
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Hasanzadeh, Kamyar; Kajosaari, Anna; Häggman, Dan; Kyttä, Marketta

A context sensitive approach to anonymizing public participation GIS data : From development to the assessment of anonymization effects on data quality

Published in:
Computers, Environment and Urban Systems

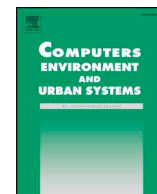
DOI:
[10.1016/j.compenvurbsys.2020.101513](https://doi.org/10.1016/j.compenvurbsys.2020.101513)

Published: 01/09/2020

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Hasanzadeh, K., Kajosaari, A., Häggman, D., & Kyttä, M. (2020). A context sensitive approach to anonymizing public participation GIS data : From development to the assessment of anonymization effects on data quality. *Computers, Environment and Urban Systems*, 83, Article 101513.
<https://doi.org/10.1016/j.compenvurbsys.2020.101513>



A context sensitive approach to anonymizing public participation GIS data: From development to the assessment of anonymization effects on data quality

Kamyar Hasanzadeh*, Anna Kajosaari, Dan Häggman, Marketta Kyttä

Department of Built Environment, Aalto University, PO Box 14100, Aalto 00076, Finland

ARTICLE INFO

Keywords:

PPGIS
Open data
GIS
Anonymization
Spatial
GDPR

ABSTRACT

Use of Public Participation Geographic Information System (PPGIS) for data collection has been significantly growing over the past few years in different areas of research and practice. With the growing amount of data, there is little doubt that a potentially wider community can benefit from open access to them. Additionally, open data add to the transparency of research and can be considered as an essential feature of science. However, data anonymization is a complex task and the unique characteristics of PPGIS add to this complexity. PPGIS data often include personal spatial and non-spatial information, which essentially require different approaches for anonymization. In this study, we first identify different privacy concerns and then develop a PPGIS data anonymization strategy to overcome them for an open PPGIS data. Specifically, this article introduces a context-sensitive spatial anonymization method to protect individual home locations while maintaining their spatial resolution for mapping purposes. Furthermore, this study empirically evaluates the effects of data anonymization on PPGIS data quality. The results indicate that a satisfactory level of anonymization can be reached using this approach. Moreover, the assessment results indicate that the environmental and home range measurements as well as their intercorrelations are not significantly biased by the anonymization. However, necessary analytical measures such as use of larger spatial units is recommendable when anonymized data is used. In this study, European data protection regulations were used as the legal guidelines. However, adaptation of methods employed in this study may be also relevant to other countries where comparable regulations exist. Although specifically targeted at PPGIS data, what is discussed in this paper can be applicable to other similar spatial datasets as well.

1. Introduction

Transparency, openness, and reproducibility are widely recognized as essential features of science (McNutt, 2014; Miguel et al., 2014; Nosek et al., 2015). In theory, most scientists embrace these features as disciplinary norms and values of science (Anderson, Martinson, & De Vries, 2007). However, as widely discussed and reviewed in a number of studies (Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; John, Loewenstein, & Prelec, 2012; O'Boyle, Banks, & Gonzalez-Mulé, 2017), as opposed to what one might expect, these valued features are not yet routine in daily practice of many researchers. A likely culprit for this mismatch is an academic reward system that does not sufficiently incentivize open practices (Nosek et al., 2015). However, this disconnect can also be attributed to the rightfully ever-tightening rules and legislations related to privacy and personal data protection. Adequate

consideration of such concerns poses technical and legal difficulties that may render the idea of open science more problematic than rewarding.

There are various aspects and levels of open science discussed in the literature (Nosek et al., 2015). Nevertheless, this paper focuses on *open data* as one of the standards of a move toward open science. Particularly, this paper focuses on data collected through participatory mapping methods. Participatory mapping approaches, applied in a variety of fields of research and practice, have raised increasing interest during the past two decades. Various terms are applied to describe these approaches, most prominently public participation geographical information systems (PPGIS), participatory GIS (PGIS), and volunteered geographic information (VGI) (Verplanke, McCall, Uberhuaga, Rambaldi, & Haklay, 2016). For simplicity, in this paper we will use the term PPGIS data to refer to data collected through such methods.

PPGIS promotes the use of GIS and modern communication

* Corresponding author.

E-mail address: kamyar.hasanzadeh@aalto.fi (K. Hasanzadeh).

<https://doi.org/10.1016/j.compenvurbsys.2020.101513>

Received 3 February 2020; Received in revised form 9 April 2020; Accepted 3 June 2020

Available online 04 July 2020

0198-9715/ © 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

technologies to engage the general public and stakeholders to inform participatory planning and decision-making, particularly in urban and regional development contexts (Sieber, 2006). A wider user community increasingly adopts participatory mapping applications and scholarly interest in PPGIS is growing, as evidenced by the increasing number of academic publications, conferences, workshops and journal special issues (e.g., Brown & Fagerholm 2015, Brown & Kyttä, 2018; Mukherjee, 2015). Consequently, a large volume of PPGIS data is increasingly available. Given the extensive amount of resources required for any data collection, including PPGIS, a wider community can potentially benefit from open access to such data. However, PPGIS data typically comprises spatial and non-spatial components that may pose risks to the individuals' privacy without proper anonymization.

Although to date there has not been any work done on PPGIS data anonymization, there exists a limited but valuable body of literature on data anonymization from a number of other fields. A small fraction of such literature has focused on spatial data anonymization. Moreover, the literature has rarely investigated the effects of spatial anonymization on data quality. Particularly, there is currently little knowledge available on whether and how the data will be usable after the spatial anonymization. Motivated by the existing opportunities and limitations, this study develops a PPGIS data anonymization approach and empirically evaluates how the anonymized data can be used for further processing and research.

1.1. Research objectives and paper structure

The objectives of this study are threefold. First, this study aims to explore and identify the risks as well as the opportunities in publishing PPGIS data. This is pursued by describing the common characteristics of PPGIS data, identifying the types of personal data, and evaluating the privacy concerns according to The European Union General Data Protection Regulation (GDPR) (European Parliament and Council, 2016).

Second, following the understanding of personal data types and potential privacy concerns, this study aims to develop a safe yet practical PPGIS data anonymization approach and strategy. For doing so, this study reviews the most common data anonymization approaches in the literature and builds upon them to develop a PPGIS data anonymization approach.

Third, in this study we empirically assess how data anonymization can affect data quality. To pursue this objective, we use real data obtained from a PPGIS survey and analyze how the measurements and research findings yielded from the original and anonymized data¹ differ from each other.

The structure of this paper is in line with these research objectives. Accordingly, we will first review PPGIS data characteristics and assess the privacy concerns associated with their publishing according to the legal documents. Next, we will review the literature to find solutions on how we can tackle these privacy concerns and make open PPGIS data possible. Subsequently, we will explain our PPGIS data anonymization method. Finally, we will evaluate how this data anonymization has affected the data quality. At the end, we will discuss the findings and limitations and make some conclusions for future work in this area.

1.2. Personal data and privacy concerns

Legislation concerning data privacy regulations varies between legislative systems. This article discusses privacy concerns and related legislation from the perspective of The European Union General Data Protection Regulation (GDPR) (European Parliament and Council, 2016) implemented since May 2018 and superseding the Data

Protection Directive of 1995 (European Parliament and Council, 1995). However, the best practices on anonymizing and publishing PPGIS data introduced in this study may be applicable in other legislative systems when necessary modifications are considered.

