Location related risk and utility

Peter-Paul deWolf*, Edwin de Jonge**

* Statistics Netherlands, pp[dot]dewolf[at]cbs[dot]nl

** Statistics Netherlands, e[dot]dejonge[at]cbs[dot]nl

Abstract. Cartographic maps have many practical uses and can be an attractive alternative for disseminating statistics with spatial characteristics. However, a detailed map may disclose private data of individual units of a population. Traditionally, a disclosure risk measure is related to the (distribution of) individual units. When publishing official statistics on a map, the location is the identifying variable and thus information about units is linked with locations. In this paper we try to formulate disclosure risk measures in terms of locations: when is it safe to publish information linked to a certain location?

The other side of the coin is obviously information loss associated with applying some disclosure control method(s). We also formulate some utility measures that may be used to assess the usability of disclosure control methods, specifically targeted at geo-coded information, displayed on a map.

1 Introduction

Visualization of statistical output is a convenient way to disseminate certain statistics. Applied appropriately, it is an appealing way to show statistics to the general public. More specifically, spatial data displayed on a map can be very helpful for regional policy makers as well, since such a visualization could highlight the areas of their interest. Think about detecting areas in a city where the number of people that receive some kind of social benefit is relatively high or about highlighting areas where criminal activity is concentrated. Spatial display of data is used in a variety of research areas, like criminal analysis (Chainey et al., 2002), epidemiology (Gatrell et al., 1996), animal and plant ecology (Dixon and Chapman, 1980 and Worton, 1989), urban modelling (Borruso, 2003) and seismic risk analysis (Danese et al. 2008). In official statistics, displaying statistical information using maps is discussed in e.g., Markkula (2003) and Young et al. (2009). More recently, plotting spatial patterns on a map is becoming popular as well, see e.g., O'Keefe (2012), de Jonge and de Wolf (2016) and Suñé et al. (2017).

A recent project, partly financed by Eurostat, deals (among other things) with disseminating census data by grid squares of size $1\,\mathrm{km}^2$. In that project the approach

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

The authors like to thank Eric Schulte Nordholt for reviewing this paper.

is essentially to deliver tabular data where the grid squares are the table cells. In our current paper we will however concentrate on the dissemination of plots of spatial distributions of certain characteristics, not necessarily (and actually preferably not) in the form of tabular data.

In de Jonge and de Wolf (2016) we introduced our first ideas on displaying spatial distributions on maps in a safe way. We introduced a kernel type estimator of the relative distribution of a dichotomous variable. We applied some top- and bottom coding, based on some ad hoc disclosure risk measure. Moreover, we did not define any utility measure that we could use to confront the effects of the proposed disclosure control method. In the current paper we will define some disclosure risk measures and some utility loss measures, that can be used in the context of displaying relative distributions of dichotomous variables on a map.

In displaying spatial patterns, location is a central concept defining the data and is sometimes considered to be either identifying information or sensitive information. In epidemiology, it is often seen as sensitive information: it is sensitive to derive the location of an ill person. In official statistics, the location is often considered to be identifying information: the location is used to identify a specific person. In case the location is considered to be identifying information, we can represent the individuals by their locations. It thus seems logical to define disclosure risk, usually connected to individuals, in such a way that it is connected to locations directly.

We will first try to relate disclosure risk to locations. We will derive some risk measures as functions of areas. We will elaborate a bit on how these measures could be used in the dissemination of spatial statistics. Since the art of statistical disclosure control lies in finding the right balance between high utility and low disclosure risk, we will also derive some utility measures for the plots we have in mind.

2 Disclosure risk

In case a map is the dissemination format of some dichotomous variable (respondents having a certain attribute or not), it seems logical to define disclosure risk related to locations. Obviously, disclosure risk is primarily connected with individual respondents. However, location is a very important identifying variable, hence dissemination of statistical information with the aid of a map directly relates disclosure risk to locations as well. For example, if a certain attribute can be associated with a small area like a house, it implies a high disclosure risk for the people living in that house.

