CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 4
29-31 October 2019, the Hague, the Netherlands 16 September 2019

Comparing methods of safely plotting variables on a map

Y. (Sapphire) Han, Peter-Paul de Wolf & Edwin de Jonge (Statistics Netherlands, The Netherlands)
y.han@cbs.nl, pp.dewolf@cbs.nl, e.dejonge@cbs.nl

Abstract and Paper

In recent years, the cartographic portrayal of quantitative information has been developed as a favourable method of displaying statistics for many users. National statistical institutes take advantage of this development by publishing the spatial distribution of social and industrial statistics. However, privacy concerns arise with the development of spatial distribution maps since they contain geographic identifiers of individuals or entities. Statistical disclosure control and risk measure methods developed in the past decades cannot be readily applied to cartographic plots. A current solution used by spatial distribution maps is, for example, to aggregate the data to large areas and suppress the information in areas that consist of too little number of individuals or entities. This method usually preserves privacy but leads to substantial information loss, which degrades the utility of the cartographic mapping or cluster detection studies. In previous studies, the authors adapted smoothing methods to safely plot variables on cartographic maps. Results from previous studies demonstrated the potential of smoothing method in both providing disclosure protection by smearing out variables over a large area and revealing spatial patterns of variables of interest. The current paper is designed to contribute to the spatial statistical disclosure control methods by comparing different methods in plotting spatial distributions. First, we construct a risk measure of the variable of interest (e.g., energy consumption) as well as a kernel density type of estimator for smoothing detected unsafe areas. Then, we evaluate the utilities of both the smoothing method and a traditional aggregation-suppression method at postal code level. We apply the methods to real data sets.

Keywords

Disclosure risk, Cartographic map, Postal-code region, Spatial distribution, Kernel density type estimator

Comparing methods of safely plotting variables on a map

Y. (Sapphire) Han[*], Edwin de Jonge[**] and Peter-Paul de Wolf[***]

[*] Statistics Netherlands, The Hague, The Netherlands, y.han@cbs.nl
[**] Statistics Netherlands, The Hague, The Netherlands, e.dejonge@cbs.nl
[***] Statistics Netherlands, The Hague, The Netherlands, pp.dewolf@cbs.nl

Abstract: In recent years, the cartographic portrayal of quantitative information has grown to be a favourable method of displaying statistics for many users. National statistical institutes take advantage of this development by publishing the spatial distribution of social and industrial statistics. However, privacy concerns arise with the development of spatial distribution maps since they contain geographic identifiers of individuals or entities. Statistical disclosure control and risk measure methods developed in the past decades cannot be readily applied to cartographic plots. A current solution used by spatial distribution maps is, for example, to aggregate the data to large areas and suppress the information in areas that consist of too little number of individuals or entities. This method usually preserves privacy but leads to substantial information loss, which degrades the utility of the cartographic mapping or cluster detection studies. In previous studies, the authors adapted smoothing methods to safely plot variables on cartographic maps. Results from previous studies demonstrated the potential of smoothing methods in both providing disclosure protection by smearing out variables over a large area and revealing spatial patterns of variables of interest. The current paper is designed to contribute to the spatial statistical disclosure control methods by comparing different methods in plotting spatial distributions. First, we construct a risk measure of the variable of interest (e.g., energy consumption) as well as a kernel density type of estimator for smoothing detected unsafe areas. Then, we evaluate the utilities of both the smoothing method and a traditional aggregation-suppression method at postal code level. We apply the methods to real data sets.

Keywords: Disclosure risk, Cartographic map, Postal-code region, Spatial distribution, Kernel density type estimator

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

1 Introduction

In recent years, the use of spatial mapping of statistics has gained increasing popularity in the field of population (O'Brien & Cheshire, 2016), energy consumption (Ramachandra & Shruthi, 2007), segregation (Raanan & Shoval, 2014) and many others. There are three advantages when presenting statistics on maps. First of all, maps and spatial information is a means of storing data. Large quantities of information can be stored in these maps. Secondly, maps and spatial information can be used to identify and study spatial patterns. Since human beings are often visually oriented, spatial

relationships can attract attention more easily on maps than on tabular data. In addition, maps are effective in presenting information and communicating findings. Maps are useful when conveying information and finding that are either difficult to express verbally or condense messages that would be lengthier to describe in words. These advantages are especially obvious when policy makers make use of the maps to explore and express their regional policies.