The aim of the GDPR is to protect the rights of natural persons - in relation to the processing of their personal data and to harmonize these rights across the EU member states. The GDPR defines personal data as any information that may lead to the direct or indirect identification of a natural person. Examples of personal data are provided, including, but not limited to, name, location data, online identifier, and factors specific to the persons physical, economic, or social identity (European Parliament and Council, 2016; Article 4). The GDPR defined the rights natural persons have concerning their personal data, including, for example, the right to be informed about the content and processing of the personal data and the right to access, rectify, or erase personal data.

However, the principals of data protection defined by the GDPR do not apply to anonymous information. Anonymity refers to a state where a person can no longer be identified or singled out from the data (European Parliament and Council, 2016; Recital 26). In other words, during an anonymization process the data must be irreversibly processed in such a way that it can no longer be used to identify a natural person by using "all the means likely reasonably to be used" by any party (European Parliament and Council, 2016). Unlike data that is pseudonymized (i.e., personal data is processed in such a manner that it cannot itself be linked to a specific person, e.g. replacing names with number codes), anonymized data guarantees that the individual person cannot be identified when all available additional information on the subject is considered (European Parliament and Council, 2016; Recital 26).

If not for any other ethical reasons or otherwise agreed with the study participants, open sharing and publishing of research data in compliance with the GDPR requires that the preconditions of anonymized personal information are met. The European advisory body on data protection and privacy outlines three criteria for an effective anonymization (Article 29 Data Protection Working Party, 2014):

- *Singling out*, the anonymization must make it impossible to isolate some or all records which identify an individual in a dataset
- *Linkability*, the anonymization must make it impossible to link records relating to an individual
- *Inference*, the anonymization must make it impossible to infer, with significant probability, the value of an attribute from values of a set of other attributes.

1.2.1. Types of personal PPGIS data

A PPGIS survey may include all components of a conventional research survey, with different elements used to collect personal and non-personal information. PPGIS data stands apart from other survey-based data by including the additional component of spatial information created by the study or survey participants. As described by Brown and Kyttä (2014), PPGIS surveys employ spatial elements to locate behaviors, functions, perceptions, or evaluations. Common to these elements is that they are located in the geographic extent of the respondent's everyday life, thus capturing the context the respondents have the most knowledge about through their lived-in experiences. A typical PPGIS dataset could contain, in addition to other survey elements, spatial elements (points, polylines, or polygons) representing spatial phenomena mapped by the respondent, for example, places the respondent visits on a regular basis or perceptions of the environment, such as, places with high natural value or places the respondent perceives as unsafe to visit.

Considering the nature of respondent-created spatial information and the possibility to identify an individual according to the GDPR, we identify three main types of PPGIS spatial data, namely (1) Primary personal spatial data, (2) Group-level spatial data, and (3) Thematic spatial data. These classes relate to different types of mapping tasks

¹ Anonymized data are available online at: <http://doi.org/10.5281/zenodo.3621321>

Table 1
PPGIS data and potential personal information (before anonymization).

	GIS entity type	Likelihood of individual Identification	Recommendations for data anonymization
1. Primary personal spatial data - Residential location(s), second homes	Point	Very likely In areas with low residential density, an individual or the individual's household could be identified from non-anonymized point data. Likely Increased risk of identification when spatial data is linked to other individual-level variables Unlikely In areas with high residential density, individual may be recognized on the level of street address	Always recommended. Increased need for anonymization when the residential location is situated in rural areas or urban areas with low population density, or when the amount of other individual-level variables increases (gender, age, occupation, etc.)
2. Group-level spatial data - locations identifiable to a limited group of individuals, e.g., place of work, university, child's kindergarten	Point, polyline	Unlikely If data is presented as such Likely Increased risk of identification when spatial data is linked to other individual-level variables	Recommended, when spatial data is linked to other individual-level variables.
3. Thematic spatial data - locations with no direct connection to the individual, e.g., environmental perceptions, places related to behavior in public or private spaces visited by a high number of people, such as, shopping centers, parks, etc.	Point, polyline, polygon	Very unlikely If data is presented as such Likely Increased risk of identification when spatial data is connected to other individual-level variables that can be used to infer individual behavior patterns, e.g., activity spaces	Anonymization is rarely needed. Recommended in specific cases, when spatial data is connected to other individual-level variables and patterns derived from thematic spatial data that can be used to identify the individual

differing on whether the respondent may be identified from the spatial data itself or from the spatial data in conjunction with other personal information. These data types and recommendations for their treatment during an anonymization process are introduced in Table 1. These recommendations are based on our interpretation of GDPR and its implications for data types typically present in PPGIS.

1.3. Data anonymization: Existing methods and approaches

Data anonymization is a type of information sanitization process (Saygin, Hakkani-Tür, & Tür, 2005), which aims to protect individuals' privacy and to satisfy the requirement of compatibility with legal and ethical grounds of further processing. Anonymization is a good strategy to keep the benefits while mitigating the risks. For spatial anonymization, this would mean that we provide privacy protection for individual addresses and precise geographical information while maintaining spatial resolution for mapping purposes (Allshouse et al., 2010). Although this is a sensitive and challenging task, its proper implantation can enable open data publication and greatly benefit the scientific community.

Broadly speaking, the existing anonymization techniques fall into two categories of *generalization* and *randomization* (Zhou, Pei, & Luk, 2008). The generalization approach consists of generalizing, or diluting, the attributes of data participants by modifying the respective scale or order of magnitude. For spatial anonymization this would mean that each individual map feature would be generalized into a bigger spatial region that contains at least $K-1$ other users (Ghinita, Zhao, Papadias, & Kalnis, 2010). For example, instead of sharing points representing individuals' homes, one would share the neighborhood, the grid cell, the region, or any other bigger corresponding spatial unit. Such techniques help prevent a data subject from being singled out by grouping them with, at least, $K-1$ other individuals. This is widely referred to as K -anonymity, which not only serves as an anonymization method, but also as an indicator of how effective an anonymization process is (Cassa, Grannis, Overhage, & Mandl, 2006).

Randomization refers to a family of techniques that alter the accuracy of data in order to weaken links between the data and the individuals. This is most commonly accomplished by addition of some

noise to the data (Zandbergen, 2014). For spatial anonymization this would mean that a map feature, for example a point, is displaced to a new location d units of distance away from its original location. Depending on the implementation, the value of d , or its direction, or both can be randomly generated. To control the characteristics of d , the operator may impose conditions for its generation. For example, the operator may define minimum and maximum values for d to control the magnitude of displacement. In this case, the displacement area will be a donut shaped ring according to the minimum and maximum parameters (Allshouse et al., 2010)(Fig. 1).

It is also possible to go further in controlling d by defining a custom function for generating it. For instance, a Gaussian function can be used to make the production of smaller displacement values more likely (Cassa et al., 2006). This can help preserve the overall spatial quality of the data by generally avoiding very large displacements. Nevertheless, satisfactory anonymization may not be achieved with very small displacements. That is why it is important to use mixed approaches. One mixed approach is to use a bimodal Gaussian displacement. This is in essence a combination of the "donut" approach and the Gaussian random generation. In other words, while values are generated using the given function, minimum and maximum conditions can be defined to control the outcomes.

A Gaussian function is blind to the feature's context and hence may not be the most suitable approach for spatial anonymization per se. For example, if we are trying to anonymize the primary personal spatial data of individuals, represented as points in a dataset, a small displacement may suffice in dense urban areas to ensure that the individual can no longer be singled out. On the other hand, in a sparsely populated area, a larger displacement may be needed to effectively deidentify an individual. Therefore, a customized and context sensitive function for randomization is a more promising approach for spatial anonymization.