In this paper we consider the publication of statistical information on a map as a dissemination of aggregated information. We do not consider the release of microdata sets that include geo-spatial variables. Actually, most of the existing literature on privacy issues involving a spatial component deals with the release of microdata sets (see e.g., Armstrong et al., 1999) or with plots of individual respon-

dents on a map (see e.g., http://research.cbs.nl/colordotmap/adam/). We on the other hand are interested in disseminating aggregated/smoothed-out statistical information plotted on a map.

Following existing practice, we will look at identity disclosure (disclosing the identity of a single respondent) and attribute disclosure (disclosing an attribute of a respondent). Attribute disclosure can also appear in the form of so-called 'group' disclosure, i.e., disclosure of the same score on an attribute for an identifiable group of respondents.

For use in spatial analysis it is convenient to write the population as $\mathcal{U} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N\}$ with $\boldsymbol{r}_i = (x_i, y_i)$ the representation of element i of the population by its coordinates (x_i, y_i). I.e., $\mathcal{U}$ is a set of points in $\mathbb{R}^2$. This also reflects the notion that the location is an *identifying* variable. Furthermore, let $\mathcal{A}$ denote an area, defined to be a subset of $\mathbb{R}^2$. An area can be anything from administrative regions to a set of points: $\mathcal{A}$ can be a house, a street, a municipality, a grid square, a county, a general polygon, etc. We will first try to link identity and attribute disclosure to a location or area. To that end, we will adopt the notion of spatial relative risk used in epidemiology, see e.g., Bithell (1990), to identify areas where the disclosure risk related to individuals that reside in that area is too high. The fact that we will use *relative* risk (relative to the population density of an area) is induced by the idea that a low number of incidents in a densely populated area yields a lower risk to the people in that area compared to a low number of incidents in a sparsely populated area.

2.1 Identity disclosure risk

Let $n(x, y)$ (or equivalently, $n(\boldsymbol{r})$) denote the areal number density of the population. See Appendix A for the representation of the number densities as a sum of Dirac delta functions. The number of people in a region $\mathcal{A}$ is thus given by

$$\iint_{\mathcal{A}} n(x, y)\, dxdy = \sum_{i=1}^{N} \mathbb{1}(\boldsymbol{r}_i \in \mathcal{A}) = |\mathcal{A} \cap \mathcal{U}|$$

with $\mathbb{1}(B) = 1$ if B is true, zero if B is false, and $|\mathcal{X}|$ the cardinality of set $\mathcal{X}$.

Assume that an attacker has knowledge that the object he is interested in, is located in region $\mathcal{A}$. Moreover, assume that he has no further information to locate the object to a more specific location. This means that the size of e.g., a rural area only indirectly plays a role in the definition of disclosure risk: only the fact that it is rural matters. Under those assumptions, the attacker would randomly assign one of the objects in that region to the object he is interested in and we thus define identity disclosure risk for region $\mathcal{A}$ as

$$\mathsf{R}_{\mathsf{I}}(\mathcal{A}) = \left(\iint_{\mathcal{A}} n(x, y)\, dxdy \right)^{-1}$$

whenever $\iint_{\mathcal{A}} n(x,y)\,dxdy \neq 0$, otherwise the risk is defined to be 0 itself. Note that $\mathsf{R_I}(\cdot)$ has the property that if $\mathcal{A}_1 \supseteq \mathcal{A}_2$ then $\mathsf{R_I}(\mathcal{A}_1) \leq \mathsf{R_I}(\mathcal{A}_2)$, i.e., enlarging the size of the region will not increase the identity disclosure risk.

