat{f}_h(x, y) = \frac{1}{h^2} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}, \frac{y - y_i}{h}\right) \tag{1}$$

where $K(x, y) = (1/2\pi) \exp\left(-(x^2 + y^2)/2\right)$ is the bivariate Gaussian kernel. The bandwidth h determines the size of the region over which the mass is smoothed out.

Note that in the current implementation in `sdcSpatial` the bandwidth is a constant value for all locations, and that the same bandwidth is used in both dimensions resulting in a symmetrically scaled kernel.

3 Risk and Utility measures

3.1 Risk assessment

Denote the area of interest by $\mathcal{A} \subset \mathbb{R}^2$ (for example the national territory or a subsection thereof). We consider N population units, geographically identified by their planar coordinates $(x_i, y_i) \in \mathcal{A},\ i = 1, \ldots, N$. The geographic grid constitutes a tiling of $\mathcal{A}$ into very small subareas of grid cells, which we denote by $C_j,\ j = 1, \ldots, M$. The cell-level count is the number of population units located in a given cell, i.e.

$$r_j = \sum_{i=1}^{N} \mathbb{1}\left[(x_i, y_i) \in C_j\right] \qquad \forall j = 1, \ldots, M$$

where $\mathbb{1}[\cdot]$ is the indicator function. We consider here a straightforward minimum count criterion, by which a cell is considered *at risk*, if it contains fewer than $k \in \mathbb{N}^+$ units (cf. de Wolf and de Jonge, 2017). The risk indicator for a cell is $R_j(k) = \mathbb{1}\left[r_j < k\right] \quad \forall j = 1, \ldots, M$. Two global risk measures can then be defined as the share of cells at risk $R^{(C)}$ and the share of population at risk $R^{(N)}$ for which we have:

$$R^{(C)}(k) := \frac{1}{M} \sum_{j=1}^{M} R_j(k) \qquad \text{and} \qquad R^{(N)}(k) := \frac{1}{N} \sum_{j=1}^{M} R_j(k) \cdot r_j \tag{2}$$

3.2 Utility assessment

3.2.1 A map seen as a table. To assess the loss of information due to a protection process, we have to compare the original map with the protected one. In a first approximation, a map can be seen as a table where the cells are described by the polygons and filled in with the count (population or households for instance) displayed in the map. Then, each utility metric relevant for frequency tables is relevant for maps. We choose the Hellinger distance for this purpose (See Shlomo (2007) for instance). A small distance implies small distortion and is therefore generally desirable.
Denote the original (unprotected) raster by $\mathbf{R}$ and the protected version by $\mathbf{R}'$, each having the same M cells with values r_j and $r'_j,\ j = 1, \ldots, M$ respectively. The Hellinger distance on the interval $[0\,, 1]$ is:

$$HD(\mathbf{R}, \mathbf{R}') = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{M} \left(\sqrt{\frac{r'_j}{\sum_{j=1}^{M} r'_j}} - \sqrt{\frac{r_j}{\sum_{j=1}^{M} r_j}} \right)^2} \tag{3}$$

3.2.2 How to assess the distortion of spatial patterns? A metric such as the Hellinger distance does not take into account the spatial distribution of the units. And, while releasing perturbed maps, to ensure that we haven't perturbed too much the spatial distribution is at least as important as to ensure a good preservation of the values cell by cell.
When the original phenomenon displayed on the map is characterized by a spatial dependency, we'd like to ensure that this dependency is not broken by the protection process. For this purpose, M. Buron & M. Fontaine suggest to compare the Moran's I for the two maps (See Buron and Fontaine (2018)). The Moran's I measures the intensity of the spatial autocorrelation of the phenomenon. Then, the more the coefficient is distorted, the more the spatial information of the phenomenon is lost. With some protection process, the spatial autocorrelation will be automatically reduced (noise injection) and with other ones it will be automatically reinforced (smoothing).

It is less unpredictable for other ones (cell suppression for instance). We, then, tend to prefer SDC methods that preserve this coefficient as well as possible.
In the same perspective to focus on the main spatial patterns of the map, de Wolf and de Jonge (2017) suggest a utility metric able to monitor the distortion of cold and hot spots due to the protection process. Actually, three utility measures are suggested, one related to the location of the spots, one related to their shape and the last one related to their size.
For this paper, we suggest to use the *Kantorovic-Wasserstein Distance (KWD)* as it is implemented in the `SpatialKWD` R package. This distance, also called Earth Mover Distance, comes from the transportation problem: What is the minimal cost to transport a mass from one distribution to another? Whereas the Hellinger distance, as many other ones, could be considered as a "bin-by-bin" distance, the KWD can be viewed as a "cross-bin" distance: each bin (our grid cells) is put in relation with all other ones (Ricciato (2023)).
Formally, KWD is defined as the solution to an optimization problem: We shift around distribution mass of Δr_{jk} between the jth and kth grid cell, until $\mathbf{R}'$ is transformed into $\mathbf{R}$ (or the other way around).[1] Each such shift is evaluated with costs equivalent to the distance covered, denoted by d_{jk}. Our implementation uses the Euclidean distance between cell centroids, measured in multiples of the cell width. We need to solve:

