Marshall Plan Scholarship

Obfuscation of Spatial Survey Data in Carinthia for Privacy Preservation

Dara Seidl
San Diego State University – UC Santa Barbara
Carinthia University of Applied Sciences

I. Introduction

A common goal in the data collection process is to share results with other researchers and the public. Access to a shared data source allows for results to be replicable and improved upon. Data sharing and integration with other sources can lead to improved knowledge production and amplify the impact of the research. However, sharing data collected under the promise of confidentiality can be restrictive. This is particularly true of spatial data, as location is a strong personal identifier. In 2012, researchers at the Carinthia University of Applied Sciences built an Energy Web GIS Portal for the collection and reporting of energy consumption data in the Hermagor District of Carinthia, Austria (Paulus et al. 2014). The portal offers a standardized questionnaire, spatial visualization, and download and reporting tools. In keeping with data confidentiality, the energy demand maps for private households offered by the portal display data aggregated to grid-like statistical units with data suppressed where population number is not high enough to protect anonymity. As an alternative to the suppression and aggregation of data, obfuscation approaches involve the slight alteration of point data to protect both spatial patterns and personal identities. Geomasking obfuscation techniques offer the potential to both preserve privacy and maintain underlying spatial distributions better than large-scale aggregation. This study evaluates the effectiveness of the recognized obfuscation techniques of grid masking, random perturbation, and weighted random perturbation in maintaining distributional integrity in the Hermagor energy data, and also evaluates the performance of a new masking procedure, Voronoi masking.

An important question for the utility of masking is whether the masked data are fit for decision support. This is a more complex test for the fitness of the data compared to simply the degree to which the masked data fit the general spatial pattern of the original data points. In order

to provide value to researchers and decision-makers, the masked data must respond to research questions with similar accuracy to the original data. This study statistically tests the clustering of household energy consumption and evaluates the performance of the obfuscated data compared to the original unmasked data. The greater significance of this research is that it is a first step in determining if masked data can serve to replace original data in decision support systems.

II. Background

Over the past ten years, there has been a surge in interest in the concept of locational privacy, or geoprivacy, among the geographic community (Zandbergen 2014). Geoprivacy is generally understood as the individual right to determine how, if, and when one's personal location information is shared with other parties (Kwan et al. 2004; AbdelMalik et al. 2008; Elwood and Leszczynski 2011; Kar et al. 2013). The links between location and other available data, while generating sensitivity to breaches of privacy and confidentiality, have great utility for emergency response, navigation, disaster relief, health, and social research purposes (Duckham and Kulik 2007). Location data sharing is also crucial to many popular location-based services, some of which could not exist without exact coordinate locations (Vicente et al. 2011). Furthermore, limits on access to fine-level data are found to impede accurate spatial analysis for proper health research, according to a survey of health professionals (AbdelMalik et al. 2008). Given the advantages of precise spatial data, the ability of individuals to control their spatial footprints can be put at risk. Kounadi and Leitner (2014) write that the disclosure of location information as a breach of geoprivacy can come from new geospatial technologies, laws that do not stringently protect privacy, and negligence by authors and publishers.

Legal Geoprivacy Protections

The right to geoprivacy is closely tied to concepts of information privacy and data protection. Legal privacy expert Sjaak Nouwt (2008) asserts that the concept of a reasonable expectation of geoprivacy exists within the European legal framework and has been recognized by the European Court of Human Rights. Having a "reasonable expectation" of geoprivacy means that in realms where individuals can reasonably expect privacy with regard to their location information, their locations cannot lawfully be disclosed. Information privacy in Austria is safeguarded under the Data Protection Act of 2000, or *Datenschutzgesetz 2000*, which has stipulations restricting further use of data collected in surveys and by other means, such as sensor networks. The act protects personal privacy with regard to data collected from research subjects and video surveillance. For example, video surveillance implementations must be documented with the Data Protection Administration, and data must be deleted after 72 hours.

In survey research, a permit can be obtained from the Data Protection Authority to use personally identifiable data for scientific analyses. If a permit is granted, data must be recoded without delay so that data subjects are unidentifiable as soon as is acceptable for analysis. Improper use of the data is punishable by an administrative penalty of up to €25,000 or 1 year in prison if the intention of data mishandling and personal identification is determined to be for harm. Despite protective laws and rulings, however, a 2010 report by the EU Agency for Fundamental Rights finds that data protection authorities are not well-equipped to investigate or intervene in privacy violations (European Agency for Fundamental Rights 2010). Data retention is a contentious concept in Europe. The EU Data Retention Directive of 2006 ordered member states to store citizen telecommunications data for at least six months, allowing police to access IP address data and all text messages when permitted by a court. The European Court of Justice

invalidated this directive in 2014 as interfering with the rights to privacy and data protection of European citizens. In this ruling, privacy was deemed more essential than the benefits of sweeping personal information for law enforcement.

Obfuscation for Privacy Preservation

If legal protections are not adequate for protection of confidentiality in surveys, it is up to authors and publishers to ensure that location data are protected. Obfuscation through geographical masks is being evaluated as a method to protect locational privacy when mapping sensitive data (Kounadi and Leitner 2014). Obfuscation degrades the quality of geographical data by introducing inaccuracy, increasing imprecision, or maintaining vagueness in terms used to describe location (Duckham and Kulik 2007). Aside from affine transformations, which translate, re-scale, or rotate a point pattern, grid masking and random perturbation are the most thoroughly documented masking techniques (Kwan et al. 2004; Leitner and Curtis 2004; Kounadi and Leitner 2014). Random perturbation displaces points a random distance and direction within a specified distance threshold. Grid masking, a variant of point aggregation, obfuscates points by snapping them to the centroid or vertex of the nearest grid cell of set size. Donut masking, which is similar to random perturbation in randomizing distance and direction, limits the area of displacement by enforcing a minimum distance. Donut masking has primarily been evaluated in masking sensitive health data (Hampton et al. 2010; Allshouse et al. 2010). These masking techniques are generally preferred over aggregation to administrative units for preserving the balance between spatial pattern and privacy. It is argued that aggregation can adversely impact cluster detection and lead to inaccurate or misleading results (Kwan et al. 2004; Kounadi and Leitner 2014).

The Energy Web GIS Portal built by researchers at the Carinthia University of Applied Sciences in 2012 releases spatial data regarding household energy consumption in the Hermagor District aggregated to statistical units (Paulus et al. 2014). The statistical units are grid cells of 125 meters, and data are repressed in cells where there are not enough households to maintain confidentiality. This research tests whether masking techniques can preserve spatial patterns that would be important for energy analysis and be as protective of privacy as aggregation. A first objective is the exploration of how well obfuscation techniques preserve the integrity of original point data and preserve household anonymity in Hermagor. A second research question is occupied with how masked point data may be used to answer realistic research questions regarding energy consumption. These results are an important first step in testing the fitness of masked data for decision-making.

III. Methods

This section describes the methods implemented to test the changes in underlying spatial structure during obfuscation and the preservation of household anonymity. Figure 1 depicts the overview of the analysis starting from the original point address data down to the masked point data and resulting statistical comparisons.