2.2 Attribute disclosure risk

For attribute disclosure risk, let $\varphi_C(\cdot)$ denote having an attribute C or not (e.g. C denotes having child care). I.e.,

$$\varphi_C(\boldsymbol{r}) = \begin{cases} 1 & \text{if element } \boldsymbol{r} \text{ does have the attribute } C \\ 0 & \text{if element } \boldsymbol{r} \text{ does not have the attribute } C \end{cases}$$

The areal number density equivalent $f_C(x,y)$ is then defined by the property that the number of elements with attribute C in region $\mathcal{A}$ is given by

$$\iint_{\mathcal{A}} f_C(x,y)\,dxdy = \sum_{i=1}^{N} \varphi_C(\boldsymbol{r}_i)\mathbb{1}(\boldsymbol{r}_i \in \mathcal{A})$$

For attribute disclosure risk in frequency count tables, group disclosure would be a useful indicator: whenever a too large group scores on the same sensitive category, it is considered to be attribute disclosure for all units in that group. In our situation that means that, for a certain region, we have to relate the number of units with attribute C to the total number of units in that region. This is similar to the so-called spatial relative risk in epidemiology (see Bithell, 1990). There the spatial relative risk is the ratio of two continuous spatial distributions: the numerator related to the occurrence of a disease at a certain location and the denominator related to the occurrence of a person at a certain location. We could similarly define a spatial relative risk function for having attribute C at location (x,y) by $f_C(x,y)/n(x,y)$. However, using *number densities* as is the case in our situation, this ratio is not properly defined (recall Appendix A). For comparison, in epidemiology the spatial relative risk is actually defined *after* the densities are estimated by continuous versions.

In our situation we could consider each number density to be the empirical density of a sample from an underlying, continuous, density function. This means that the observed population $\mathcal{U}$ is considered to be a sample $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$ from a continuous spatial distribution describing a 'super population'. We will denote the underlying, continuous, densities by $\phi_C(\boldsymbol{r})$ and $\eta(\boldsymbol{r})$. Obviously, we would need the properties that

$$\iint_{\mathcal{A}} \phi_C(x,y)\,dxdy = \iint_{\mathcal{A}} f_C(x,y)\,dxdy \quad \text{and} \quad \iint_{\mathcal{A}} \eta(x,y)\,dxdy = \iint_{\mathcal{A}} n(x,y)\,dxdy$$

We can then define the attribute risk for area $\mathcal{A}$ as

$$\mathsf{R_A}(\mathcal{A}) = \frac{\iint_{\mathcal{A}} \phi_C(x,y)\,dxdy}{\iint_{\mathcal{A}} \eta(x,y)\,dxdy}$$

whenever $\iint_{\mathcal{A}} \eta(x,y)\,dxdy \neq 0$ (i.e., area $\mathcal{A}$ is inhabited) and consider it to be 'undefined' otherwise. In plotting the attribute risk, undefined risk would be plotted transparent to distinguish that situation from the situation where risk is actually zero (when the numerator equals 0 while the denominator is positive). In practice we would need to estimate these underlying densities, e.g., by using kernel density estimates. See e.g., de Jonge and de Wolf (2016) for a first attempt.

2.3 Using the disclosure risk

Note that the identity disclosure risk and the attribute disclosure risk usually get different thresholds. In case of identity disclosure, we completely identify the person and we thus want a very low risk threshold. In case of attribute disclosure (in the interpretation of group disclosure) we 'only' get information on the probability of having an attribute when we know that a person lives in a certain area. For attribute disclosure risk we thus use a high(er) threshold: only in case the majority in a certain area has the attribute, we can assign that attribute with high probability to a specific individual that we know lives in that area.

We now have a way to link disclosure risk (related to a dichotomous variable) to a general area. However, we still need to show how to use this notion when producing statistics on a map. We suggest two approaches:

1. Start with a general estimator of the relative distribution of the dichotomous variable that will be plotted on the map. Then use the introduced notion of disclosure risk to identify the maximum zooming factor.

2. Derive the 'optimal' parameters for the estimator (e.g., the bandwidth in case of a kernel smoothing estimator) using the notion of disclosure risk.