$$
\begin{aligned}
KWD(\mathbf{R}, \mathbf{R}') = \min_{\Delta r_{jk}} \quad & \frac{1}{\sum_{j=1}^{M} r_j} \sum_{j=1}^{M} \sum_{k=1}^{M} \Delta r_{jk} d_{jk} \\
\text{s.t.} \quad & \sum_{k=1}^{M} \Delta r_{jk} = r_j && \forall j = 1, \ldots, M \\
& \sum_{j=1}^{M} \Delta r_{jk} = r'_k && \forall k = 1, \ldots, M \\
& \Delta r_{jk} \geq 0 && \forall j, k = 1, \ldots, M
\end{aligned}
\tag{4}
$$

The following example gives an intuition on how KWD takes the spatial distribution of error into account. Consider a 4×4 grid, symbolized by $\mathbf{A}$, which maps the ground truth.

$$
\mathbf{A} = \begin{matrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}
$$

$\mathbf{B_1}, \ldots, \mathbf{B_4}$ are different alterations, loosely corresponding to protection mechanisms. For instance, $\mathbf{B_1}$ and $\mathbf{B_2}$ are aggregations, whereas $\mathbf{B_3}$ and $\mathbf{B_4}$ are randomly shifting some cell values.

$$
\mathbf{B_1} = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}, \quad
\mathbf{B_2} = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}, \quad
\mathbf{B_3} = \begin{matrix} 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}, \quad
\mathbf{B_4} = \begin{matrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{matrix}
$$

Intuitively, we will tend to prefer $\mathbf{B_1}$ over $\mathbf{B_2}$, since while both come with the same level of aggregation, the former preserves the original's center of gravity, while the latter does not. Similarly, considerations of utility would lead us to prefer, in most cases, $\mathbf{B_3}$ over $\mathbf{B_4}$: while both shift two cell values, the first shifts each of them only to direct neighbors, the other relocates to the edges of the map. It is easy to verify, however, that bin-by-bin utility measures, such as the Hellinger distance, do not cover this rationale as we get $HD(\mathbf{A}, \mathbf{B_1}) = HD(\mathbf{A}, \mathbf{B_2})$ and $HD(\mathbf{A}, \mathbf{B_3}) = HD(\mathbf{A}, \mathbf{B_4})$. They are ignorant of geographic considerations. The Kantorovic-Wasserstein distance, on the other hand, gives for the above cases the desired $KWD(\mathbf{A}, \mathbf{B_1}) < KWD(\mathbf{A}, \mathbf{B_2})$ as well as $KWD(\mathbf{A}, \mathbf{B_3}) < KWD(\mathbf{A}, \mathbf{B_4})$.

[1] Being a proper distance, we have $KWD(\mathbf{R}, \mathbf{R}') = KWD(\mathbf{R}', \mathbf{R})$.

Our choice of the KWD was in part inspired by Ricciato and Coluccia (2023). The `SpatialKWD` package provides an implementation of such a distance to compare two maps, a map being seen as a 2D-distribution. As its exact computation is very time-demanding, the package implements a "very tight approximation", following the work of Bassetti et al. (2020).
Several considerations have to be made before using this tool in an effective and consistent way.

- Which distance to use for assessing the cost of transporting a value from point A to point B? The Euclidean distance is the default, but actually the choice depends on the physical interpretation of this transportation. For instance, if we were comparing a map of residential locations with a map of working locations, the physical interpretation of the transportation problem is related to commuting and a travel time distance could be more appropriate. In the case of an SDC protection process, the physical interpretation is not so obvious. Since we consider the use case of a thematic map, we choose, as a first approximation, to interpret transportation analogous to eye movement of an observer, for which the Euclidean distance seems a sensible pick.
- Geographical areas are not always convex. In that case, the Euclidean distance could be replaced by the least internal path to join two points of the area. Here, we choose to keep the Euclidean distance even in non-convex areas.
- The KWD is fitted to compare two maps displaying the same total mass. Some protection methods (suppression for instance) modify the total mass displayed. One way to deal with this is to set a fixed per-unit mass penalty cost as the maximum distance between two points in the map for the remaining or lacking mass. Another is to virtually re-insert missing mass before computing KWD at one or more sensible proxy-locations.