At the same time, a parallel line of research has occasionally sought alternative approaches for spatial data anonymization. Obfuscation is an example of such alternative approaches that replaces an individual's location with a near-by intersection or building to obscure the real location (Ardagna, Cremonini, Damiani, Di Vimercati, & Samarati, 2007; Duckham & Kulik, 2005). Furthermore, Zhang and colleagues develop a

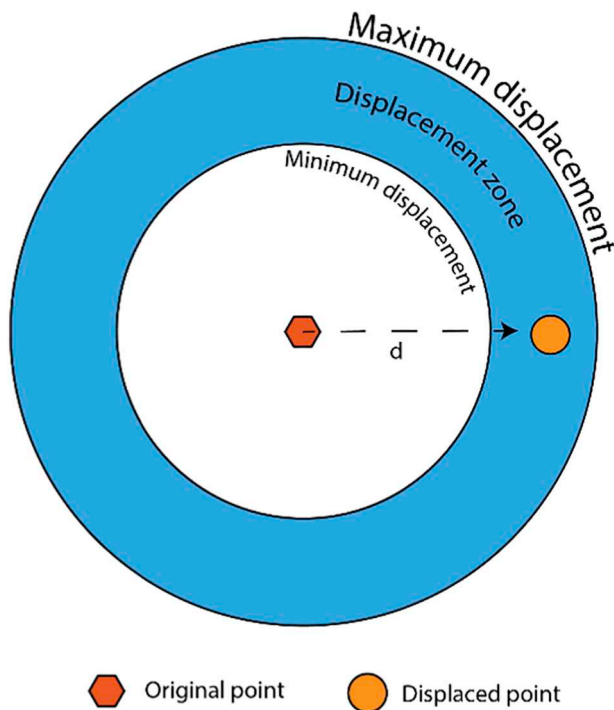


Fig. 1. “Donut” displacement. An exemplary random displacement with minimum and maximum displacement value “d”.

more complex spatial anonymization method called “Location swapping” (2017), which aims to replace an original location with a masked location selected from all possible locations with similar geographic characteristics within a specified neighborhood. A common limitation of obfuscation methods is that they can significantly degrade the quality and usability of the anonymized data. This is especially problematic when there are no appropriate targets around the individual, thus the substitute location will be far from that of the individual. Additionally, the notion of “similar geographic characteristics” in methods such as location swapping can be subjective and hard to define and operationalize.

1.4. Effects of anonymization on spatial analysis results

Research on the effects of anonymization on the analytical results is essential in order to determine whether an anonymization method reaches a meaningful balance between the privacy protection and the ability to derive relevant results and patterns in data (Zandbergen, 2014). Apart from few examples involving visual perception of spatial patterns (Leitner & Curtis, 2004, 2006), the effects of spatial anonymization on data quality has mostly been examined using specific spatial analytical procedures. Examples of such technical procedures include the assessment of anonymization on results from clustering analysis (Cassa et al., 2006; Kwan, Casas, & Schmitz, 2004), built environmental measurements (Clifton & Gehrke, 2013), and kernel density estimation (Shi, Alford-Teaster, & Onega, 2009). Results vary between these studies suggesting that different anonymization processes may affect the utility of different measures and analytical procedures differently. Nevertheless, all findings are in line indicating a consistent tradeoff between the amount of displacement and the accuracy of analytical results. In other words, with larger displacement distances a gradual reduction in the usability of data for certain analytical purposes can emerge. These findings are consistent with the body of literature on the effects of positional errors on geocoded data (e.g. Duncan, Castro, Blossom, Bennett, & Steven, 2011; Zandbergen, 2009).

2. Methods

2.1. Test data

The data was collected using an online PPGIS method that combines Internet maps with traditional questionnaires (Brown & Kyttä, 2014). A random sample of 5000 residents of Helsinki metropolitan area, in Finland, aged between 55 and 75 was obtained from Finnish Population Register Center and an invitation was sent to participants' home addresses in October 2015. The dataset included personal information as described in Table 1. In the survey (Fig. 2), respondents used an online interface to answer a number of questions about themselves and mark their everyday important places. This included their living location as well as their daily destinations such as, leisure and recreational activity places, shopping, services, and sport facilities. There were 1139 responses in total. After deleting incomplete submissions, data from 844

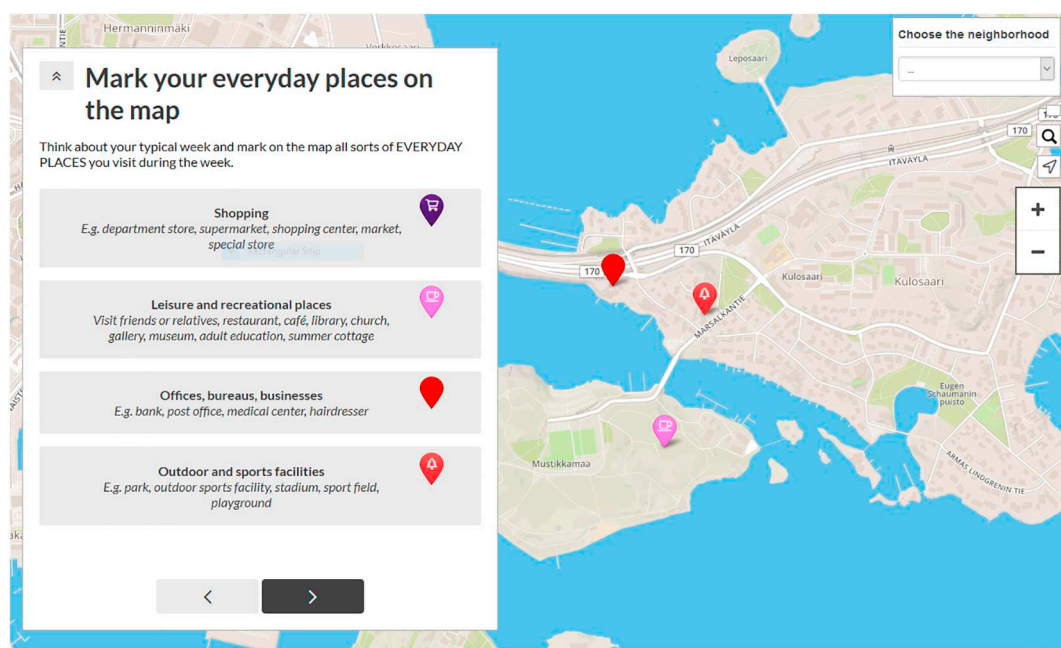


Fig. 2. An instant of ‘Me and my everyday environment’ PPGIS survey (accessible at: www.app.maptionnaire.com/825).

participants was used for the study. A summary of the attributes in the data is presented in Appendix 2.

2.2. Data anonymization

In this study, we develop a unique approach to anonymize personal spatial data while maintaining the overall quality of the datasets. In this approach, group level and thematic spatial data, such as daily destinations, are displaced using a donut spatial anonymization. Accordingly, these points are randomly displaced by a minimum distance of a and maximum distance b , to a random direction. However, for primary spatial data, such as home locations, a more complex approach was implemented. This will be described in the following section.

All the anonymization procedure described below was implemented in Python language using ESRI's ArcPy module. According to legal documents, to ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments" (European Parliament and Council, 2016; Recital 26). Therefore, for safety reasons we do not share the exact anonymization scripts with the reader. However, the detailed description of the procedure as well as the pseudocode provided as appendix should make it possible for the interested reader to easily reimplement the algorithm.

2.2.1. Anonymizing primary personal spatial data

For anonymizing primary personal spatial data, such as home locations, we developed a customized bimodal Gaussian displacement algorithm. The use of Gaussian function was to ensure that the points are displaced as little as possible as long as the minimum anonymization requirements are met. To ensure all points are displaced from their original locations, the Gaussian function was coupled with donut anonymization. Hence, the algorithm ensures that all points are displaced at least by a minimum distance. To ensure deidentification of individuals in sparsely populated areas and to avoid unnecessarily large displacements in dense areas where identification of individuals is reasonably unlikely, the displacements were adjusted according to the population density and feature density around the original location of the feature. Accordingly, features located in less dense areas are more likely to deviate more from their original place than features located in dense areas. Fig. 3 shows the anonymization process.