In the first suggestion, we need to specify a partition of the map in different regions, e.g., a grid square. We then can use the disclosure risk measures to identify whether it is feasible to publish the information on the dichotomous variable for each particular element of the partition. This could be done in such a way that certain regions could zoom to a more detailed partition compared to other regions.

In the first suggested situation, the disclosure risk measure could also be used to top-code the estimator: for a particular location (e.g., the location of a mode of the relative distribution) find the smallest neighbourhood (area) of that location with the maximum allowed disclosure risk and top-code the distribution for that area.

The second suggestion would take advantage of the notion of the disclosure risk in constructing the estimator. That way, the estimator would intrinsically deal with the disclosure risk. This type of estimator can be considered to estimate the underlying continuous spatial densities. E.g., in kernel density estimators the choice of an appropriate bandwidth could be driven by disclosure risk considerations. In the next section, we will discuss the utility aspect of estimators of spatial distributions.

We will see that in kernel density estimators the bandwidth plays an essential role in controlling utility as well.

3 Utility

Statistics are meant to be useful and give quantitative insight in phenomena. In case of spatial data a map can be an excellent means to present spatial patterns. Regional policy makers, health care takers and law enforcement often are very interested in spatial differences and spatial distributions. For example a map helps in finding out which locations have a high risk for health care problems, or have a low incident rate of vandalism.

In this paper we focus on estimating the spatial distribution for displaying it on a map. Most cartographic maps display spatial data using either areal or raster data. Areal data are aggregated on administrative regions, e.g. municipalities, provinces. Areal data typically is coarse or unequally balanced, which distorts spatial patterns: the associated areas may wildly differ in size or number of inhabitants causing the data to be less comparable.

Raster data on the other hand, place a grid on the geography and calculate a statistic per grid cell. Each grid cell has the same size. In both methods, the areas used are given and are not constructed taking the (spatial) distribution and disclosure risk of the data into account. The only way to deal with utility and risk is in using certain levels of hierarchical structures defined for those type of areas. E.g., plotting on municipality versus street level or plotting on grid squares of size $1\,\mathrm{km} \times 1\,\mathrm{km}$ versus grid squares of size $100\,\mathrm{m} \times 100\,\mathrm{m}$.

Note that starting with areal or raster data may increase the effect disclosure control methods have on utility. For example, suppressing sensitive values in raster data may yield maps with 'gaps'. Allowing different zoom levels for different regions may result in maps with blocks of different sizes that additionally are typically not aligned with the regional pattern.

To make utility more concrete first a description is given on the qualitative aspects of a spatial distribution map for a dichotomous variable.

3.1 Spatial patterns: hot and cold spots

Utility is in the eye of the beholder. It is difficult to quantify utility, since it depends on the goals a user wants to achieve. The most important aspect for estimating and displaying the spatial distribution of a variable is to allow its users to spot spatial patterns: how does the (relative) frequency vary by location? Most users will be interested in finding high incidence locations –*hot spots*– and low incidence locations –*cold spots*– which allows for further analysis or policy making. A secondary goal often is to find unexpected distributions, e.g. no spatial variation while the demographics are known to be very different, or a level of incidence that contradicts

expectation.

It is desirable to locate hot and cold spots as precise as possible and to determine their size. A policy maker researching household income would rather like to see which streets have a high rate of people living on welfare, then on neighbourhood or city level. *Resolution* is therefore an important aspect of creating a map with high utility. From a utility perspective it is desirable to have high detail and high resolution: it should allow the user to pinpoint hot and cold spots and to determine their sizes. On the other hand, we are interested in the spatial patterns and not in the locations of individuals. This restricts resolution on the high end, since very high resolution will reveal individual objects. Furthermore a spatial density with too much resolution exhibits spatial noise: it highlights the statistical noise instead of showing the spatial distribution. This is in analogy with one-dimensional histograms and kernel density estimators: a too small bin or bandwidth shows statistical noise and obscures the underlying distribution.