3.3 Focus areas

The spatial patterns to compare depend on the extent of the map. The larger the map the more complex the spatial patterns. Furthermore, a user will rarely view the whole map, but rather show interest in some subsection, for instance their home region. It therefore makes sense to base utility measures not (only) on large-scale maps, but to consider a selection of smaller maps. We call any subarea $\mathcal{A}_i \subseteq \mathcal{A}$ a *focus area*. When assessing risk and utility metrics for one such focus area, we consider only the corresponding part of the population grid, i.e. the set of cells $\{j = 1, \ldots, M : (x_j, y_j) \in \mathcal{A}_i\}$, where (x_j, y_j) are the planar coordinates of the center point of the jth grid cell. For simplicity, we employ here only square focus areas that are at the same grid resolution as the overall area.
The R package lets us choose to compute the Kantorovic-Wasserstein distance only on some $\mathcal{A}_i$, while taking into account the whole $\mathcal{A}$. Hence, the transportation of masses is computed only within the focus area, but, if needed, some masses can be transported from inside to outside and vice versa.

4 Experiments

The aim of the experiment is to protect a map of the number of persons or households per grid cell. The grid is based on the INSPIRE (2014) standard ETRS89-LAEA for geographic grid systems. This is done for different countries. The protection will be applied using the R package `sdcSpatial` and results are compared across countries using risk and utility measures.

4.1 Datasets

We start from a microdata set containing persons or households as well as X and Y coordinates and raster cells which follow the INSPIRE (2014) standard. Depending on the country the grid cells are either 500m × 500m or smaller. Table 1 shows example data from the Austrian use case. For reasons suggested above, we focus our

analysis on selected regions with specific characteristics regarding terrain and population distribution. Table 2 gives an overview of the focus areas used for each contributing country. Below we give a short overview of the characteristics of each contributing microdata set by source country.

TABLE 1. Example of microdata from Austrian use case

PID	L000500	Y	X
00000001	500mN28215E46275	2821500	4627500
00000002	500mN28085E47890	2808500	4789000
00000003	500mN28025E47925	2802500	4792500
00000004	500mN28120E47985	2812000	4798500
00000005	500mN27605E47900	2760500	4790000
00000006	500mN28040E47940	2804000	4794000

4.1.1 CBS. For the CBS application the population register for 2020 was used, which was extracted from the social statistical system of Statistics Netherlands (CBS). The dataset includes amongst other variables age, sex and educational attainment on a 100m × 100m geographic raster. Each registered inhabitant is assigned to a raster cell. To facilitate analysis the raster was coarsened to a grid of 500m × 500m.
Spatial distributions for urbanized and rural areas are very different and often take different tuning parameters. To indicate how the statistical disclosure methods differs between different regions, four different regions within the Netherlands were selected. First, the capital city Amsterdam, which is a densely populated and urbanized area. Second, the medium sized and young city of Almere where many inhabitants are commuters to different cities. Urbanized, but enclosed by nature and rural areas. Third the rural area of Drenthe, which is sparsely populated and last the region Parkstad, which contains cities and rurals area near the Aachen region in Germany.

4.1.2 Destatis. For the German application Census 2011 results for persons were used. Demographic variables like age, sex or religion are collected on a fine-grained 100m × 100m geographic raster. Household addresses are used for the assignment. Subsequently, each person is located at the centroid of its assigned raster cell. The cells are coarsened to the 500m × 500m resolution for analysis.
Focus areas are chosen to reflect a range of spatial structures. The first is the Ruhr valley area, composed of a cluster of multiple, integrated cities and homogeneously high population density. The second are the twin cities of Mainz and Wiesbaden with their corresponding surroundings. They form a diverse collection of urban, rural, forest and river parts. The third focus area is centered between the town of Stralsund and the island of Rügen at the Baltic Sea coast; it is characterised by an intricate mix of settlements and uninhabitable water areas. Finally, a map section close to the Alps in the historical region of Allgäu is chosen, in which farming and small settlements create an overall homogeneous, low population density.

4.1.3 INSEE. For the French use case, we used the 2017 tax data from the so-called Filosofi system. The data are available on the website of Insee[2]. Socio-demographic information are displayed on a 200m × 200m squares grid. We focused our analysis on the department of La Réunion. We chose to focus our attention on 4 areas, three dense areas picked along the coast of the island (Saint-Denis, Saint-Gilles and Saint-Pierre) and one sparse area picked in its center (La Plaine) (See figure 2). Saint-Denis in the north-east of the island is the most dense area with nearly 700 inhabitants per km^2 whereas the density of the area called La Plaine, in the rural and steeper center of the island, is less than 40 inhabitants per km^2. The areas of Saint-Gilles (north-east) and Saint-Pierre (south-east) have been chosen so as to include a large conurbation of cities around a populated center. Hence, both are quite larger than the two others and mix urban centers and some rural areas connected to them.