As cartographic software has become more powerful, it is now possible to easily create maps containing huge number of datasets using statistical environment such as R or Python. We expect that plotting variables on maps will become more and more common in coming years. However, a spatial distribution plot may be exploited to link information to a single unit of interest. Traditional disclosure risk measures and disclosure control methods cannot be readily applied to this type of data dissemination. De Wolf & de Jonge (2017) proposed disclosure risk measures and utility measures in terms of locations. Researchers have also developed kernel density estimation smoothers as a statistical disclosure methods either on dichotomous variables (de Jonge & de Wolf, 2016) or on continuous variables (de Wolf & de Jonge, 2018). Kernel density estimation smoothers as a novel statistical disclosure control methodology can also help in identifying and reserving spatial patterns (de Wolf & de Jonge, 2018). This is different from traditional statistical disclosure control methodology used by National Statistical Offices such as publishing data on postcode level using statistical disclosure control method such as cell suppression. To our knowledge, there is no research yet comparing the kernel density estimation smoother method with post-tabular methods such as cell suppression in the aspect of protection effects. Therefore, in this research, we aim to compare the differences of using kernel density estimation smoother and post-code map data publication using cell suppression in the sense of utility. This kind of comparison can be also useful when deciding which parameters to choose in kernel density estimation smoother.

The structure of this article is as following. In the coming section, a review of statistical disclosure control theories and applications on map data is presented. The third section is about data and method used in this research. In the result section, primary results of comparisons of the differences using kernel density estimation smoother are presented. At the last section, we discuss the conclusions of this research and suggestions for future research.

2 Statistical Disclosure Control on Map Data

The visualization of statistical output on a map is an effective way to publish geographic-related statistics. However, it also poses challenges for statistical disclosure control. One of the risks is that on a map the location of an enterprise (or an individual) is a direct identifying variable. To implement statistical disclosure control measures, a widely-adopted method of plotting spatial data is to average the variable of interest on predefined administrative regions. These regions can be a province, a municipality or at

a certain post code level: in the Netherlands post code level four or post code level six are used. Then, the mean values of the variables of interest are treated as cell values in a tabular data and standard risk measures developed for tabular data, for example the minimal rule and p%-rule (Hundepool et al., 2012), can be used as risk measures. Once a risky region is identified by the risk measure, one can either apply cell suppression (no color on the risky region) or merge the risky region into a larger administrative region (de Wolf & de Jonge, 2018). This method is straightforward and simple and therefore it is widely used by national statistics offices. A trade-off of this method is the information loss. When zooming-in is not allowed, sub-regions that may contain different values are not distinguishable. This trade-off makes plotting variables on maps less desirable since the spatial pattern may become less obvious.

De Wolf & de Jonge (2018) has constructed a kernel-type estimation smoother as a statistical disclosure method to reduce disclosure risk as well as increase utility by keep the spatial pattern of the variables on maps. By tuning the resolution and bandwidth parameters of their kernel estimation smoother, one can control the percentage of the unsafe grid cells to a desired range.

However, there is no study yet comparing the effect of the proposed kernel-type estimation smoother with traditional administrative region methods. In addition, the tuning of resolution and bandwidth parameters leads to different results in terms of risk and utility. At last, since national statistical offices have been publishing spatial data with traditional administrative region so far. The overlapping area between these two methods may leads to disclosure risks. In this study, we present primary result in comparing the traditional administrative region methods with the kernel density estimation smoother. Our procedure is useful when statisticians are considering choosing between the traditional method and the kernel-type estimation smoother.

More methodology description will be added to this article before the work session in October 2019.