A disclosure perspective strives to protect individuals by removing or perturbing details. This may be in conflict with utility. Indeed, as an extreme case, a map in which each location has the same value/colour is a safe map, but has very low utility: it displays a uniform spatial distribution without cold and hot spots.

3.2 Relative frequency maps

From a utility point of view it is tempting to plot the density of incidences on a map. It may seem useful to find out which locations have many incidences, and for some purposes it is. But more often then not, the resulting spatial density is heavily biased by the population density: there are many incidences at locations where many people live, and few incidences where the population density is low. In that sense many density maps are basically population density maps. In many cases, it is more interesting to see the relative frequency on a location. Which locations have a high incidence rate? Users of such a map are often interested in these types of hot spots, because population density hot spots are already known and dealt with. This observation can be paraphrased as 'All frequency maps are alike, each relative frequency map is different in its own way'.

In subsection 2.2 we discussed the use of relative risk as the basis for an attribute disclosure risk measure. Note that the relative frequency by location for a dichotomous variable as discussed from a utility point of view is similar to the (relative) attribute disclosure risk: it calculates for an area the fraction of occurrences relative to the underlying population in that area.

3.3 Bandwidth selection

Spatial point data is often noisy and patterns may only appear after spatial aggregation or smoothing. Consequently, to users who are interested in finding spatial patterns or the location of hot or cold spots, the spatial distribution is a more use-

ful tool. A central question then is: What is the optimal resolution or smoothing parameter for displaying the data? In Davies and Hazelton (2010) two approaches are compared in estimating the relative risk function in epidemiology: using a fixed bandwidth versus using an adaptive bandwidth. In the first approach, a rule of thumb is used to define the 'optimal' bandwidth and in the second approach that rule of thumb is extended to the situation of using an adaptive bandwidth.

This is a general problem of creating a spatial density estimator: there is no unique way for finding the optimal bandwidth for a kernel density estimator, there are only rules of thumb. Most of these guidelines assume normal distributions for the relative risk elements, which is disputable for spatial data. Without defining a specific 'optimal' bandwidth in our current paper, we can still discuss the interaction between utility and disclosure risk when we consider the resolution and the bandwidth of a spatial kernel density estimator plotted on a map. We can postulate having an 'optimal' bandwidth in the sense of utility and determine what it means when the resulting plot is confronted with disclosure risk.

Choosing what bandwidth is most appropriate to increase utility often is an interactive task. It involves tuning the smoothing parameters for the estimation so that it is most clear what the patterns in the data are. In the case considered in the current paper, it means that hot and cold spots are clearly seen. To illustrate the idea, Figure 1 shows two estimates of relative frequencies, using the same data but with different bandwidths. On the left a spatial kernel density estimate is shown