[2] https://www.insee.fr/fr/statistiques/6215138?sommaire=6215217

TABLE 2. Focus areas

country	focus area	size (cells)	size (km^2)	initial risk % cells	initial risk % pop.
AT	Vienna & Suburbs	85×85	1806.25	10.09	0.03
	Bregenz	39×39	380.25	9.57	0.08
	Alps in Tyrol	73×73	1332.25	17.10	0.59
	Krems an der Donau	41×41	420.25	12.38	0.28
DE	Ruhr valley	55×55	756.25	3.9	0.02
	Mainz & Wiesbaden	41×41	420.25	9.2	0.04
	Strelasund region	75×75	1406.25	22.2	0.64
	German Allgäu	55×55	756.25	24.4	1.06
FR	Saint-Denis	45×45	81.00	31.4	2.43
	Saint-Pierre	109×109	475.24	49.1	8.63
	La Plaine	41×41	67.24	72.2	31.63
	Saint-Gilles	71×71	201.64	51.0	10.31
NL	Amsterdam	59×46	678.50	12.1	0.04
	Almere	47×42	493.50	13.9	0.07
	Drenthe	89×111	2469.75	21.7	0.64
	Parkstad	31×44	341.50	11.1	0.09

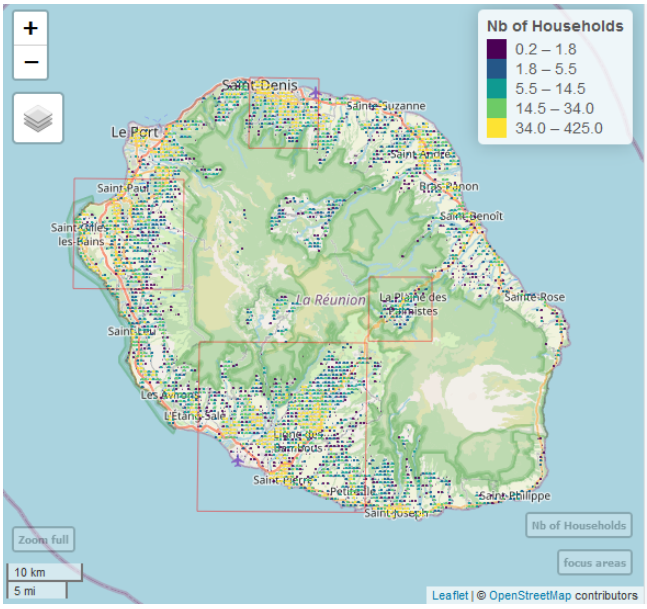

FIGURE 2. Number of Households by 200×200 meters grid squares in La Réunion. Red squares are the focus areas

4.1.4 Statistics Austria. For the use case at Statistics Austria we used the Austrian *Rich-Frame*. This internally compiled data set contains the reference frame for persons and households and is used as the sampling frame to draw a number of surveys. The data set is compiled from various administrative data sources and holds information for each person from the Central Population Register (such as age, gender, citizenship), the Building and Housing register as well as income-, tax-, education- or employment related information. Furthermore, it contains regional information at a very detailed geographical level. The data set is pseudo-anonymized which means that no direct identifying variables or coordinates are contained in the data set. The frame is updated quarterly and it takes around 6 weeks after a quarter change until the data set for the reference quarter has been finalized. For our use case we chose 4 focus areas. The first is the capital *Vienna* including the surrounding suburbs. The second is *Bregenz* including bordering municipalities containing the country border and the border to the Bodensee. Thrid and furth are regions with lower population density, namely part of the *Alps in Tyrol* and rural area in lower Austria including *Krems an der Donau* and surrounding municipalities.

4.2 Application

For applying the protection methods we mainly used the R package `sdcSpatial`. Example code can be found on https://github.com/sdcTools/sdcSpatialExperiment.
Our cases applied the following protection methods:

- Cell removal ~ `sdcSpatial:::remove_sensitive`
- Quad tree protection ~ `sdcSpatial:::protect_quadtree`
- Kernel density smoothing ~ `sdcSpatial:::protect_smooth`

There are slight differences in the use cases of the different countries:

- *CBS* For the Dutch (NL) focus regions, a minimum of 5 contributors in a grid cell was used to test for sensitivity. Four protection methods have been applied
  - Suppression of sensitive cells ('removal')
  - Quadtree with zoom levels 2 ('quad tree I') and 3 ('quad tree II')
  - Spatial smoothing with a Gaussian kernel and a bandwidth of 500m ('smoothing')
- *Destatis* Risk was assessed by the minimum count criterion, where a cell is considered sensitive if it contains fewer than 5 persons. Four protection methods were considered for the German data set:
  - Suppression of the sensitive cells ('removal');
  - Quadtree with a maximum zoom of 2 ('quad tree I') and 3 ('quad tree II');
  - Spatial smoothing with a Gaussian kernel and a bandwidth of 500m ('smoothing').
- *INSEE* For the French case (FR), four protection methods have been considered:
  - Suppression of the sensitive cells ('removal');
  - Quadtree with two different maximum of zoom of 3 ('quad tree I') and 4 ('quad tree II');
  - Spatial smoothing with a Gaussian kernel and a bandwidth of 200m ('smoothing').
- *Statistics Austria* Grid cells with a cell count below 5 were considered sensitive. For the use case at Statistics Austria the following protection methods were considered:
  - Suppression of the sensitive cells ('removal');
  - Quadtree with two different maximum of zoom of 2 ('quad tree I') and 3 ('quad tree II');
  - Spatial smoothing with a Gaussian kernel and a bandwidth of 500m ('smoothing').