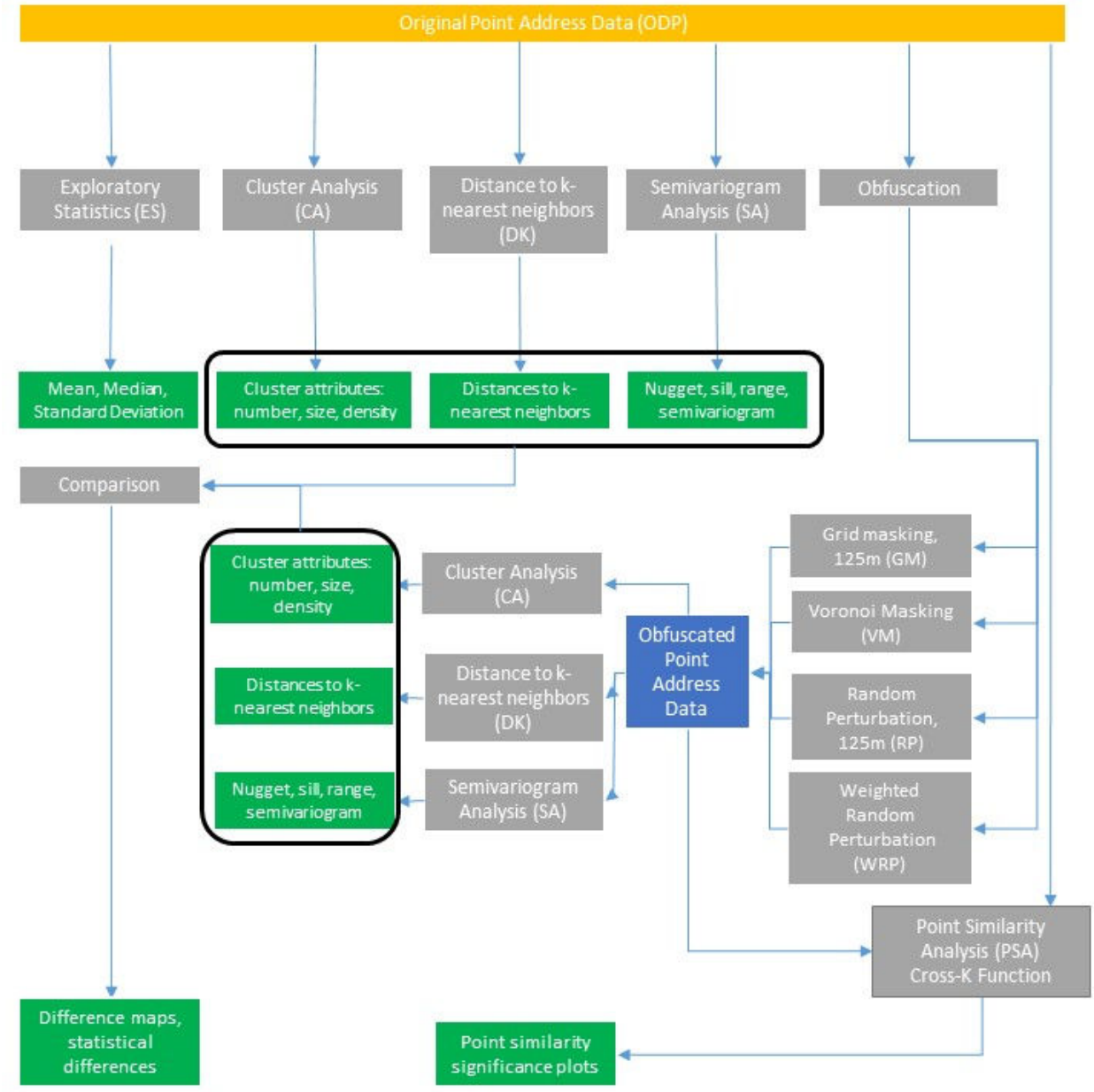


Figure 1. Overview of obfuscation analysis

Study Area

This study employs energy use data collected for every household in the Hermagor district of Carinthia in southern Austria. The district of Hermagor is situated in southwestern Carinthia, and

the data are centered in the Hermagor borough. Compared to study regions in other masking research (Kwan et al. 2004; Leitner and Curtis 2004), Hermagor has a very low population density. The total population in the district as of the first quarter of 2014 is 18,547, and the population density is 22.95 persons per square kilometer (STATCube - Statistical Database of Statistics Austria 2014). The lower population density in this region makes individual residences more vulnerable to identification. The data set for this study includes 1,945 residential records in the District of Hermagor. Data points are situated on the centroids of the physical buildings with the number of primary residents at each building included as a variable.

The mean warm water energy consumption for each household in the data set is 2.71 megawatt hours per annum with a standard deviation of 2.16 MWh/a. Per capita, the mean warm water energy consumption is 0.97 MWh/a with a standard deviation of 0.69 MWh/a. The highest per capita warm water energy consumption is in the central part of the district, as well as towards the northeast of the region. Figure 2 displays the kernel density estimation (KDE) of the original data points for warm water energy consumption with cell size of 250 meters. The southern portion of the district is primarily uninhabited in the Schwarzwald mountain region, which explains the absence of household data towards the south.

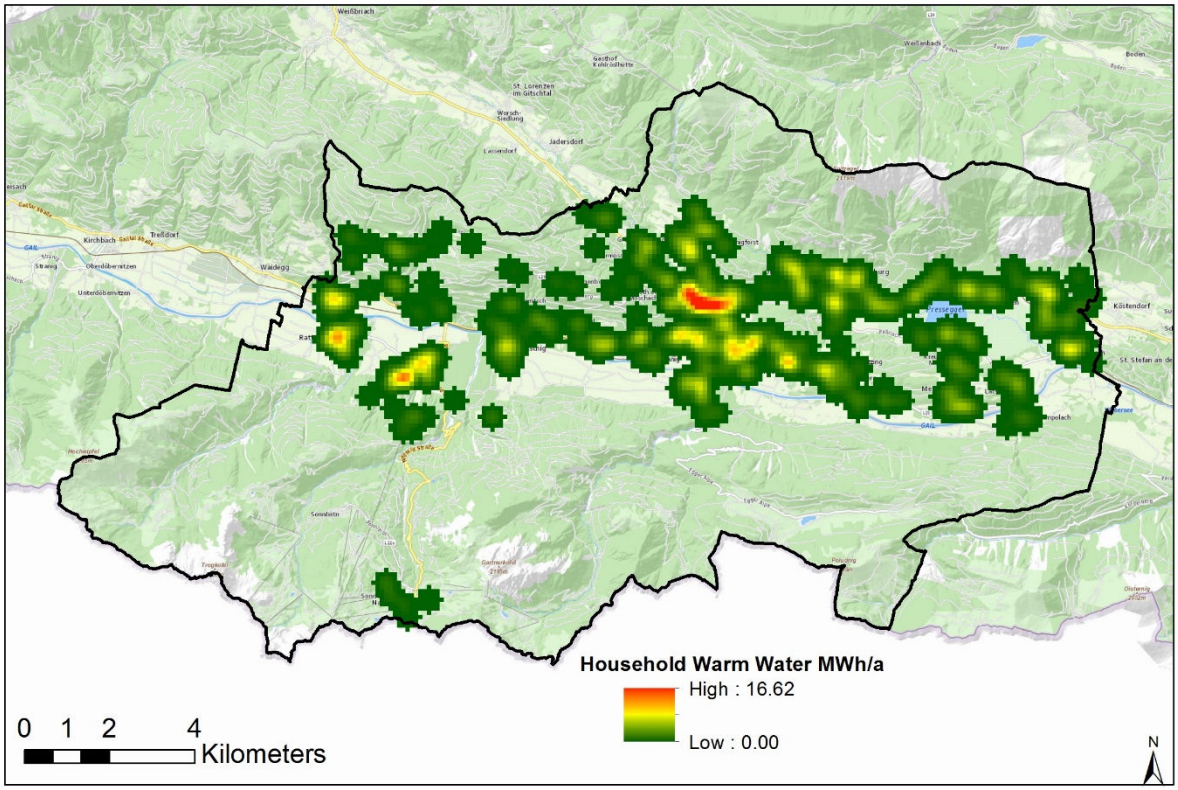


Figure 2. Kernel density estimation of per capita warm water energy consumption, Hermagor

Spatial Analysis of Original Data

A principal question for the utility of masked data is how it can be applied in decision-making. Can masked point data be used to accurately answer research questions in place of original point data? This study focuses first on the original data (ODP) with methods typically used by an energy analyst in a decision-making process. This study relies on testing the original data first for spatial patterns and subsequently applying identical methods to the masked data sets for comparison. The test research question applied for this study is to determine the locations of clusters of high and low energy consumption by household.