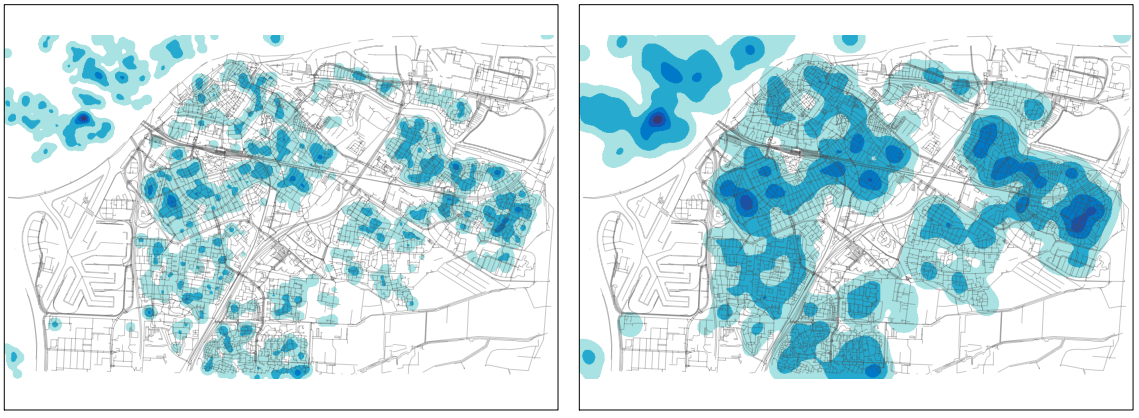

Figure 1: A kernel density estimator with a kernel of 50m (left) and 100m (right)

with a bandwidth of 50m. We see that the overall spatial distribution is rather noisy or 'spotty': it has many hot and cold spots. The map on the right, using a bandwidth of 100m, shows a smoother spatial density in which many spots are merged. Comparing the two maps also reveals some cold spots that are visible on the left but not on the right map. Manual tuning the bandwidth is required to

find the 'optimal' bandwidth, especially since the criterion for 'optimal' seems to be related to the interpretation of the views of the maps.

3.4 Utility loss measures

We assume that cold and hot spots are the most important features of a relative frequency map. Moreover, we assume that the size, the shape and the location of a spot determines the utility of such a map. As usual, we are interested in determining the loss of utility, when disclosure control techniques are applied. When a map is adjusted to control for disclosure risk, there will be a change in the characteristics of the spots: their size will likely be increased, the location may have been shifted and its shape may have been changed.

To determine information loss measures, we will need to be able to compare the utility before and after disclosure control adjustments have been applied. We will assume that the situation before application of Statistical Disclosure Control methods (the reference-map, denoted by $\mathcal{M}^U$) is a map with a relative frequency estimate based on a postulated 'optimal' bandwidth that has been determined by e.g., an interactive procedure. I.e., the map $\mathcal{M}^U$ has 'optimal' utility.

Denote the set with spots on the reference-map $\mathcal{M}^U$ by $\mathcal{S}^U$, denote the map with the relative frequency estimate after Statistical Disclosure Control methods have been applied by $\mathcal{M}^D$ and the set of spots on that protected map by $\mathcal{S}^D$. Note that e.g., top and bottom coding (see also discussion in subsection 2.3) means that the high and low rates are truncated, which results in spots in $\mathcal{S}^D$ with a larger area and shape compared to the same spots in $\mathcal{S}^M$. In the spatial density this is observed in contiguous areas that are top or bottom coded. These areas correspond with the interesting spots. Note that the number of spots on map $\mathcal{M}^U$ may differ from the number of spots on map $\mathcal{M}^D$: like in Figure 1, spots can be merged.

We first need to be able to link spots in $\mathcal{S}^U$ with spots in $\mathcal{S}^D$. We define the location of a spot as the center (or centroid) of the spot and denote it with $\dot{S}_i$ for spot S_i. We will define three utility loss measures related to location, size and shape respectively.

Location

The utility loss measure related to the location of the spots can be defined as

$$\mathsf{U}_{\mathsf{loc}} = \sum_{i \in \mathcal{S}^U} \min_{j \in \mathcal{S}^D} \left\{ \sqrt{(\dot{S}_i - \dot{S}_j)^2} \right\}$$

summing the Euclidean distances between the centroids of the spots on map $\mathcal{M}^U$ with their closest spots on map $\mathcal{M}^D$. Alternatively, one could use the maximum Euclidean distance between the centroids of nearest spots:

$$\tilde{\mathsf{U}}_{\mathsf{loc}} = \max_{i \in \mathcal{S}^U} \min_{j \in \mathcal{S}^D} \left\{ \sqrt{(\dot{S}_i - \dot{S}_j)^2} \right\}$$

Note that the minimum utility loss ($\mathsf{U}_{\mathsf{loc}} = 0$ or $\tilde{\mathsf{U}}_{\mathsf{loc}} = 0$) occurs when the locations of the spots did not change. Which definition matches the intuitive notion of information loss related to the location of spots best, is to be determined empirically.