Mathematically speaking, for each feature f , the displacement D_f was calculated as below:

$$D_f = CM_f \times G_{x,y}$$

Where, G is the Gaussian function output generated separately for x and y coordinates, for a random value of σ , as below:

$$G(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

and CM is the combined multiplier calculated as:

$$CM = C \times [\Delta \times FDM + (1 - \Delta) \times PDM]$$

Where, C is a constant parameter with non-linear effect on model's skewness, and Δ is a constant between 0 and 1 controlling the weight of FDM on anonymization. FDM and PDM are the feature and population density multipliers and are calculated as below:

$$PDM = \frac{\text{Average pop. D.}}{\text{Pop D. at location}}$$

$$FDM = \frac{\text{Average NO. of neighbors}}{\text{NO. of Neighbors}}$$

2.2.2. Calculating k -anonymity

We estimate the expected level of k -anonymity for each individual point by multiplying the local population density by a circular ring area approximation of the Gaussian probability distribution function (Fig. 4) (Cassa et al., 2006). Since 68.26% of individual cases should fall, on average, within the first standard deviation, σ meters in radius from where they were originally located, we can multiply the local population density by the area, $\pi\sigma^2$, and by the probability that the point would have been moved into that area, 0.6826. Subsequently, we add to this the next ring's population density multiplied by its area and its probability that the point would be relocated into that area, 0.2718. Finally, we add the area of the third ring multiplied by its local population density by its probability density, 0.0428. The sum of these three values provides an estimation of k -anonymity achieved for a specific point in a dataset (Cassa et al., 2006). Consequently, k -anonymity achieved for the whole dataset can be calculated as the minimum k -anonymity of all individual points. This, together with other basic statistics measured such as average, maximum, median, and standard deviation of the k -anonymity for all individuals, can provide a computationally tractable expectation of k -anonymity and an overall assessment of anonymization achieved for a dataset.

It should be noted that this estimation of k -anonymity is based on the assumption that no other external knowledge of an individual is available. Therefore, if for instance we know the gender for each individual, assuming that half of the population are males and the other

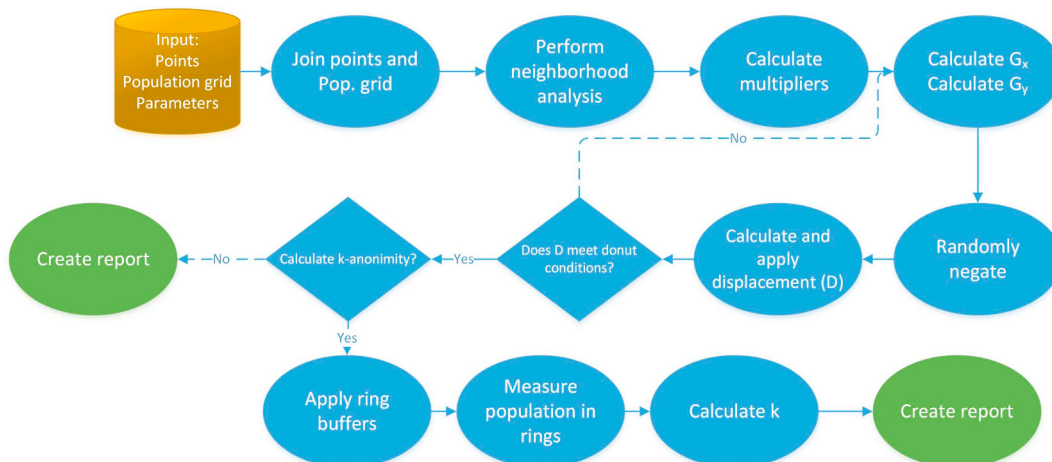


Fig. 3. An overview of the spatial anonymization process developed in this study.

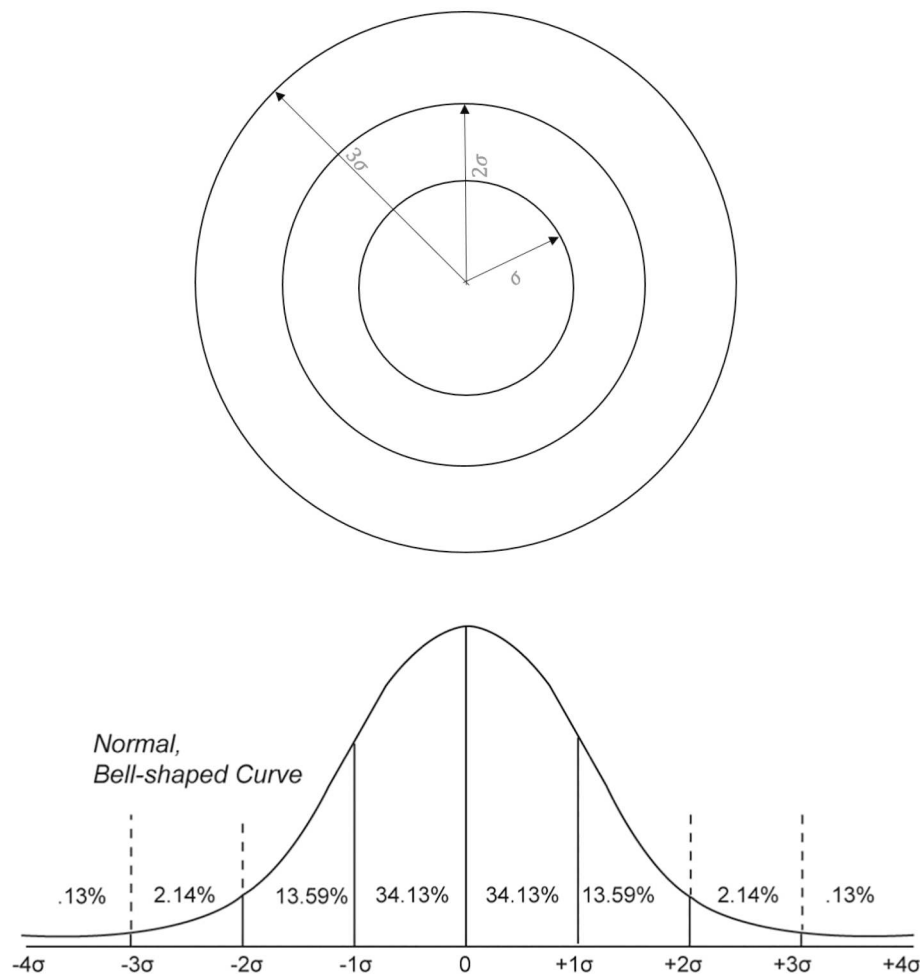


Fig. 4. Estimating k-anonymity. Using the dataset's standard deviation of displacements, σ , an estimate of achieved k-anonymity is calculated.

half females, the actual achieved k-anonymity will only be half of the measured value.

2.2.3. Anonymizing non-spatial attributes

A dual approach was taken to anonymize sensitive non-spatial attributes that can be used to single out individuals, or compromise achieved k-anonymity beyond an acceptable threshold. The very sensitive personal information that could be used to identify the individuals are removed from the data. Examples of such attributes include open comments by individuals or specific information about their job and household characteristics. As a second approach, some of the other attributes were kept in the data in an aggregated form to ensure they cannot be used to single out individuals. For example, age of individuals, which was available as a numeric value in the original data, was aggregated into large categories. The used categorization was adopted from Statistics Finland's public datasets and included three groups of individuals within following age ranges (<15 , $15-65$, and >65).