4.3 Results

The initial risk assessment for focus areas is included in table 2 in the two rightmost columns. Tables 3 to 6 show risk and utility measures after applying SDC methods. Columns '% cells' and '% pop.' also include grid cells which were not populated in the original data set.
Notably, while computation times for the Hellinger Distance were negligible, calculating the Kantarovich-Wasserstein Distance for a large number of grid cells will typically be more time-consuming. Other than with

TABLE 3. Results for the Dutch data set by focus area

focus area (NL)	method	residual risk		utility	
		% cells	% pop.	HD	KWD
Amsterdam	removal	0	0	.01	.004
	quad tree I	8.1	0.01	.08	.015
	quad tree II	0.8	< .01	.13	.054
	smoothing	1.1	< .01	.22	.257
Almere	removal	0	0	.02	.009
	quad tree I	15.9	0.03	.09	.018
	quad tree II	1.8	< .01	.13	.054
	smoothing	1.3	< .01	.25	.316
Drenthe	removal	0	0	.06	.080
	quad tree I	13.2	.13	.16	.062
	quad tree II	0.3	< .01	.23	.164
	smoothing	0.6	< .01	.31	.407
Parkstad	removal	0	0	.02	.007
	quad tree I	6.6	0.01	.13	.039
	quad tree II	0	0	.20	.124
	smoothing	0	< .01	.27	.352

TABLE 4. Results for the German data set by focus area

focus area (DE)	method	residual risk		utility	
		% cells	% pop.	HD	KWD
Ruhr valley	removal	0	0	.009	.004
	quad tree I	0.7	< .001	.079	.015
	quad tree II	0	0	.095	.025
	smoothing	0	0	.280	.381
Mainz & Wiesbaden	removal	0	0	.014	.006
	quad tree I	9.5	.012	.124	.034
	quad tree II	0.3	< .001	.227	.162
	smoothing	0	0	.365	.493
Strelasund region	removal	0	0	.057	.176
	quad tree I	21.1	.207	.165	.064
	quad tree II	2.2	.003	.222	.136
	smoothing	0	0	.451	.627
German Allgäu	removal	0	0	.073	.229
	quad tree I	10.6	.161	.188	.083
	quad tree II	0	0	.239	.161
	smoothing	0	0	.415	.625

bin-by-bin measures, it takes longer to calculate a large KWD than a short one. The approximation we used guarantees a result within 1.29% of the true value as per Gualandi (2022).

TABLE 5. Results for the French data set by focus area

focus area (FR)	method	residual risk		utility	
		% cells	% pop.	HD	KWD
St-Denis	removal	0	0	0.111	0.284
	quad tree I	3.48	0.01	0.242	0.196
	quad tree II	0	0	0.244	0.201
	smoothing	0	0	0.237	0.273
St-Pierre	removal	0	0	0.210	1.340
	quad tree I	5.64	0.06	0.310	0.338
	quad tree II	0	0	0.334	0.451
	smoothing	0	0	0.248	0.267
La Plaine	removal	0	0	0.416	1.547
	quad tree I	11.98	0.36	0.429	0.616
	quad tree II	0	0	0.456	0.851
	smoothing	0	0	0.304	0.311
St-Gilles	removal	0	0	0.230	1.167
	quad tree I	10.54	0.1	0.359	0.467
	quad tree II	0	0	0.394	0.655
	smoothing	0	0	0.286	0.340

TABLE 6. Results for the Austrian data set by focus area

focus area (AT)	method	residual risk		utility	
		% cells	% pop.	HD	KWD
Vienna & Suburbs	removal	0	0	0.014	0.008
	quad tree I	0.030	$< .001$	0.086	0.022
	quad tree II	0.002	$< .001$	0.125	0.062
	smoothing	$< .001$	$< .001$	0.273	0.351
Bregenz	removal	0	0	0.042	0.009
	quad tree I	0.032	$< .001$	0.097	0.050
	quad tree II	0.002	$< .001$	0.154	0.178
	smoothing	0	0	0.304	0.419
Alps in Tyrol	removal	0	0	0.056	0.053
	quad tree I	0.042	$< .001$	0.151	0.061
	quad tree II	0.004	$< .001$	0.210	0.147
	smoothing	0.001	$< .001$	0.381	0.577
Krems an der Donau	removal	0	0	0.042	0.029
	quad tree I	0.037	$< .001$	0.174	0.078
	quad tree II	0.002	$< .001$	0.261	0.213
	smoothing	0	0	0.471	0.664

4.4 Discussion

We notice that KWD and HD yield the same rank-ordering of SDC methods and parameterizations between quad tree and smoothing, but judge removal quite differently. Indeed, KWD seems to penalize more the removal method than the quad tree or the smoothing, whereas HD penalizes more the quad tree and the smoothing than the removal.