3 Data and Methods

To compare the differences of using kernel density estimation smoother and post-code map data publication using cell suppression in the sense of utility, we apply the proposed methods to a dataset describing the energy consumption of enterprises used previously by de Wolf & de Jonge (2018). The dataset contains detailed energy consumption of enterprises in a region of the Netherlands called ‘Westland’ in 2014. The region of ‘Westland’ is known for its commercial greenhouses as well as enterprises from the Rotterdam industrial area. The ‘Westland’ region belongs to South Holland province (in Dutch: provincie Zuid-Holland) of the Netherlands and contains municipality, namely, Westland, the Hague, Rotterdam and many others. The undisclosed microdata are enterprise data and contain for each enterprise its detailed energy consumption and location ((x-y coordinates). The ‘Westland’ region has 2190 unique post code six units

and in this research only units with complete information on geographic coordinates (x-y coordinates) and energy consumption are taken into accounts. There are in total 7016 enterprises in 'Westland' region selected. Furthermore, there is a published aggregated dataset available in which the p%-rule is applied to the microdata on the zip-code (postcode) level.

The analyses are performed using the R package sdcSpatial developed by de Jonge & de Wolf (2019).

4 Results

In this section, we present primary results comparing post-code six level and kernel density estimation smoother.

First, per post code region, when applying the combination of minimal rule and p% rule, the unsafe cells, i.e., unsafe post code regions are presented in **Tab 1**. The unsafe post code regions are at least more than 60% under the given condition for minimal rule and p% rule. This is because most of the enterprises are large in area and therefore, there are few enterprises in one post code region. The number of unsafe enterprises is much smaller compared to unsafe post code region, as shown in **Tab 2**. However, in regards to safely plotting variables on a map, the number of unsafe post code region is more of a concern than the unsafe enterprises.

Min\ p%	**80**	**85**	**90**	**95**
3	68%	65%	64%	62%
4	75%	74%	73%	72%
5	81%	80%	79%	79%

Tab 1. percentage of unsafe post code regions when applying tabular statistical disclosure control minimal rule and p% rule.

Min\ p%	**80**	**85**	**90**	**95**
3	30%	27%	25%	23%
4	36%	34%	32%	31%
5	42%	41%	40%	39%

Tab 2. percentage of unsafe enterprises when applying tabular statistical disclosure control minimal rule and p% rule.

The kernel density estimation smoother is performed on grid cells based on x-y coordinates of the enterprise. The number of unsafe grids based on minimal rule and p% rule also depends on the size of the grid, i.e., resolution. The visualization of plotting the

energy consumption variable on a raster map with a resolution of 500×500 m^2 is shown in **Fig 1**. In **Fig 1**, the number of unsafe grids under the given minimal rule (n = 3) and p% rule (p = 90) is clearly smaller than 64% as in **Tab 1** when plotting the energy consumption variable on a predefined post code area. Since the number of unsafe grids depends on the resolution, the change of unsafe grids of the given data is shown in **Fig 2**. The trend of **Fig 2** is that the number of unsafe grids drops as the increase of resolution. When the resolution is bigger than 300×300 m^2, the percentage of unsafe grids are smaller than the percentage of unsafe post code area (64%) on a map under the condition of minimal rule n = 3 and p% rule p = 90.