2.3. Real data implementation and evaluation of effectiveness

To evaluate its performance, the procedure described above was applied to a PPGIS dataset collected from Helsinki metropolitan area, Finland. Subsequently, we measured the statistics on how the procedure has performed in anonymizing the data. Additionally, we tested how the anonymization has affected the measurements derived from the anonymized data compared to the original dataset.

The data in this part was anonymized using the same procedure as

described earlier. However, it should be noted that as an extra safety measure, home points located in areas less densely populated than 200 per square kilometer were removed prior to the anonymization. Additionally, any point, which failed to meet a minimum K-anonymity of 300 after anonymization, was also removed.

2.3.1. Measures and variables used for evaluation

PPGIS data is widely used in urban and environmental studies to assess various levels and aspects of person-environment relationships. Therefore, in order to evaluate the usability of anonymized data for research, we used a home range model as the spatial unit of analysis and calculated a number of mobility and environmental characteristics for each individual. The home range model was adopted from an earlier study using PPGIS data (Hasanzadeh, Broberg, & Kyttä, 2017). The model is an individualized customized minimum convex polygon containing the home location of everyday destinations of the individuals. In this study, to avoid overly large polygons, destinations further than 10 km from an individual's place of residence were excluded from the modeling.

Using this model as the spatial unit, a number of variables were calculated for each individual in both original and anonymized datasets for comparison. These are some common variables that are adopted from previous research (Table 2).

In addition to the calculated variables described above, a number of other variables were directly taken from the survey. This includes four perceived wellbeing measures, namely health, quality of life (QoL), capability of functionality, and happiness, which were directly asked in the survey using a five-point Likert scale that ranged from very bad to

Table 2
List of variables used for comparison.

Variable	Measurement method	Example of previous use
Area	Area of home range	(Hasanzadeh, Laatikainen, & Kyttä, 2018)
Elongation	Length to width ratio of the smallest rectangle enclosing the home range	(Ramezani, Laatikainen, Hasanzadeh, & Kyttä, 2019)
Greenness	The percentage of home range area covered by open green spaces e.g. forests and parks. (using Corine land cover data)	(Broberg, Salminen, & Kyttä, 2013)
Distance to destinations	Average distance from home to all everyday destinations for each individual.	(Perchoux et al., 2014)
Population density	Average population density within home range boundary (using Statistics Finland population grid data 2017)	(Hasanzadeh, 2019)
Orientation	The orientation of the longer side of the smallest rectangle enclosing the home range. Orientation angles are measured in decimal degrees clockwise from north.	(Sherman, Spencer, Preisser, Gesler, & Arcury, 2005)

very good. Additionally, four background variables, namely age education, gender, and income, are also included in the analysis.

It should be noted that all spatial analyses and simulations were conducted in ArcMap 10.6 and Python 2.7 environment mainly using ESRI's ArcPy module.

2.3.2. Statistical methods

Paired sample *t*-tests were utilized to examine whether significant differences exist between variables calculated using the two sets of data, i.e., the original and the anonymized data. The significance of comparison results is adjusted for type I error using Bonferroni correction. To examine the associations between variables within the two datasets, a Pearson correlation analysis was performed on each dataset. Subsequently, we compared the two correlation analyses to evaluate how the anonymization has affected the correlations between variables. We did this by transforming the correlation coefficients into Z scores using Fisher's transformation. Consequently, the significance of differences was tested using Z test statistic. For the Z test, the null hypothesis, H_0 , was that the correlation coefficients from each corresponding pair in the two datasets are equal ($\beta_1 = \beta_2$). H_A , the alternative hypothesis, was that $\beta_1 \neq \beta_2$. All the statistical analyses were conducted using SPSS Statistics 26 and Python 2.7.

3. Results

3.1. Data anonymization output, key figures and stats

A total of 844 home points were available in the original data. This number dropped to 824 points after anonymization as a number of points were removed to protect these individuals' privacy¹. This included points located in sparsely populated areas, as well as those which failed to meet a minimum k-anonymity requirement of 300 after data anonymization. As it can be seen in Table 3, all the points have been displaced for between 50 and 727 m from their original location and a minimum k-anonymity of 330 is achieved in the process. As illustrated in Fig. 5, the displacements have preserved the normal curve shape with a higher concentration of smaller values.

As seen in Fig. 6, population density multiplier (PDM), has affected the anonymization process, with the greatest displacements occurring in the least populated areas. A very small value of 0.05 was used for Δ , hence the feature density multiplier (FDM) has minimal effect on the anonymization process.

Table 3
Anonymization statistics. Rounded to the nearest integer. SD: Standard deviation ($n = 824$).

	Minimum	Maximum	Mean	Median	SD
Displacement (m)	50	727	137	126	69
K-anonymity	330	93,603	53,456	58,442	27,869

3.2. Effects of data anonymization on measurements

Fig. 7 illustrates examples of how home range models derived from the two datasets overlap with one another. On average, the anonymized and original home ranges overlap for around 92% of their areas. However, there is some variation between individuals with overlap percentages varying from as little as 31% up to 99%, with a standard deviation of 6%. As shown in Fig. 7, the poorest overlaps occur when large displacements are applied to small spatial units. Such poor cases were most common in less densely populated areas and among individuals who had marked few or no other points than their living location.

The *t*-test analysis indicates that none of the measurements has significantly changed after the anonymization (Table 4). Green area percentage on average has diverged from its original value for only 0.17%, which is a small value. Further, the area of home range on average has changed for 0.1 km² after anonymization. Similarly, the changes incurred to distance, elongation, and orientation were also small.

It is worth mentioning that interpretation of significances needs to be made cautiously and in a context-sensitive manner. For example, the 0.1 km² average change caused to the areas seem insignificant compared to the 11.95 km² average area of the home range units in this study (0.8% change). This change equates to a bigger proportion of a smaller spatial unit such as a 500 m in radius circular buffer (12%).

3.3. Effects of data anonymization on associations between personal, environmental, and home range variables

Table 5 shows the results of Z test on the correlation coefficients derived from the two datasets, original and anonymous, assessing how significantly they differ from each other. According to these results, the changes caused by data anonymization are insignificant on most correlation coefficients. However, the few significant differences are likely caused by the generalization of attribute data. For example, in the original dataset income was presented in 16 categories. This was aggregated into only two categories of below and above average in the anonymization process.

4. Discussion

Use of PPGIS for data collection has been growing rapidly over the past few years in different areas of research and practice. With the growing amount of data, there is little doubt that a potentially wider community can benefit from open access to them. Additionally, open data add to the transparency of research and can be considered as an essential feature of science. However, open data comes with significant legal and ethical challenges as PPGIS datasets typically contain sensitive personal information, which need to be protected prior to any publication.

PPGIS data have special characteristics that make them different

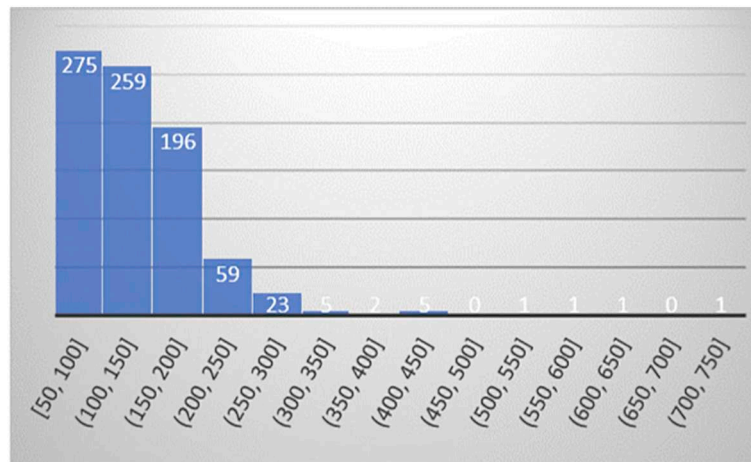


Fig. 5. Frequency of different displacement distances. Distances are in meters.