Some obfuscation studies have focused solely on preservation of spatial pattern as a test of maintaining the integrity of spatial data. For example, Shi et al. (2009) generate kernel density

surfaces of original and masked point data and then test for Pearson's correlations between the rasters. The current study departs from these spatial pattern correlation tests by utilizing more robust measures of difference. Underlying point distributions are evaluated using the metrics of clustering, nearest neighbor distance, and spatial autocorrelation.

As a first step, exploratory statistics (ES) are run on the original data points (ODP), including the mean and median centers with standard deviations. A nearest neighbor hierarchical cluster analysis (CA) is performed to determine the number of first order clusters present in the original data, the density of the clusters, and their standard deviational ellipses. The distance to the *k* nearest neighbors (DK) for each point is also calculated. This provides a step towards spatial weighting for masking, where the distance threshold is varied according to how vulnerable to identification each point is based on its neighbors. The concept of *k*-anonymity in privacy research refers to ensuring that each individual cannot be distinguished from *k-1* other individuals in the data (Sweeney 2002). Hampton et al. (2010) prepare spatial weights in their donut masking study to vary the distance thresholds by the radii required to reach *k* neighbors around each original point. Semivariogram analysis (SA) tests for spatial autocorrelation in the original data set. If the energy consumption data are spatially autocorrelated, a similar correlation should ideally be present in the masked patterns. These techniques provide the baseline for determining clusters of high warm water energy consumption and for quantifying the underlying spatial distributions of the original data.

Geomasking Techniques

The next step in the analysis is to obfuscate the original residence data points. Statistics Austria releases spatial data aggregated to statistical units, so called regional statistical units. These units represent regularly spaced vector grid cells of 100, 125, or 250 meters side length,

omitting information where there are too few households falling in a unit. In this case study, the statistical units are sized at 125 meters. Aggregations such as these reduce the spatial resolution of data and thereby reduce the ability of researchers to detect underlying patterns, such as disease risk (Hampton et al. 2010; Kwan et al. 2004). Zandbergen (2014) echoes that spatial analysis techniques, including cluster detection and point pattern analysis, become less accurate with aggregated data. Similarly, Luo et al. (2010) note that the smoothing effects of aggregation adversely impact the estimations of statistical models compared to data of finer resolutions. The modifiable areal unit problem (MAUP), which states that patterns witnessed on an aggregated level do not match those found at finer resolutions, makes it difficult to analyze phenomena that move across boundaries. Aggregation to grid cells is preferable to aggregation to larger geographies such as tracts or zip codes for the preservation of spatial pattern. However, the size of the grid cells used for aggregation can cause points to be moved or smoothed across a greater distance than is actually necessary to preserve privacy. Geomasking offers a solution that can allow researchers to maintain a fine point resolution for analysis while also moving the points the bare minimum distance necessary to maintain confidentiality.

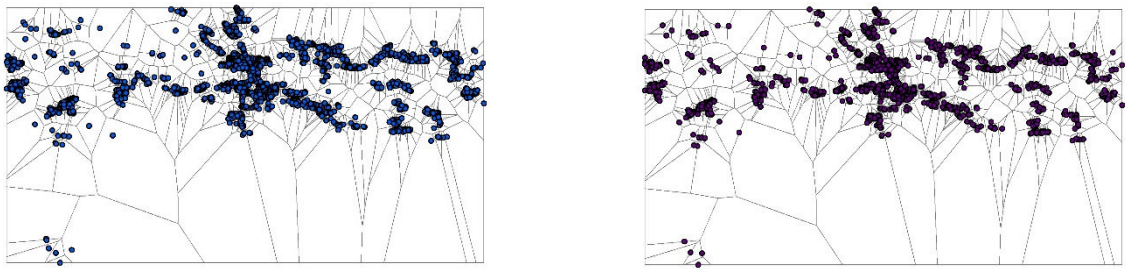
The masking techniques employed in this study include:

- Grid masking
- Voronoi masking
- Random perturbation
- Weighted random perturbation.

Grid masking (GM) involves snapping each original data point to the centroid of grid cells of a given size (Curtis et al. 2011; Leitner and Curtis 2004; Krumm 2007). This method is most similar to the aggregation technique the Austrian government uses in summarizing points into

grid-like statistical units. If the statistical units used for the energy survey results (125 meters by 125 meters) were converted to points, they would closely resemble a grid-masked version of the original data points with a distance threshold of 125 meters. This study tests grid masking with a distance threshold of 125 meters as representative of the currently implemented official statistical unit aggregation. Since all of the analyses in this study are based on point data, this is necessary for uniformity.

This study introduces a new form of obfuscation referred to here as Voronoi masking (VM). Voronoi polygons, or Thiessen polygons, define areas where the boundaries are equidistant between the surrounding points, or where inside the polygons is closer to the corresponding point than to any other point. In Voronoi masking, each point is snapped to the edges of the Voronoi polygons, making them equidistant between original point data. An advantage of this technique is that where the density of original points is higher, the points are moved a shorter distance on average, resulting in patterns that more closely resemble the original data set. Another advantage of Voronoi masking is that some points in adjacent polygons will be snapped to the same location, which can increase the k-anonymity for those points. Finally, if the data set incorporates all residences in a study area, no relocated point is placed on an actual residence. None of the relocated points remain in their original locations, or at the centroids of other residences. This permits no false identification of household points. In areas of sparse data points, it is expected that some points will be moved large distances with this method, which could disrupt patterns. However, if there are at least two households close to each other in a remote region, the points will potentially be moved a shorter distance than they would be with other masking techniques that do not account for underlying settlement patterns. Sample results of Voronoi masking are shown in Figure 3.



(a) (b)

Figure 3. Original point data (a) and points obfuscated by Voronoi masking (b)

A third masking technique for this study is random perturbation (RP). Like with grid masking, this is applied uniformly with a 125-meter distance threshold. In random perturbation, each point is relocated a random distance within a distance threshold and in a random direction (Kwan et al. 2004). A fourth measure is weighted random perturbation (WRP), where the distance threshold varies according to the distribution of residences. The weighted distances are based on the radii generated in the DK process, or the distance to *k* nearest neighbors in the original point data set. The weights are thus local measures that vary from point to point.

These random perturbation methods are supported by previous research, but have yet to be tested in a rural Austrian environment. Kwan et al. (2004) implement weighted random perturbation, deriving the weight from the population density of the Census tracts where the points resided, dividing the densities into ten equal interval classes and assigning a weighting factor from 1 to 5.5 from highest to lowest density. Allshouse et al. (2010) also weight the distance thresholds in their donut masking study based on the number of households and area of the Census block group of the original points. The maximum distance threshold is determined by displacement by a specified maximum number of k households. Actual k-anonymity in the Allshouse et al. study is measured as the number of households with a shorter distance to the

original household point than the distance between the original and masked point. Allshouse et al. use a minimum k of 5 households for illustration purposes. This study finds that the average distance to the 5th nearest neighbor in the original spatial data (ODP) is 117 meters. Since grid masking (GM) and random perturbation (RP) are applied at a blanket distance of 125 meters, a k=5 households threshold is set for WRP so that the average distance of displacement approximates that of the other methods. The distance of displacement is randomized between 0 and the distance to the 5th nearest neighbor for each point.