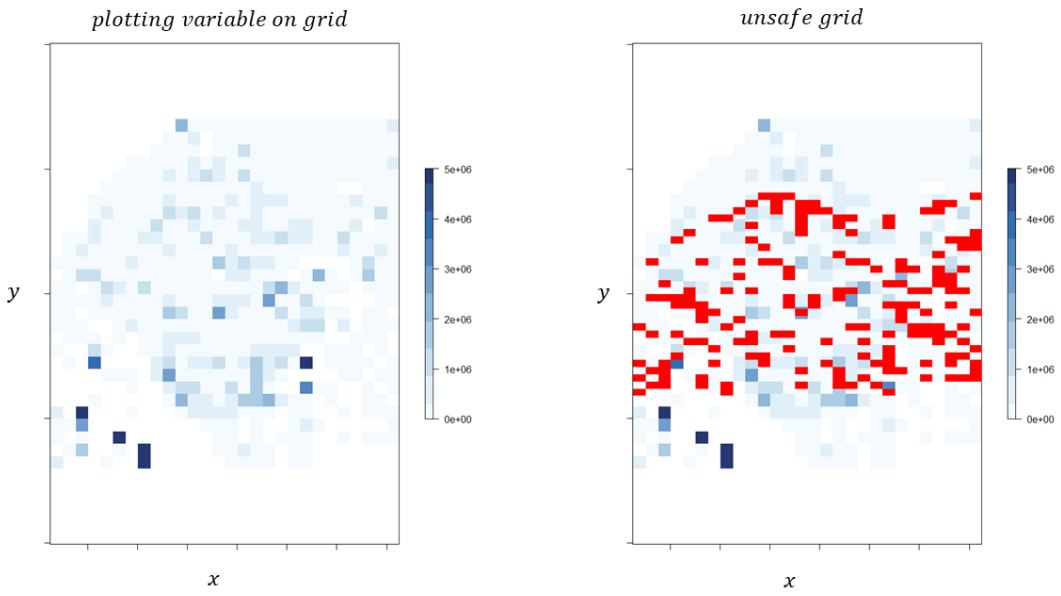

Fig 1. left, plotting energy consumption on a raster map (resolution: 500×500 m^2); right, overlapping the number of unsafe grids (minimal rule n = 3 and p% rule p = 90).

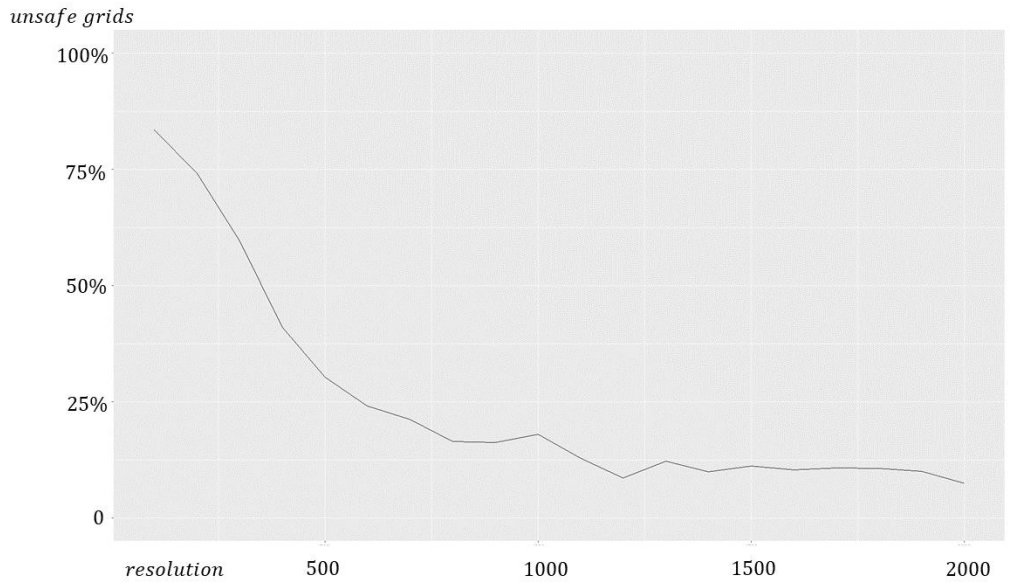

Fig 2. the percentage of unsafe grids change (under minimal rule n = 3 and p% rule p = 90) with different resolutions (from 100×100 m^2 to 2000×2000m^2).

The effect of kernel density estimation smoother proposed by de Wolf & de Jonge (2018) depends on the resolution as well as bandwidth. The change of the number of unsafe grids under the resolution of 500×500 m^2 (as in **Fig 1**), minimal rule (n = 3), and p% rule (p = 90) is shown in **Fig 3**. Without applying kernel density estimation smoother, the percentage of unsafe grids is 30%. One can reduce the number of unsafe grids to 14% when applying the kernel density estimation smoother with a bandwidth of 1000.