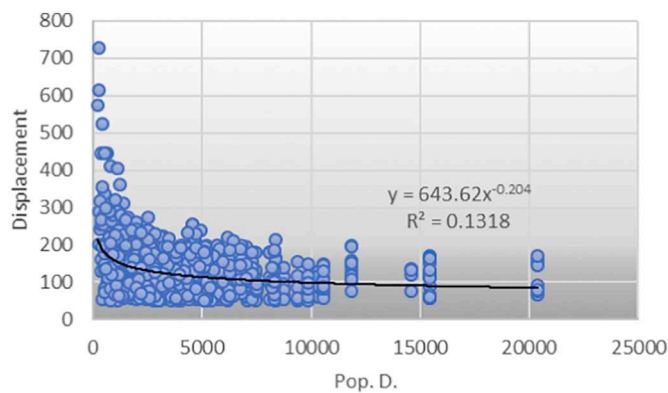


Fig. 6. Distribution of displacement distances by population density. Distances are in meters.

from many other sources of data, hence, special measures need to be taken for anonymizing them. PPGIS datasets are often collected on an individual level, which adds to the sensitivity of this data. Additionally, PPGIS data typically include both spatial and non-spatial personal information. Therefore, not a single method but a strategy comprising of various methods is required in order to protect different levels of personal information in this data. In response to these needs, in this study we developed a PPGIS data anonymization strategy comprising various anonymization methods to enable opening PPGIS data.

Using the anonymization algorithm developed in this study, the home locations in the test data were displaced for 137 m on average from their original locations. This is nearly half the average displacement distance reported by a previous study using a comparable bimodal

Gaussian approach (Cassa et al., 2006). Because of the included context-based parameters, the largest displacements generally occurred in areas where population density and participation were lower. A minimum and average K-anonymity of respectively 330 and 53,456 were accomplished for the home locations after the anonymization. This is a considerably higher K-anonymity compared to some previous studies. In a study carried out in three counties in the US, respectively 17 and 23% of cases yielded k-anonymity values of less than 20 after anonymization with random noise addition and location swapping techniques (Zhang, Freundsuh, Lenzer, & Zandbergen, 2017). In another study carried out in Portland metropolitan region in the US, 5 out of 10 studied neighborhoods yielded minimum k-anonymity values of less than 50 when a donut approach with maximum distance of around 1600 m (1 mile) was used (Clifton & Gehrke, 2013). It is worth noting that interpretation of k value as a measure of anonymity should be made with caution. The true k-anonymity actualized in an anonymization may be considerably lower than the measured value depending on what other attributes are included in the data. For example, the k-anonymity reported in this study does not take into account any attributes other than gender. Depending on the level of generalization, including information such as income and education for each individual may significantly lower the actual K. Therefore, it is recommended to aim for a greater K-anonymity in the anonymization process in order to ensure confidentiality of participants.

We compared the measurements derived from the two datasets based on a number of home range calculations adopted from earlier studies (Hasanzadeh et al., 2017). On average, the home range of individuals after anonymization showed a roughly 92% match with the one before anonymization. This indicates that anonymized and original home ranges substantially overlap, suggesting potentially insignificant

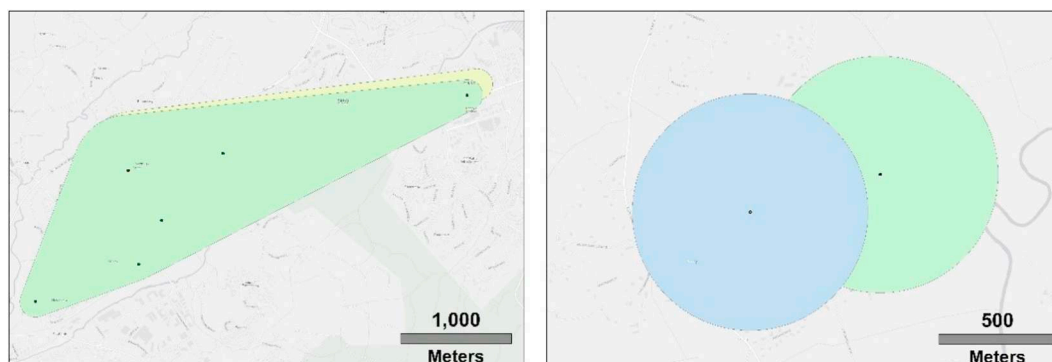


Fig. 7. Good (left) and poor (right) examples of how original and anonymized individual home ranges overlap.

Table 4

t-test results. df = 1666. Bonferroni correction did not influence the significance any of the comparisons.

	t	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Area (km ²)	0.08	0.92	0.10	1.14	− 2.15	2.35
Greenness (%)	0.32	0.74	0.17	0.54	− 0.89	1.23
Distance (m)	0.02	0.97	36.19	1352.81	− 2617.19	2689.58
Elongation	0.56	0.57	0.06	0.11	− 0.15	0.27
Orientation (degrees)	0.54	0.58	1.36	2.48	− 3.51	6.23

Table 5Z test statistics showing the significance of changes to correlations between variables in the two datasets. The results in this table are in the following format: Z (n_{original}, n_{anonymous}). Significant values are bolded in the table. * $p < .05$ ** $p < .01$. NA: not applicable

	Age	Education	Gender	Income	Pop. D.	Greenness	Area	Elongation	Orientation	Health	Functioning	QoL	Happiness
Age	1												
Education	1.41 (890,752)	1											
Gender	0.64 (1105,815)	1.04 (887,879)	1										
Income	2.34** (1065,790)	5.28** (859,847)	0.96 (1061,1061)	1									
Pop. D.	1.83* (836,824)	0.19 (775,752)	0 (835,815)	0.40 (815,790)	1								
Greenness	0.81 (836,824)	0.98 (775,752)	0.40 (835,815)	0.8 (815,790)	0.59 (844,824)	1							
Area	0.81 (836,824)	1.37 (775,752)	0.20 (835,815)	0.41 (815,790)	0 (844,824)	0 (844,824)	1						
Elongation	0.40 (836,824)	0.19 (775,752)	0 (835,815)	0.60 (815,790)	0.2 (844,824)	0.41 (844,824)	0.20 (844,824)	1					
Orientation	0.20 (836,824)	0.19 (775,752)	0.61 (835,815)	0.6 (815,790)	0.2 (844,824)	0.61 (844,824)	0 (844,824)	0.20 (844,824)	1				
Health	0.20 (896,751)	1.48 (892,864)	0.21 (894,912)	2.35** (865,866)	0.19 (781,751)	0.19 (781,751)	0.59 (781,751)	0.19 (781,751)	0.97 (781,751)	1			
Functioning	0.8 (888,746)	1.06 (884,862)	0.42 (886,883)	1.92* (857,848)	0.19 (773,746)	0.19 (773,746)	0.39 (773,746)	0.58 (773,746)	1.16 (892,742)	NA	1		
QoL	0.8 (892,742)	2.11** (888,860)	0.42 (890,873)	1.09 (861,840)	0 (777,742)	0.58 (777,742)	0.78 (777,742)	0 (777,742)	0.78 (777,742)	NA	NA	1	
Happiness	0.40 (893,751)	1.47 (889,869)	0 (891,881)	1.30 (863,847)	0.19 (779,751)	0.39 (779,751)	0.19 (779,751)	0.39 (779,751)	0.39 (779,751)	NA	NA	NA	1

environmental measurement biases after the anonymization. However, in few cases the overlap percentage was as low as 30%. These individuals have relatively small home ranges and/or have not reported any other points than their home locations in the survey. Therefore, scarcity of map responses and the size of spatial units used need to be taken into consideration for future use of the anonymized data.