Spatial Analysis of Masked Data

Just as with the original data points (ODP), spatial analysis techniques are implemented on the obfuscated data. These include nearest neighbor hierarchical cluster analysis (CA), the determination of distance to *k*-nearest neighbors (DK) as a test of how anonymous the masked data are, and semivariogram analysis (SA) for spatial autocorrelation. These results are compared to those of the original dataset. Difference maps of kernel density for household warm water energy consumption are included for the comparison. The metrics here help to assess differences in the measured patterns of warm water energy consumption between the ODP and masked results.

Point Similarity Analysis

The final step of the process is a statistical point similarity analysis (PSA). This step implements cross-K functions to test between the original and masked point patterns. These tests run simulations to see if the similarities between the point patterns are due to random chance or are statistically similar. Kwan et al. (2004) implement the cross-k function in their masking study, which tests whether differences observed between point patterns are significantly similar

or different compared to random simulations. The cross-k functions here are run with 99 simulations to test a confidence envelope at the 99% confidence interval.

IV. Results

This study employs four principal obfuscation techniques: grid masking (GM), Voronoi masking (VM), random perturbation (RP), and weighted random perturbation (WRP). Figure 4 displays an example subset of original data and obfuscated results from the masking techniques used in this analysis. Visually in this example, GM appears to least preserve the spatial pattern exhibited by the original data, while VM and WRP maintain the outer shape of the original data extent better than RP. GM is thus expected to least preserve the spatial integrity of the original data points.

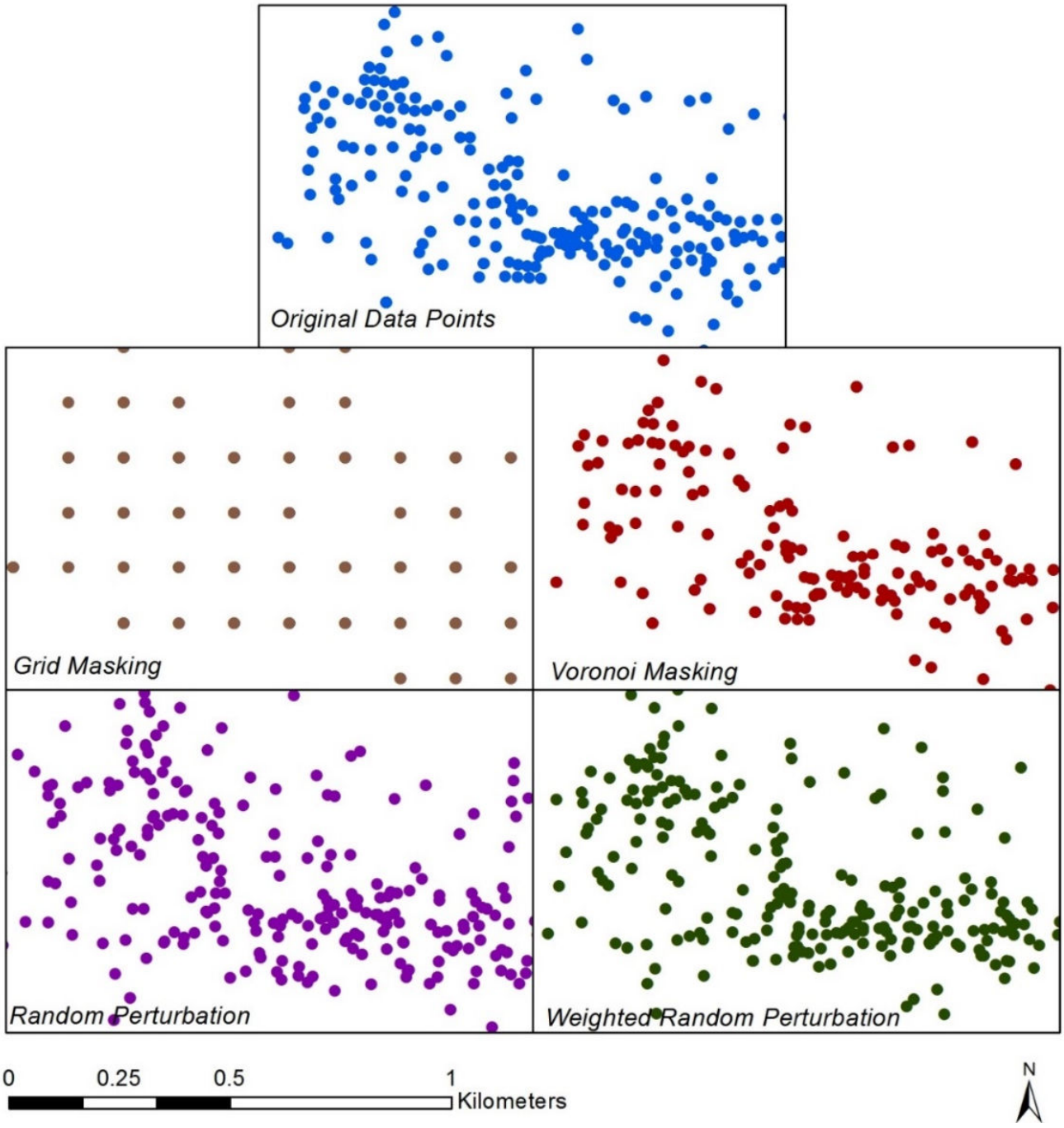
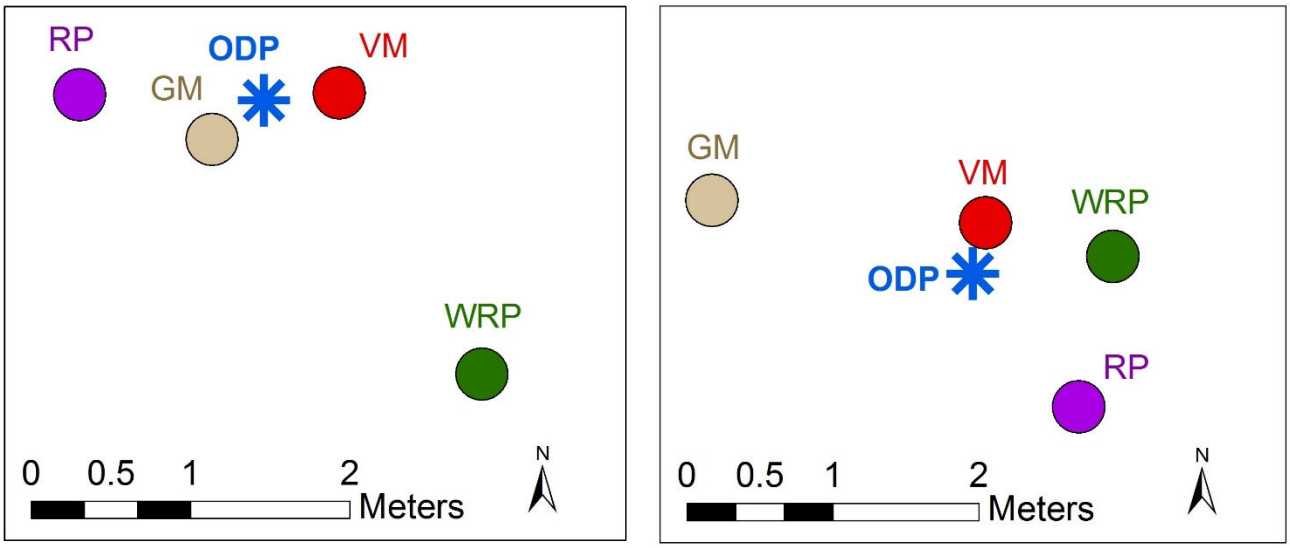