To understand the different behavior of the two utility measures, consider moving distribution mass of amount Δr between cells. The KWD associated with this change is proportional to the minimum-cost way of reversing it, 'costing' $d\Delta r$. On the other hand, HD and other bin-by-bin measures scale only with Δr, independent of distance. If SDC methods cause changes with wide variations in d, we expect no clear connection between the two types of measures. If, however, d is within predictable bounds, which is always the case when an SDC mechanism acts locally, i.e. shifts mass exclusively (or preferably) within geographic neighborhood, we will see a close association.

Consider, for instance, a single aggregation step of the quad tree method, described in 2.2. It consists of redistributing mass between cells of a four-cell square. The KWD associated with reversing such a step is proportional to $d\Delta r$, where $d \in [1\,,\sqrt{2}]$ is tightly bounded. Such a small variation of d can basically be treated as noise, compared to the amount of mass shifted Δr. A similar intuition applies for the smoothing approach: Distribution mass is 'smeared' out locally, so that shifting it back can be viewed as a localized transport problem, where the weighted distance d is again closely bounded. The bound depends on the kernel bandwidth and tail of the kernel function. Typically we find that for both protection mechanisms (quad tree and smoothing) KWD scales mostly with Δr. Judging them via bin-by-bin utility measures versus the cross-bin KWD metric therefore yields overall similar rankings of SDC methods.

With cell removal, on the other hand, distribution mass is deleted at various points of the map. The KWD depends on how this missing mass is treated. Applying a constant cost c per removed unit of mass implies total KWD of $c\Delta r$, which again would scale up with Δr. The ranking of cell removal in comparison to other SDC method depends then entirely on how we set c, i.e. how we judge the information loss from removing mass compared to shifting it. If we consider instead virtually re-inserting the missing mass, KWD depends on the distances of the virtual bin to the points of the map where mass was deleted. In that case, d can vary widely and we do not expect to see a correlation of bin-by-bin measures and KWD. For instance, in the results for Germany (table 4), KWD preferred cell removal for the two more densely populated focus areas (Ruhr valley, Mainz & Wiesbaden), but judged it second-worst for the two sparsely populated focus areas. In comparison, HD favored removal throughout. Hence, we get a divergence between what difference utility metrics recommend.

Throughout, we find that KWD after smoothing is highest. This does not necessarily disqualify the smoothing approach, however. To see why, consider the setting shown below. Let $\mathbf{C}$ be the ground truth; $\mathbf{D_1}$ is a situation as would result from the quad tree method, $\mathbf{D_2}$ might result from smoothing.

$$\mathbf{C} = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 50 & 0 & 0 \\ 0 & 0 & 50 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}, \quad \mathbf{D_1} = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 25 & 25 & 0 \\ 0 & 25 & 25 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}, \quad \mathbf{D_2} = \begin{matrix} 4 & 4 & 4 & 0 \\ 4 & 25 & 5 & 4 \\ 4 & 5 & 25 & 4 \\ 0 & 4 & 4 & 4 \end{matrix}$$

Upon visual inspection, we can claim that $\mathbf{D_2}$ preserves some *qualitative* properties of $\mathbf{C}$ better: The statement 'distribution mass is centered along the main diagonal', for instance, is easily learned from $\mathbf{D_2}$, but hidden in $\mathbf{D_1}$. The small masses in the former would easily be filtered out visually in a heat map, but they do influence utility metrics. Specifically, for the given example we will find that $KWD(\mathbf{C}, \mathbf{D_1}) < KWD(\mathbf{C}, \mathbf{D_2})$. Generally speaking, the more tail-heavy the kernel and the broader the bandwidth, the higher the measured utility loss from smoothing as compared to, for instance, the quad tree method. Generally speaking, we do not find that our utility metrics cover such qualitative aspects well.