None of the included measurements, namely area, greenness, distance, orientation, and elongation of home range, had statistically significantly changed from their original values. Furthermore, with the little effects of anonymization observed on these measurements, as expected, the associations were statistically unchanged in most cases. The only significant changes were caused by the generalization of non-spatial personal attributes such as income and education. Although measured by different sets of structural variables, previous research using donut approach has generally shown more environmental measurement errors caused by the anonymization process (Clifton & Gehrke, 2013). However, the errors have been previously found to be associated with the amount of displacement as well as the structural characteristics of the area (Clifton & Gehrke, 2013).

At the same time, for the interpretation of these results it is important to note that the statistical differences depend on the geographical scale and the size of spatial unit of analysis. For instance, a 0.1 km² change in the area of home ranges is insignificant compared to the large size of home ranges in this study. Obviously, for a smaller spatial unit of analysis this may turn out to be a significant deviation from the original. Therefore, when using anonymized data, it is

recommendable to opt for larger spatial units of analysis, such as activity spaces and home ranges (Hasanzadeh et al., 2017; Laatikainen, Hasanazadeh, & Kytä, 2018), rather than smaller, often home based, units of analysis such as circular buffers (Kytä, Broberg, Haybatollahi, & Schmidt-Thome, 2015; Seliske, Pickett, Boyce, & Janssen, 2009). This is comparable to a previous finding suggesting that a larger search radius or bandwidth needs to be used when applying kernel density estimation on geomasked data (Shi et al., 2009).

Overall, the results from this study show that the PPGIS data can be safely anonymized while maintaining its overall quality for many purposes. However, the data anonymization inevitably comes with some level of data accuracy and quality degradation. Therefore, for reuse of anonymized data, it is important to first, know the quality losses, second, learn how they can affect the study results, and third, take necessary analytical measures to mediate these impacts. Further, it should be noted that this study has discussed the anonymization of usual PPGIS data with certain personal information. However, anonymization of PPGIS with a clear focus on highly sensitive personal information, such as medical history or other health records, most likely requires additional steps in order to ensure the privacy of the respondents.

This study was conducted in EU and thus European data protection regulations have been used as the legal guidelines. However, comparable regulations exist in other parts of the world and thus, adaptation of methods employed in this study may be relevant to other countries as well. Further, although the methods discussed in this paper are directly targeted to PPGIS, many of them can be applicable to other sources of spatial data as well.

5. Conclusions

This study developed a PPGIS data anonymization approach that can be used to protect the sensitive spatial and non-spatial information in data. In its core, this approach offered a context sensitive spatial anonymization method for protecting primary spatial information such as home locations of participants.

The application of this anonymization method on real data obtained from a PPGIS survey showed that using this method, PPGIS data could be safely anonymized while maintaining its overall quality. Although, some quality loss is the price we must pay for privacy protection, the data seems to be still usable for many purposes. Particularly, when a lot of individual level spatial data is available and larger spatial units such as activity spaces are used, the anonymization is unlikely to cause significant biases to the results. Despite these forward steps in PPGIS

data anonymization, future research can potentially benefit from other optimized anonymization approaches. Further, additional evidence on effects of anonymization on data quality is needed to further examine the reusability of anonymized data and promote use of open PPGIS data.

Acknowledgment

This research is primarily funded by Aalto University as part of internal seed funding aimed at Digitalization. This research is also partially funded by Finnish academy as part of PLANhealth Project (13297753). We would like to thank Ilari Lähteenmäki, Maria Söderholm, Maria Reh binder, and Tiina Laatikainen for their valuable contribution to our work.

Appendix 1: Pseudo-code of the algorithm

```
# This program displaces sensitive individual locations using a
# customized Gaussian function
Import required modules and libraries
Import input files and specify outputs
Define headers for writing excel report file
Function averagePopulation (A=Input points, B=population grid)
{
  SJ=Spatial join (A, B)
  Iterate SJ and calculate AVG_POP
  Return AVG_POP
}
Function GaussianDisplacor (Sigma)
{
  G1=Calculate G value
  G2=Randomly negate G
  CM= Calculate combined multipliers
  Return CM*G
}
## Here actual anonymization begins
Copy Original point data to a new file for processing
Iterate through each row of the copy file and update coordinates
{
  Sigma= Generate random Sigma within defined range
  New_X_coordinate= OLD_X+ GaussianDisplacor(Sigma)
  Sigma= Generate random Sigma within defined range
  New_Y_coordinate= OLD_Y+ GaussianDisplacor(Sigma)
  D= Calculate distance between original and anonymized point
  If D within donut conditions:
    Proceed to next row
  Else:
    Redo Gaussian displacement for the row
}
Write report to Excel file
If Calculate_K_Anonimity= FALSE
All files saved
Code ends here
ELSE:
{
  Apply ring buffer
  Calculate population in each ring [Spatial join]
  Calculate K according to formula
  Write to Excel report
  Save all files
  Code ends here
}
```

Appendix 2: List of attributes in data (before anonymization)

Attribute name	Responses
Gender	Categorical: Male, Female
Birthyear	Numeric
Income	Categorical: 16 categories
Retired	Categorical: Yes, No
Personal goals (19 items)	Likert scale: 0–6
Frequency of visit to destinations	Categorical: Daily, several times a week, once a week, a few times a month, less often
Usual mode for visiting destinations	Categorical: Foot, bike, car, public transport, other (entered as text)
Perceived Health – 4 items: overall health, ability to function, quality of life, state of happiness	Categorical: very bad, bad, average, good, really good
Do you have pets?	Categorical: Yes, No
Do you do sports regularly?	Categorical: Yes, No
Do you have other hobbies?	Categorical: Yes, No
Do you have grandchildren whom you meet regularly?	Categorical: Yes, No
Education	Categorical: 6 categories
Marital status	Categorical: single, married, divorced/separated, widow, registered partnership
Family type	Categorical: single with kids, single without kids, couple with kids, couple without kids
Residential type	Categorical: apartment, semi-detached house, detached house
Ownership	Categorical: 8 categories
Other open comments	Text