Figure 4. Masking technique examples

Exploratory Statistics (ES)

The mean center of the original data points is found at (2999.7, 164876.4)[1] along Gailtal Straße between Major Troje Weg and Friedhofweg Hermagor. This is nestled between the northern and southern settlement patterns, which run east-west along the district. The median center of the original data points is situated north of the mean center along Gailtal Straße at (3071.2, 165050.9). The mean centers, median centers, and standard deviational ellipses between the ODP and masking results are within two meters of each other and do not vary much according to technique. The mean and median center results are shown in Figure 5. The VM mean and median centers are among the closest to those of the original data points at less than one meter away. Two meters of distance between the mean and median centers is not enough of a difference to underscore any major change in underlying spatial pattern between the unmasked and masked data.



a) *b)*

Figure 5. Mean centers (a) and median centers (b) of original and obfuscated data

[1] Projected in MGI Austria GK Central

The standard deviational ellipses for each masked point pattern also exhibit low variation and on a map appear to completely overlap. The first standard ellipse statistics are shown in Table 1. The rotation of the ellipse only varies .01 degrees for random perturbation (RP), and even less for the other methods compared to the original data. The vertical and horizontal standard distances vary most for grid masking (GM) compared to the other methods, but are still within 5 meters of the unmasked standard deviational ellipse. For these general summary statistics, there are only very small differences between the unmasked and masked data sets. Further examination of the clustering patterns and spatial relationships is needed to determine the similarity of the underlying patterns.

Table 1. First standard deviational ellipse statistics by obfuscation method, distance in meters

	X Standard Distance	Y Standard Distance	Rotation
Unmasked (ODP)	1625.97	6639.41	88.09
Grid masking (GM)	1630.42	6640.55	88.09
Voronoi masking (VM)	1627.21	6639.82	88.09
Random perturbation (RP)	1625.95	6637.51	88.08
Weighted random perturbation (WRP)	1626.29	6640.17	88.09

Difference Maps

The difference maps displayed in Figure 6 were creating using kernel density estimations of the warm water energy consumption by household. A cell size of 100 meters and a search distance of 250 meters were used to smooth patterns and make them visible at the scale of the entire district of Hermagor. The absolute value of the difference between the ODP and masked result rasters for warm water energy consumption is the value depicted in the maps. All difference maps are symbolized with the same breaks. GM and RP demonstrate the highest levels of divergence from the ODP kernel estimation. RP also has the worst performance with

the few isolated points at the south of the region. The errors appear more widespread in RP, which is expected with the random displacement of all points to some random distance.

The clear best performance for this metric comes from the VM density surface. There are few visible cells for VM that reach the top category for highest difference from the ODP kernel density raster. Weighted random perturbation (WRP) also fares better than GM and RP due to shorter-distance displacements of points where there is a high point density. A problem area where all the methods resulted in great difference from the ODP is in the town of Tröplach, where the error is centered on a few sparse data points with high warm water energy consumption. There is a similar area of error in the eastern portion of the town of Hermagor, where there are no data points, but the nearby points have higher consumption records. This result may say more about errors with interpolation rather than errors resulting from obfuscation, however.

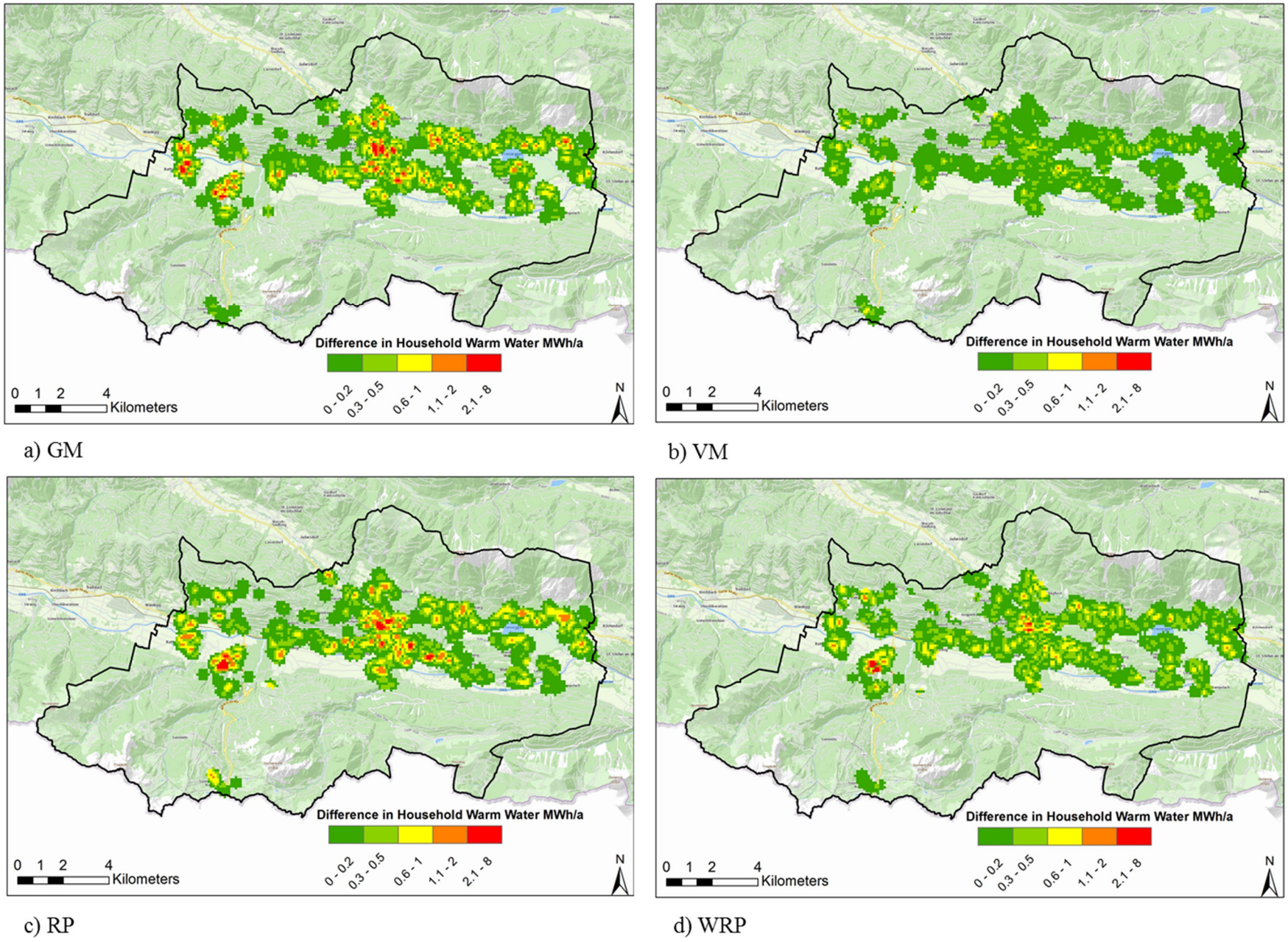


Figure 6. Difference maps for kernel density estimation of household warm water consumption

Distance to k-Nearest Neighbors (DK)

The concept of k-nearest neighbors is implemented both as a means of comparing spatial pattern across obfuscation methods and evaluating privacy. Obfuscation must balance between maintaining a similar number of neighbors around each point to maintain pattern, while simultaneously increasing the number of neighbors to enhance k-anonymity. Table 2 depicts the mean distance in meters to the 1st, 5th, 10th, and 20th nearest neighbors in the original and masked data sets. The original unmasked point data in Hermagor have a mean distance of 44 meters to the 1st, 116 meters to the 5th, 192 meters to the 10th, and 329 meters to the 20th neighbor.