Size

The utility loss measure related to the size of the spots we define as

$$\mathsf{U}_{\mathsf{size}} = \frac{|\bar{\mathcal{A}}(\mathcal{S}^D) - \bar{\mathcal{A}}(\mathcal{S}^U)|}{\max\{\bar{\mathcal{A}}(\mathcal{S}^D), \bar{\mathcal{A}}(\mathcal{S}^U)\}}$$

with $\bar{\mathcal{A}}(\mathcal{S})$ denoting the average area defining the spots in set $\mathcal{S}$, i.e.,

$$\bar{\mathcal{A}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathcal{A}(S_i)$$

where $\mathcal{A}(S_i)$ denotes the area defining spot S_i. Note that the utility loss measure takes values in $[0, 1]$ and attains it minimum when the average spot-size did not change. Moreover, in practice we expect that $\bar{\mathcal{A}}(\mathcal{S}^D) \geq \bar{\mathcal{A}}(\mathcal{S}^U)$.

Shape

Shape changes can be complex, so we restrict ourselves to looking at the change in circumference of the spots:

$$\mathsf{U}_{\mathsf{shape}} = \frac{|\bar{\mathcal{C}}(\mathcal{S}^D) - \bar{\mathcal{C}}(\mathcal{S}^U)|}{\max\{\bar{\mathcal{C}}(\mathcal{S}^D), \bar{\mathcal{C}}(\mathcal{S}^U)\}}$$

with $\bar{\mathcal{C}}(\mathcal{S})$ denoting the average circumference of spots in set $\mathcal{S}$, i.e.,

$$\bar{\mathcal{C}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathcal{C}(S_i)$$

where $\mathcal{C}(S_i)$ is the smallest circle that contains the complete spot S_i. Note that the utility loss function takes values in $[0, 1]$ and attains its minimum when the average spot-shape did not change. Again, in practice we expect that $\bar{\mathcal{C}}(\mathcal{S}^D) \geq \bar{\mathcal{C}}(\mathcal{S}^U)$.

4 Discussion

In case of disseminating statistics as a relative frequency on a map, the spatial patterns can be very interesting for policy makers. Obviously, detailed information is interesting from a utility point of view, but may jeopardize the privacy of individual respondents. In this paper we focused on disclosure risk and utility loss in case of displaying a relative frequency of a dichotomous variable on a cartographic map.

In our view, the individual respondents can be represented by their location in the situation we describe. Hence, it seems natural to link the disclosure risk and

the utility to locations. We define some disclosure risk measures following the ideas of the relative risk function used in epidemiology. Moreover, we define some utility measures related to so-called hot and/or cold spots: areas on a map where a certain characteristic is concentrated or is absent. Both the disclosure risk and the utility should in our view be considered relative to the underlying population density. The utility is measured using the location, the size and the shape of the hot and/or cold spots.

The introduced measures can be used to either adjust plots based on the raw data to take care of too high disclosure risk or to adjust the estimation procedure itself to construct an estimator that intrinsically takes care of the disclosure risk. The utility loss measures should of course be used to measure the impact of either approach. At the time of writing this paper, we still need to apply our measures to some practical situations to see how the measures behave in practice. Moreover, we need to investigate further the theoretical properties of the proposed measures.

Since the success of visualizing statistics depends on the individual perception of the users, it is difficult to define some mathematical measure to capture the 'true' information or utility of a visualization. Future research could be in that direction: how to capture such subjective opinions about the success of a visualization?

We have restricted ourselves to maps of relative frequencies of dichotomous variables. Obviously, other types of variables can be plotted on cartographic maps as well. It would be interesting to investigate how the proposed measures behave under different settings: continuous variables and multinomial variables.

References

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

Bithell, J.F. (1990), "An application of density estimation to geographical epidemiology", *Statistics in Medicine*, **9**, 691–701.