Finally, some practical issues relating to spatial SDC can be learned from Fig.3, which maps a small part of the German data set after protection by three different methods. Overlayed in grey are zones that count as generally *uninhabited*: rivers and lakes as well as some forms of vegetation (forests, swamps, marshes). We

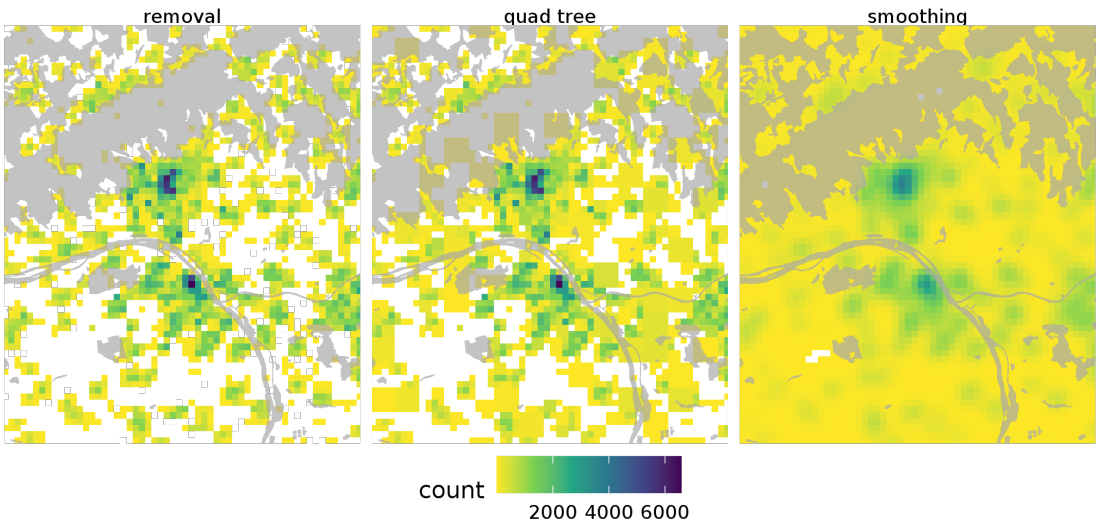

FIGURE 3. Raster maps of a subset of German data after SDC methods have been applied; shaded sections indicate forest and river areas.

can see that quad tree aggregation may intrude into these zones, populating previously unpopulated locations. It can, for instance, aggregate cells from different sides of a river, creating artificial crossings, or make forest areas seem more inhabited than they are. Smoothing, similarly, will often assign positive distribution mass to these implausible locations. Whether this constitutes a problem of consistency for users will depend on the planned application. A more serious question from the point of view of disclosure protection is, in how far knowledge of implausible locations can be used to attack and partially reverse protection methods. In Fig.3 we have used shape files of uninhabited areas, but the opposite is also feasible: many states offer *Open Data terrain models* that explicitly demarcate settlement areas. Outside of these areas, valid person addresses may often be implausible. Overlaying protected raster maps with such auxiliary geographic data could be used to revert some of the changes made. These risks should be further investigated.

5 Conclusion

The experiment we carried out enabled us to test and compare several methods implemented in the `sdcSpatial` package for protecting geo-referenced grid data against the risk of disclosure. The risk we measure here is essentially a re-identification risk, considering a cell to be sensitive as soon as it doesn't reach a certain population threshold. The finer the geographical level of distribution, the more serious the risk. Other types of risk could be investigated in the future. For example, the distribution of several maps on different categories of the population could generate problems of disclosure of group attributes. Another very realistic risk is the risk of differentiation with other maps displaying the same information on different zonings, such as administrative zonings. As these are not, in general, an exact sum of tiles, differentiation cannot be reduced to a problem of nested hierarchies. Costemalle (2019) provides an elegant way of detecting such problems. Future work could involve integrating the suggested analysis of differentiation problems into risk measurement.
The three protection methods suggested by the `sdcSpatial` package (cell suppression, quad tree and smoothing) all have advantages and disadvantages, as summarized in table 7. A suppressive method seems reasonable when the number of sensitive cells is low, particularly in densely populated areas. The quad tree and smoothing methods protect sensitive information by diluting it in the neighborhood. Despite the creation of implausible locations that they both generate, if the zoom factor for the quad tree or the smoothing radius are not too large, the usefulness of the outputs, qualitatively speaking, remains interesting. In addition, smoothing generates less

TABLE 7. Advantages and disadvantages of SDC methods

SDC method	advantages	disadvantages
Cell removal	straightforward and irreversible, hot spots kept intact by design, no artificially inhabited cells	loses mass, low-density regions are deleted from the map
Quadtree	resulting cells assure k-anonymity, small measured distance metrics	overly blocky structure, can artificially enlarge hot spots, can result in implausible locations
Smoothing	secondary utility as a visualization aid, diminishes spatial noise	may lose mass at edges, can result in implausible locations, potential of reversal attacks