Random perturbation (RP) and weighted random perturbation (WRP) increase the distance to the nearest 1, 5, 10, and 20 neighbors. This decreases k-anonymity when k-anonymity is measured within a masked dataset, rather than against true housing patterns. Grid masking (GM) and Voronoi masking (VM) exhibit lower average distances to the 1st nearest neighbors (24 meters and 17 meters, respectively) due to the tendency of these methods to snap very close points together. With RP and WRP, no masked points share identical coordinates, which explains the higher average distances to the 1st nearest neighbors. At 10 and 20 nearest neighbors, the average distances for GM and VM to the *k*th nearest neighbors are closer to the results of the ODP, although VM maintains lower average distance to all four neighbor totals tested. This is because for Voronoi masking, even the most remote points are snapped to one other point if they are not along bounding geometry for the data set. For the k-anonymity measure of privacy then, VM maintains privacy better than the other obfuscation methods and the original data pattern.

Table 2. Mean distance (meters) to k-nearest neighbors by masking method

	Neighbors			
	1	5	10	20
Unmasked (ODP)	43.8	116.8	192.0	328.8
Grid masking (GM)	24.4	116.5	201.9	335.1
Voronoi masking (VM)	16.9	109.7	188.6	326.5
Random perturbation (RP)	50.6	129.3	199.6	335.6
Weighted random perturbation (WRP)	52.0	122.9	192.6	329.6

Semivariogram Analysis (SA)

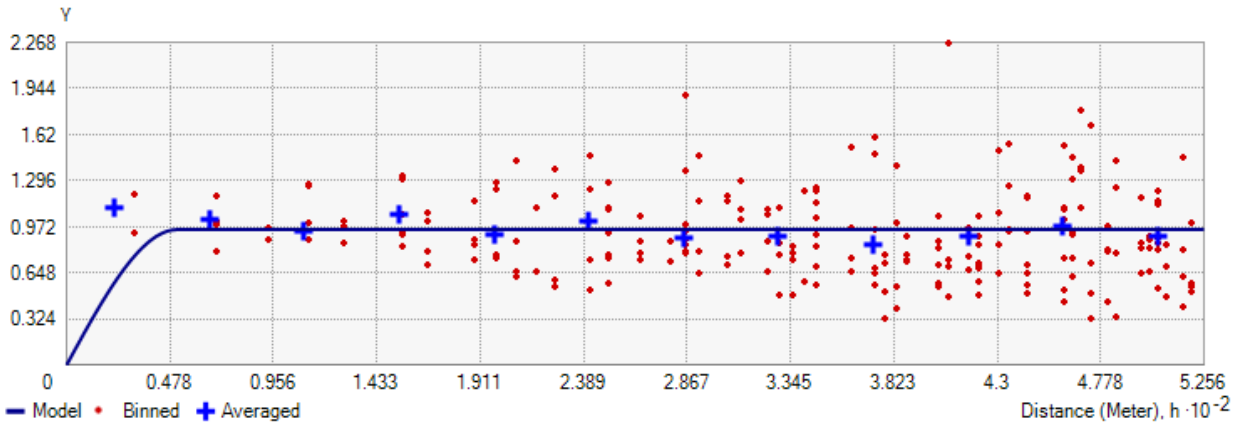
The semivariogram analysis (SA) is intended to compare the semivariogram patterns of the original data set with those of the masked data sets, including the nugget, range and sill. Semivariograms often exhibit an upward trend that levels out as the distance increases. This pattern is characteristic of spatial autocorrelation, where points that are closer together are more similar than points that are farther apart. A global test for spatial autocorrelation (Moran's I) with inverse distance in the original data points is significant with 99% confidence and a z-score of 2.67. This pattern is based on warm water energy consumption by household. A summary of the Moran's I results for warm water energy consumption in the original and masked data sets is shown in Table 3.

Table 3. Mean distance (meters) to k-nearest neighbors by masking method

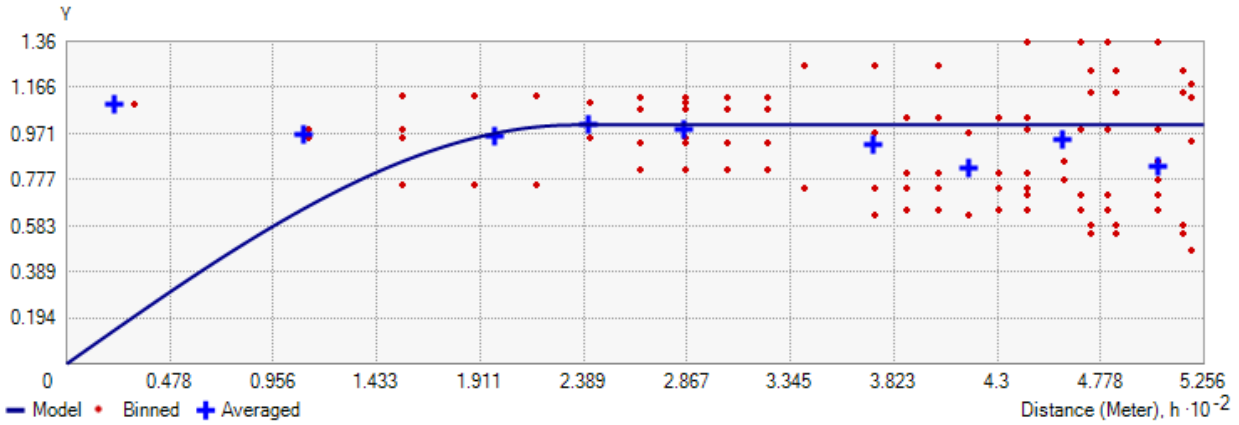
	Moran's I	z-score	p-value
Unmasked (ODP)	0.044	2.673	0.008
Grid masking (GM)	0.034	2.518	0.012
Voronoi masking (VM)	0.026	1.192	0.233
Random perturbation (RP)	0.027	4.569	0.000
Weighted random perturbation (WRP)	0.009	1.803	0.071

The warm water energy consumption point patterns for the unmasked and random perturbation (RP) reach significance for clustering at the $p < 0.01$ level. The z-scores for grid masking (GM) and weighted random perturbation (WRP) suggest that they exhibit globally clustered patterns with GM significant at the $p < 0.05$ level and WRP weakly significant at the $p < 0.10$ level. Voronoi masking (VM) is the only point pattern for which there is a decidedly random global distribution of warm water energy consumption points. One explanation for this is the snapping of somewhat distant points with varying warm water energy consumption from neighboring Voronoi polygons to be right atop each other. This snapping therefore appears to erase some spatial autocorrelation that existed before points with varying consumption levels were snapped to each other.