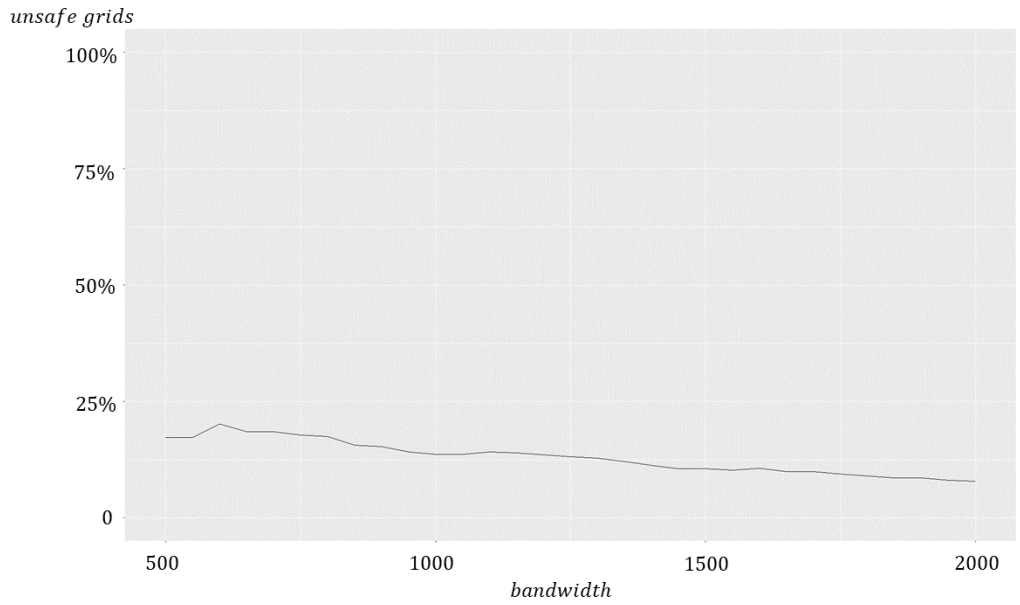

Fig 3. the percentage of unsafe grids change (under minimal rule n = 3 and p% rule p = 90, and resolution of 500×500 m^2) with different bandwidth (from 500 to 2000).

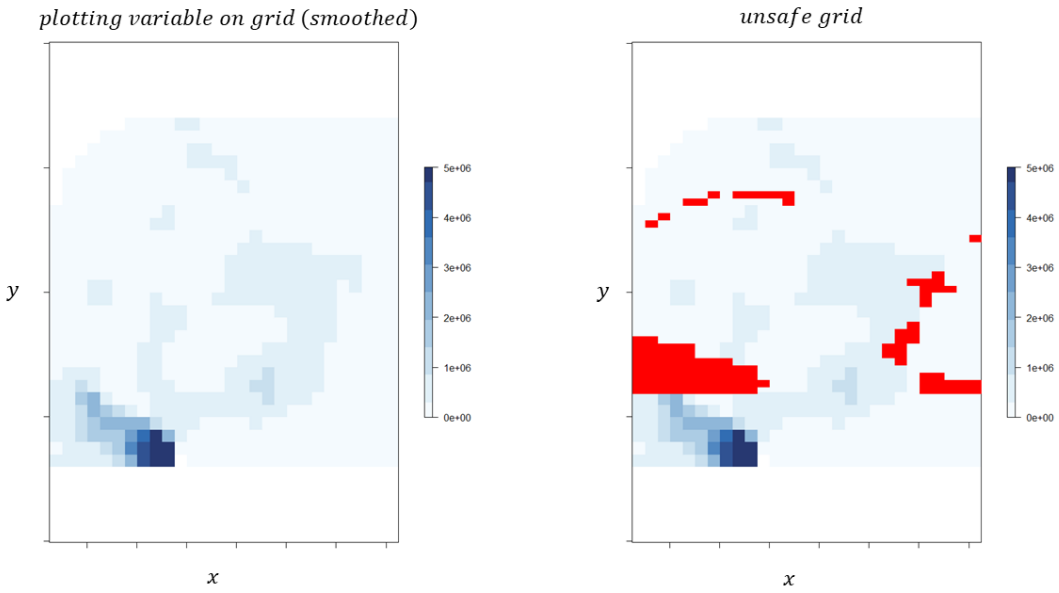

Fig 4. left, plotting the kernel density estimation smoothed energy consumption on a raster map (resolution: 500×500 m^2, bandwidth = 1000); right, overlapping the number of unsafe grids (minimal rule n = 3 and p% rule p = 90).