Borruso, G. (2003), "Network density and the delimitation of urban areas.", *Transaction in GIS*, **7(2)**, 177–191.

Chainey, S., Reid, S. and Stuart, N. (2002) "When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime", in Kidner, D., Higgs, G., White, S. (Eds.) *Innovations in GIS 9: Socio-economic applications of geographic information science*, 21–36, Taylor and Francis.

Danese, M., Lazzari, M. and Murgante, B. (2008), "Kernel densit estimation methods for a geostatistical approach in seismic risk analysis: the case study of

Potenza Hilltop town (southern Italy)", in *ICCSA 2008, Part I*, Springer Verlag Berlin, LNCS 5072, 415–429.

Davies, T.M. and Hazelton, M.L. (2010), "Adaptive kernel estimation of spatial relative risk", *Statistics in Medicine*, **29**(23), 2423–2437.

Dixon, K.R. and Chapman, J.A. (1980), "Harmonic mean measure of animal activity areas", *Ecology 61*, 1040–1044.

Gatrell, A.C., Baily, T.C., Diggle, P.J. and Rowlingson, B.S. (1996), "Spatial point pattern analysis and its application in geographical epidemiology", *Transaction of Institute of British Geographer* 21, 256–271.

Jonge, E. de and Wolf, P.P. de (2016), "Spatial smoothing and statistical disclosure control", *Privacy in statistical databases 2016*, LNCS 9867, 107–117, Springer.

Markkula, J. (2003), "Geographic Personal Data, Its Privacy Protection and Prospects in a Location-Based Service Environment", PhD thesis, University of Jyväskylä, https://jyx.jyu.fi/dspace/handle/123456789/13227.

O'Keefe, C.M. (2012), "Confidentialising maps of mixed point and diffuse spatial data", *Privacy in statistical databases 2012*, LNCS 7556, 226–240, Springer.

Suñé, E., Rovira, C., Ibáñez, D. and Farré, M. (2017), "Statistical disclosure control on visualising geocoded population data using quadtrees", extended abstract at NTTS 2017, http://nt17.pg2.at/data/x_abstracts/x_abstract_286.docx.

Worton, B.J. (1989), "Kernel Methods for Estimating the Utilization Distribution in Home-Range Studies", *Ecology*, **70**(1), 164–168.

Young, C., Martin, D. and Skinner, C. (2009), "Geographically intelligent disclosure control for flexible aggregation of census data", *International Journal of Geographical Information Science*, **23 (4)**, 457–482.

Appendix A: Representation of the number densities

Both $f_C(\cdot)$ and $n(\cdot)$ are defined by their integrals over areas that count the number of elements in these areas. This can be represented as the sum of so-called Dirac delta functions that place infinite mass at certain locations.

Define a uniform distribution on an ϵ-neighbourhood of location $\boldsymbol{r}_0 = (x_0, y_0)$ in $\mathbb{R}^2$ as

$$\delta_\epsilon(\boldsymbol{r} - \boldsymbol{r}_0) = \begin{cases} 1/\epsilon^2 & ||\boldsymbol{r} - \boldsymbol{r}_0|| < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

The Dirac delta function for location $\boldsymbol{r}_0$ can then be defined as $\delta(\boldsymbol{r}-\boldsymbol{r}_0) = \lim_{\epsilon \downarrow 0} \delta_\epsilon(\boldsymbol{r}-\boldsymbol{r}_0)$. This allows us to write n and f_C as

$$n(x,y) = n(\boldsymbol{r}) = \sum_{i=1}^{N} \delta(\boldsymbol{r}-\boldsymbol{r}_i)$$

and

$$f_C(x,y) = f_C(\boldsymbol{r}) = \sum_{i=1}^{N} \varphi(\boldsymbol{r})\delta(\boldsymbol{r}-\boldsymbol{r}_i)$$