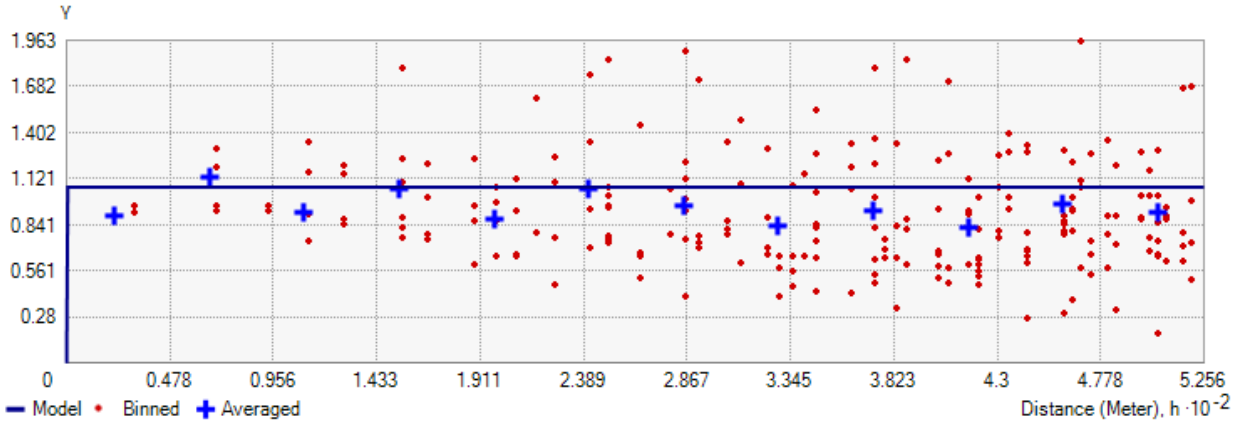
Knowing that the original data points (ODP) are globally clustered, but that not all the masked point patterns exhibit clustering, the semivariograms are expected to show some differences between the different point patterns. A subset of 400 points was selected from the original data set to examine the semivariogram. This random selection represents approximately 20% of the total data set and was chosen to examine finer patterns within the autocorrelated data. The same subset of features was then selected from the masked data sets to produce semivariograms for comparison. The original data were transformed by the normal score, and given the average nearest neighbor measurement of 43.8 meters, were given a lag size of 43.8 meters with 12 lags to match. Without modeling the nugget (giving the nugget a value of 0), the partial sills of these values are shown in Table 4. The model type of the semivariograms is spherical, and the graphs are displayed in Figure 7.


Y
2.268
1.944
1.62
1.296
0.972
0.648
0.324
0
0.478
0.956
1.433
1.911
2.389
2.867
3.345
3.823
4.3
4.778
5.256
Model
Binned
Averaged
Distance (Meter), h ·10 -2

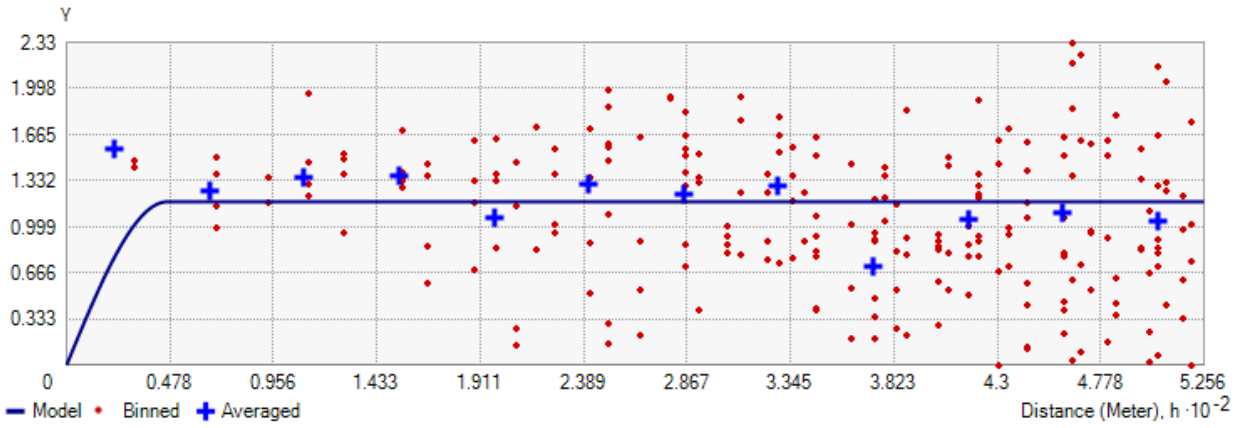
(a) ODP


Y
1.36
1.166
0.971
0.777
0.583
0.389
0.194
0
0.478
0.956
1.433
1.911
2.389
2.867
3.345
3.823
4.3
4.778
5.256
Model
Binned
Averaged
Distance (Meter), h ·10 -2

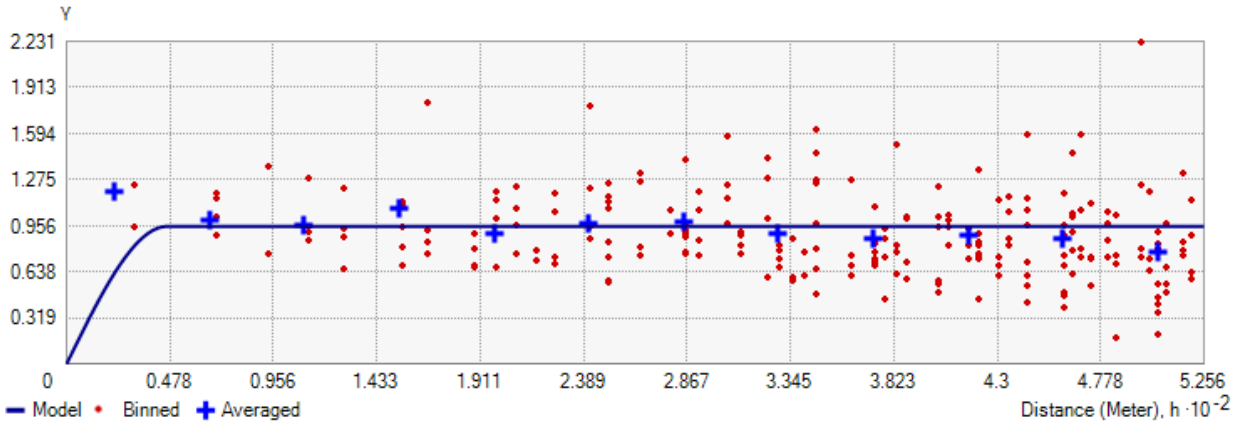
(b) GM


Y
1.963
1.682
1.402
1.121
0.841
0.561
0.28
0
0.478
0.956
1.433
1.911
2.389
2.867
3.345
3.823
4.3
4.778
5.256
Model
Binned
Averaged
Distance (Meter), h ·10 -2

(c) VM



(d) RP



(e) WRP

Figure 7. Semivariograms with lag size of 43.8 meters and 12 bins

Table 4. Partial sill values for semivariograms

	Partial sill
Unmasked (ODP)	0.952
Grid masking (GM)	1.008
Voronoi masking (VM)	1.069
Random perturbation (RP)	1.179
Weighted random perturbation (WRP)	0.952

The averaged values for all the semivariograms appear to approximate horizontal lines. This suggests that there is little spatial correlation in the data and that data points that are far away from given points have similar values. The grid masking semivarigoram appears most different from the other graphs with sparser points plotted at the specified lag parameters. This is likely due to the snapping of dissimilar points to each other. The partial sill of the WRP sermivariogram is identical to that of the original data points at 0.952, suggesting a closer approximation of the semivariogram model to the ODP than any of the other masking techniques. Random perturbation exhibits the highest partial sill value, and the plotted binned points appear more randomized along the semivariogram model than in the other masked representations. Directional influences did not appear in the semivariograms when tested. Due to the horizontal trend of the semivariograms, this test did not offer insights into whether obfuscation would impact exhibited spatial correlation in this manner.