So far, the number of unsafe cells are compared between the traditional post code region method and the kernel density estimation method. It is shown that the number of unsafe cells are smaller in a raster map than a post code map. In addition, with kernel density estimation smoother, one can reduce the number of unsafe cells to a desired range by tuning the resolution and bandwidth parameters. It is useful to further compare the utility of difference of these two map statistical disclosure control methods. One can cross-tabulate whether one enterprise is classified as safe or unsafe using these two methods. The cross tabulation of the number of safe and unsafe cells is shown in **Tab 3**. Most of the enterprises (72%) are considered safe under the condition of minimal rule n = 3 and p% rule p = 90. There are more unsafe cells (19% more) when plotting the energy consumption on a post code six level map than a raster map with a resolution of 500×500 m^2. When cross-tabulating the safe and unsafe cells of smoothed raster map (resolution of 500×500 m^2, bandwidth = 1000) and post code six level in **Tab 4**, the effect of this kernel density estimation smoother is more observable. 75% of the enterprises are considered safe by both methods. There are less unsafe enterprises when plotting the energy consumption on a smoothed raster maps. This shows that not only on a cell level, on an individual enterprise level, plotting the energy consumption on a smoothed raster map gives higher utility.

Raster map (resolution of 500×500 m^2)\post code six	Safe	Unsafe
Safe	72%	22%
Unsafe	3%	3%

Tab 3. Cross tabulation of safe and unsafe enterprises (in percentage) when plotting the energy consumption on a raster map (resolution of 500×500 m^2) or post code six level under the condition of minimal rule n = 3 and p% rule p = 90.

Smoothed raster map (resolution of 500×500 m^2, bandwidth = 1000) \post code six	Safe	Unsafe
Safe	75%	25%
Unsafe	0	0

Tab 4. Cross tabulation of safe and unsafe enterprises (in percentage) when plotting the energy consumption on a smoothed raster map (resolution of 500×500 m^2, bandwidth = 1000) or post code six level under the condition of minimal rule n = 3 and p% rule p = 90.

More analyses and results will be added to this article before the work session in October 2019.

5 Discussion and Future work

In this research, we set out to compare two statistical control methodologies for plotting variables on a map. A geographic dataset containing energy consumption in the region of 'Westland' in the Netherlands has been used to test the risk measure differences of traditional post code six plotting and the new kernel density estimation smoother methods. Our results show that the smoothed raster map is more effective in plotting enterprise since the number of unsafe cells and the number of unsafe enterprises are both lower than post code six plotting. In addition, the number of unsafe cells (also linked to unsafe enterprises) can be reduced to a desired risk range by tuning the resolution and bandwidth parameters.

The post code level maps are popular among different institutes and there are various spatial data dissemination using this method. When one intends to also implement the kernel density estimation smoother method, it is necessary to consider the cells and enterprises of the edge areas from overlapping.

More discussion will be added to this article before the work session in October 2019.

References

de Jonge, E., de Wolf, P.-P. Spatial smoothing and statistical disclosure control. In: Domingo-Ferrer, J., Peji´c-Bach, M. (eds.) PSD 2016. LNCS, vol. 9867, pp. 107–117. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45381-1 9

de Jonge, E., de Wolf, P.-P. (2019). sdcSpatial: Statistical Disclosure Control for Spatial Data. R package version 0.1.1. https://CRAN.R-project.org/package=sdcSpatial

de Wolf, P.P., de Jonge, E. Location related risk and utility. Presented at UNECE/Eurostat Worksession Statistical Data Confidentiality, 20–22 September, Skopje (2017). https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/3LocationRiskUtility.pdf

de Wolf, P. P., & de Jonge, E. (2018, September). Safely Plotting Continuous Variables on a Map. In International Conference on Privacy in Statistical Databases (pp. 347-359). Springer, Cham.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P. P. (2012). Statistical disclosure control. John Wiley & Sons.

O'Brien, O., & Cheshire, J. (2016). Interactive mapping for large, open demographic data sets using familiar geographical features. Journal of Maps, 12(4), 676-683.

Raanan, M. G., & Shoval, N. (2014). Mental maps compared to actual spatial behavior using GPS data: A new method for investigating segregation in cities. Cities, 36, 28-40.

Ramachandra, T. V., & Shruthi, B. V. (2007). Spatial mapping of renewable energy potential. Renewable and sustainable energy reviews, 11(7), 1460-1480.