References

- Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., ... Miller, W. C. (2010). Geomasking sensitive health data and privacy protection: An evaluation using an E911 database. *Geocarto International*, 25(6), 443–452. <https://doi.org/10.1080/10106049.2010.496496>.
- Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of u.s. scientists. *Journal of Empirical Research on Human Research Ethics: JERHRE*, 2(4), 3–14. <https://doi.org/10.1525/jer.2007.2.4.3>.
- Ardagna, C. A., Cremonini, M., Damiani, E., Di Vimercati, S. D. C., & Samarati, P. (2007). Location privacy protection through obfuscation-based techniques. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, LNCS, 4602, 47–60. https://doi.org/10.1007/978-3-540-73538-0_4.
- Article 29 dData pProtection wWorking pParty (2014). *Anonymisation techniques. Working Party Opinions*, April. 1–37. <https://doi.org/10.1016/j.jpeds.2005.10.042>.
- Broberg, A., Salminen, S., & Kytä, M. (2013). Physical environmental characteristics promoting independent and active transport to children's meaningful places. *Applied Geography*, 38, 43–52. <https://doi.org/10.1016/j.apgeog.2012.11.014>.
- Brown, G., & Kytä, M. (2014). Key issues and research priorities for public participation GIS (PPGIS): A synthesis based on empirical research. *Applied Geography*, 46, 126–136. <https://doi.org/10.1016/j.apgeog.2013.11.004>.
- Brown, Gregory, & Fagerholm, Nora (Jun 2015). Empirical PPGIS/PGIS mapping of ecosystem services: A review and evaluation. *Ecosystem Services*, 13, 119–133. <https://doi.org/10.1016/j.ecoser.2014.10.007>.
- Brown, G., & Kytä, M. (2018). Key issues and priorities in participatory mapping: Toward integration or increased specialization? *Applied Geography*, 95, 1–8. <https://doi.org/10.1016/j.apgeog.2018.04.002>.
- Cassa, C. A., Grannis, S. J., Overhage, J. M., & Mandl, K. D. (2006). A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *Journal of the American Medical Informatics Association*, 13(2), 160–165. <https://doi.org/10.1197/jamia.M1920>.
- Clifton, K. J., & Gehrke, S. R. (2013). Application of geographic perturbation methods to residential locations in the Oregon household activity survey. *Transportation Research Record: Journal of the Transportation Research Board*, 2354(1), 40–50. <https://doi.org/10.3141/2354-05>.
- Duckham, M., & Kulik, L. (2005). A formal model of obfuscation and negotiation for location privacy. *Lecture Notes in Computer Science*, 3468, 152–170. https://doi.org/10.1007/11428572_10.
- Duncan, D. T., Castro, M. C., Blossom, J. C., Bennett, G. G., & Steven, L. G. G. (2011). Evaluation of the positional difference between two common geocoding methods. *Geospatial Health*. <https://doi.org/10.4081/gh.2011.179>.
- European Parliament and Council (1995). *Directive 95/46/EC of the European Parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*.
- European Parliament and Council (2016). *Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, (and repealing Directive 95/46/EC)*.
- Ghinita, G., Zhao, K., Papadias, D., & Kalnis, P. (2010). A reciprocal framework for spatial K-anonymity. *Information Systems*, 35(3), 299–314. <https://doi.org/10.1016/j.is.2009.10.001>.
- Hasanzadeh, K. (2019). Exploring centrality of activity spaces: From measurement to the identification of personal and environmental factors. *Travel Behaviour and Society*, 14, 57–65. <https://doi.org/10.1016/J.TBS.2018.10.001>.
- Hasanzadeh, K., Broberg, A., & Kytä, M. (2017). Where is my neighborhood? A dynamic individual-based definition of home ranges and implementation of multiple evaluation criteria. *Applied Geography*, 84. <https://doi.org/10.1016/j.apgeog.2017.04.006>.
- Hasanzadeh, K., Laatikainen, T. E., & Kytä, M. (2018). A place-based model of local activity spaces: Individual place exposure and characteristics. *Journal of Geographical Systems*, 20(3), 227. <https://doi.org/10.1007/s10109-017-0264-z>.
- Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences (Vol. 18, Issue 5, pp. 235–241)* Elsevier Ltd <https://doi.org/10.1016/j.tics.2014.02.010>.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>.
- Kwan, M. P., Casas, I., & Schmitz, B. C. (2004). Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica*, 39(2), 15–28. <https://doi.org/10.3138/X204-4223-57MK-8273>.
- Kytä, M., Broberg, A., Haybatollahi, M., & Schmidt-Thome, K. (2015). Urban happiness: Context-sensitive study of the social sustainability of urban settings. *Environment and Planning B: Planning and Design*, 43(1), 34–57. <https://doi.org/10.1177/0265813515600121>.
- Laatikainen, T. E., Hasanzadeh, K., & Kytä, M. (2018). Capturing exposure in environmental health research: Challenges and opportunities of different activity space models. *International Journal of Health Geographics*, 17(1), 29. <https://doi.org/10.1186/s12942-018-0149-5>.
- Leitner, M., & Curtis, A. (2004). Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives*, (49), 22–39. <https://doi.org/10.14714/cp49.439>.
- Leitner, M., & Curtis, A. (2006). A first step towards a framework for presenting the location of confidential point data on maps—Results of an empirical perceptual study. *International Journal of Geographical Information Science*, 20(7), 813–822. <https://doi.org/10.1080/13658810600711261>.
- McNutt, M. (2014). Reproducibility. *Science (Vol. 343, Issue 6168, p. 229)* American Association for the Advancement of Science <https://doi.org/10.1126/science.1250475>.
- Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... Van der Laan, M. (2014). Promoting transparency in social science research. *Science (Vol. 343, Issue 6166, pp. 30–31)* American Association for the Advancement of Science <https://doi.org/10.1126/science.1245317>.
- Mukherjee, F. (2015). Public participatory GIS. *Geography Compass*, 9(7), 384–394. <https://doi.org/10.1111/gec3.12223>.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science (Vol. 348, Issue 6242, pp. 1422–1425)* American Association for the Advancement of Science <https://doi.org/10.1126/science.aab2374>.
- O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The Chrysalis effect. *Journal of Management*, 43(2), 376–399. <https://doi.org/10.1177/0149206314527133>.
- Perchoux, C., Kestens, Y., Thomas, F., Van Hult, A., Thierry, B., & Chaix, B. (2014). Assessing patterns of spatial behavior in health studies: Their socio-demographic determinants and associations with transportation modes (the RECORD Short Study). *Social Science and Medicine*, 119, 64–73. <https://doi.org/10.1016/j.socscimed.2014.07.026>.
- Ramezani, S., Laatikainen, T., Hasanzadeh, K., & Kytä, M. (2019). Shopping trip mode choice of older adults: An application of activity space and hybrid choice models in understanding the effects of built environment and personal goals. *Transportation*,

- 1–32. <https://doi.org/10.1007/s11116-019-10065-z>.
- Saygin, Y., Hakkani-Tür, D., & Tür, G. (2005). Sanitization and anonymization of document repositories. *Web and information security* <https://doi.org/10.4018/978-1-59140-588-7.ch007>.
- Seliske, L. M., Pickett, W., Boyce, W. F., & Janssen, I. (2009). Association between the food retail environment surrounding schools and overweight in Canadian youth. *Public Health Nutrition*, 12(9), 1384. <https://doi.org/10.1017/S1368980008004084>.
- Sherman, J. E., Spencer, J., Preisser, J. S., Gesler, W. M., & Arcury, T. A. (2005). A suite of methods for representing activity space in a healthcare accessibility study. *International Journal of Health Geographics*, 4(1), 24. <https://doi.org/10.1186/1476-072X-4-24>.
- Shi, X., Alford-Teaster, J., & Onega, T. (2009). Kernel density estimation with geographically masked points. 2009 17th International Conference on Geoinformatics. *Geoinformatics*, 2009. <https://doi.org/10.1109/GEOINFORMATICS.2009.5292881>.
- Sieber, R. (2006). Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers*, 96(3), 491–507. <https://doi.org/10.1111/j.1467-8306.2006.00702.x>.
- Verplanke, J., McCall, M. K., Uberhuaga, C., Rambaldi, G., & Haklay, M. (2016). A shared perspective for PGIS and VGI. *The Cartographic Journal*, 7041(November), 1–10. <https://doi.org/10.1080/00087041.2016.1227552>.
- Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass*, 3(2), 647–680. <https://doi.org/10.1111/j.1749-8198.2008.00205.x>.
- Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine*, 2014, 1–14. <https://doi.org/10.1155/2014/567049>.
- Zhang, S., Freundsuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, 44(1), 22–34. <https://doi.org/10.1080/15230406.2015.1095655>.
- Zhou, B., Pei, J., & Luk, W. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter*. <https://doi.org/10.1145/1540276.1540279>.