noisy maps that are easier to read. It is not so obvious for quad tree. In the future, we could also explore the possibility of combining the different protection methods to increase the usefulness of the outputs.
The `sdcSpatial::protect_smooth` function displays a smoothing with a Gaussian kernel which doesn't take into account the borders or the presence of natural barriers. This could be implemented as in the `btb` package which uses a quadratic kernel, which takes into account only points within the bandwidth. In addition, an edge-correction (a Diggle correction) is implemented in the `btb::btb_smooth` function, to deal with edge-effects (see Sémécurbe et al. (2018)). The method is then more conservative than the one implemented in the `sdcSpatial` package and could lead to improve the utility of the resulting map.
In a future work, we could also try other SDC methods, beginning with some classical ones as swapping or the cell key method for instance. Note that swapping and the cell key method could both be considered as 'pre-map' methods: they are applied to get safe data *before* using a method to plot the data on a map. The Quadtree and Smoothing methods act on the unprotected data directly and supply protection when plotting the data on a map. Moreover, swapping and the cell key method in their basic forms do not take the spatial characteristics into account.
Any method of protection generates a loss of utility. It is therefore a question of choosing the method which, while protecting sufficiently, will be able to preserve the most original information. However, geo-referenced data cannot be assimilated to simple tables because the spatial distribution of the data is information as important as the data themselves. In this work, we thus used two metrics, one appropriate for comparing tables (the Hellinger distance), the other more suitable for comparing maps (the Kantorovic-Wasserstein distance). It was the first opportunity for us to use the, latter which requires a pronounced attention to certain details such as the mass-mismatching or the convexity of the zonings.
We could think about other utility metrics, especially from the ones that can grasp the spatial patterns information, as the Moran's I (Buron and Fontaine (2018)) or as the characteristics of cold and hot spots (de Wolf and de Jonge (2017)).
An additional problem that should be addressed in future work is the relation between risk and utility measures and the resolution of the map in question. The resolution of a map in some sense determines the level to which a user could zoom in on the map. Obviously, the more zoomed-in a user is looking at the map, the more detailed locations could be determined. Thus the identification risk becomes higher. Future work should include recommendations on how to deal with this feature of being able to zoom in on the map, in relation to the disclosure risk and the utility. See for some discussion in this direction e.g., de Wolf and de Jonge (2018).

References

Bassetti, F., S. Gualandi, and M. Veneroni (2020). On the computation of Kantorovich–Wasserstein distances between two-dimensional histograms by uncapacitated minimum cost flows. *SIAM Journal on Optimization 30*(3), 2441–2469.

Behnisch, M., G. Meinel, S. Tramsen, and M. Diesselmann (2013). Using quadtree representation in building stock visualization and analysis. *Erdkunde 67*(2), 151–166.

Buron, M. and M. Fontaine (2018, Oct). *Confidentiality of spatial data* (Insee Methodes ed.)., Chapter 14, pp. 349–373. Paris.

Costemalle, V. (2019, Dec). Detecting geographical differencing problems in the context of spatial data dissemination. *Statistical Journal of the IAOS 35*(4), 559–568.

de Jonge, E. and P.-P. de Wolf (2016). Spatial smoothing and statistical disclosure control. In J. Domingo-Ferrer and M. Pejić-Bach (Eds.), *Privacy in Statistical Databases. UNESCO Chair in Data Privacy International Conference, PSD '16, Dubrovnic, Croatia, September 14-16, Proceedings*, Springer Lecture Notes in Computer Science, LNCS 9867, pp. 107–117.

de Jonge, E. and P.-P. de Wolf (2022). *sdcSpatial: Statistical Disclosure Control for Spatial Data*. R package version 0.5.2.

de Wolf, P.-P. and E. de Jonge (2017). Location related risk and utility. In *UNECE - Expert Meeting on Statistical Data Confidentiality*.

de Wolf, P.-P. and E. de Jonge (2018). Safely plotting continuous variables on a map. In J. Domingo-Ferrer and F. Montes (Eds.), *Privacy in Statistical Databases. UNESCO Chair in Data Privacy International Conference, PSD '18, Valencia, Spain, September 26-28, Proceedings*, Springer Lecture Notes in Computer Science, LNCS 11126, pp. 347–359.

Gualandi, S. (2022). *SpatialKWD: Spatial KWD for Large Spatial Maps*. R package version 0.4.1.

INSPIRE (2014). *Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems, D2.8.I.2 Data Specification on Geographical Grid Systems - Technical Guidelines*. European Commission Joint Research Centre.

Ricciato, F. (2023, Mar). Kantorovich-Wasserstein distances for spatial statistics: The Spatial-KWD library. Presentation at the NTTS 2023 conference.

Ricciato, F. and A. Coluccia (2023). On the estimation of spatial density from mobile network operator data. *IEEE Transactions on Mobile Computing 22*(6), 3541–3557.

Sémécurbe, F., L. Genebes, and A. Renaud (2018, Oct). *Spatial Smoothing* (Insee Methodes ed.)., Chapter 8, pp. 205–229. Paris.

Shlomo, N. (2007, Aug). Statistical disclosure control methods for census frequency tables. *International Statistical Review 75*(2), 199–217.

Suñé, E., C. Rovira, D. Ibáñez, and M. Farré (2017). Statistical disclosure control on visualising geocoded population data using a structure in quadtrees. NTTS 2017.

Wand, M. and M. C. Jones (1994). *Kernel smoothing*. CRC Press.