Cluster Analysis (CA)

The next step in the analysis was a nearest neighbor hierarchical spatial clustering test, computed in CrimeStat. The parameters set for this test were a random nearest neighbor distance and a minimum threshold of 20 points per cluster. This minimum cluster size was selected to limit the quantity of clusters produced and view more generalized cluster patterns. First standard deviational ellipses of the resulting first order clusters for the original and masked data were overlaid in GIS. The original data points (ODP) generated 20 first order clusters, encompassing an average of 27.7 points each. The clustering statistics for the original and masked data are shown in Table 5.

Table 5. Cluster attributes from nearest neighbor hierarchical clustering

	First order clusters	Mean cluster points	Mean cluster density
Unmasked (ODP)	20	27.7	0.00045
Grid masking (GM)	17	28.0	0.00103
Voronoi masking (VM)	21	28.0	0.00102
Random perturbation (RP)	13	27.6	0.00093
Weighted random perturbation (WRP)	20	27.4	0.00105

The ODP, VM, and WRP results all produced a similar number of clusters, and the mean number of points in each output cluster is similar. RP produced the fewest first order clusters at 13, which is 7 fewer clusters than the ODP generated. Random perturbation, through random changes in distance and direction, does not tend to snap any nearby points together, as grid masking and Voronoi masking do, which can lead to comparatively less dense clustering patterns at fine scales. With a smaller number of clusters detected, RP without weighting may prove a less useful obfuscation technique for cluster analysis than the other masking methods. With weighted random perturbation, clusters are more likely to remain intact, as with a higher number of neighbors, a point is only moved a short distance. The mean density of the clusters detected for all of the obfuscation methods was higher than the mean cluster density for the ODP, indicating that the masking techniques tend to strengthen the cohesion of existing clusters.

The map in Figure 8 highlights similarities and differences of the locations and sizes of output clusters. For most of the ODP cluster ellipses shown, the VM ellipses appear to best approximate their location and size. This is particularly true of the northern portion of the map, where only the VM cluster ellipses match up with those of the ODP, and no other ellipses are closeby. Outliers by location for the group are emphasized towards the west by grid masking and weighted random perturbation. Unweighted RP fares the worst in this map, as RP ellipses are absent in

many cases where there is an ODP cluster. Based on the cluster analysis, VM demonstrates the best performance in approximating the underlying cluster patterns of the ODP.

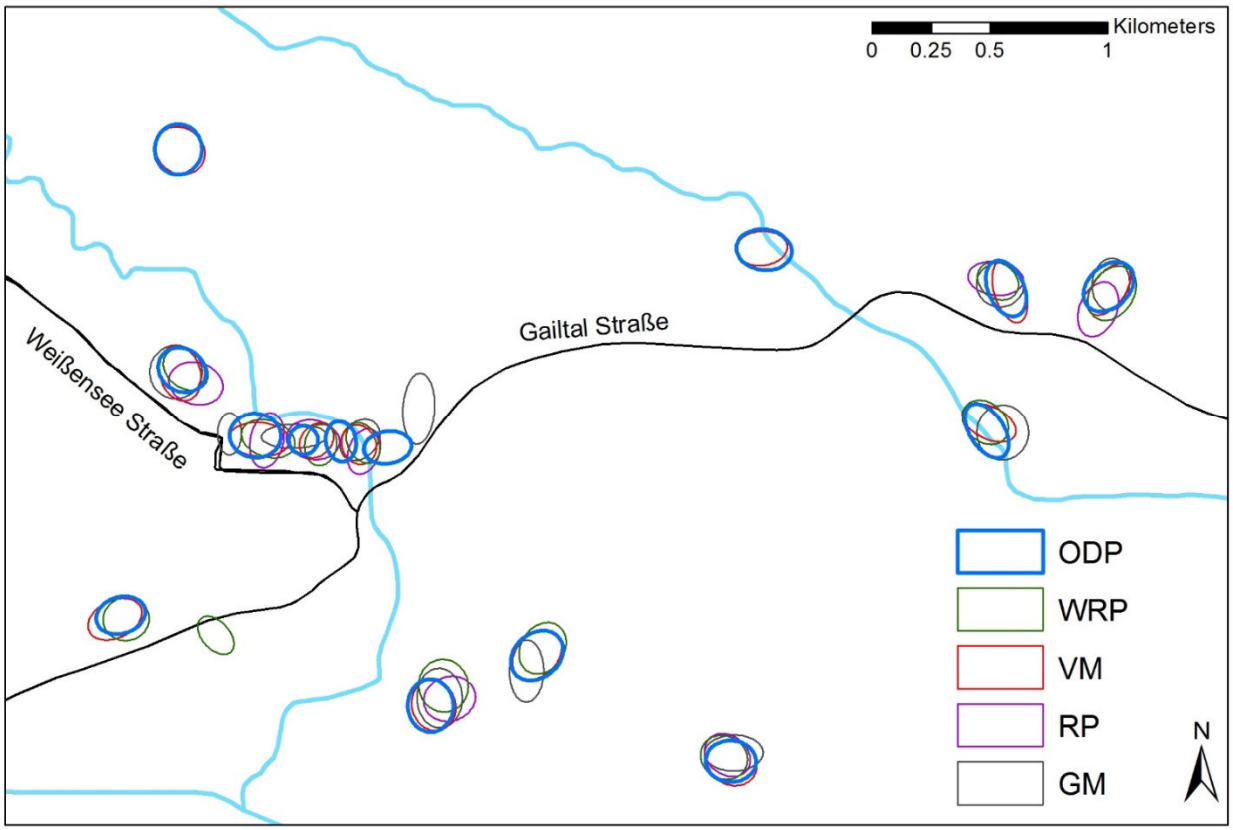


Figure 8. Subset of standard deviational ellipses for first order nearest neighbor hierarchical clusters

Point Similarity Analysis (PSA)

The point similarity analysis involved a cross-k analysis between the ODP and each of the obfuscated data sets. To set up the data for the cross-k function, the ODP were subsequently merged with each masked data set. The ODP points were set as the type *i* events with the masked data set as type *j* events so that the cross-k was run four times, once for each masking method. The cross-k function tests the degree to which the masked points are clustered around the ODP. Border correction was implemented using the administrative boundary of the Hermagor district.

The first set of results from the cross-k demonstrates strong clustering between all four masked data sets and the ODP. These results are shown in Figure 9. The point distributions of the *i* and *j* events are highly similar in all four resulting graphs, far exceeding the theoretical distributions to be found with random points, particularly with border correction. Differences between the masking techniques in this regard are difficult to detect from the plots. From low to high distances, the cross-k demonstrates that the underlying point distributions remain highly linked to each other.

The graphs displayed in Figure 10 plot the results of the cross-k for each obfuscation technique along with the simulation envelopes. The number of random simulations chosen for these data sets was 99 to demonstrate results at the $p<0.01$ confidence level. The results again show significant dependence between the ODP and masked data sets. At the scales of these plots, slight differences in performance between these masking techniques are difficult to detect. This suggests that cross-k plots and measuring the dependence between masked data sets may not capture finer differences in underlying spatial pattern. It places confidence in the masking techniques that the results should be spatially dependent on the original data points, but the methods may not be ideal for highlighting difference.

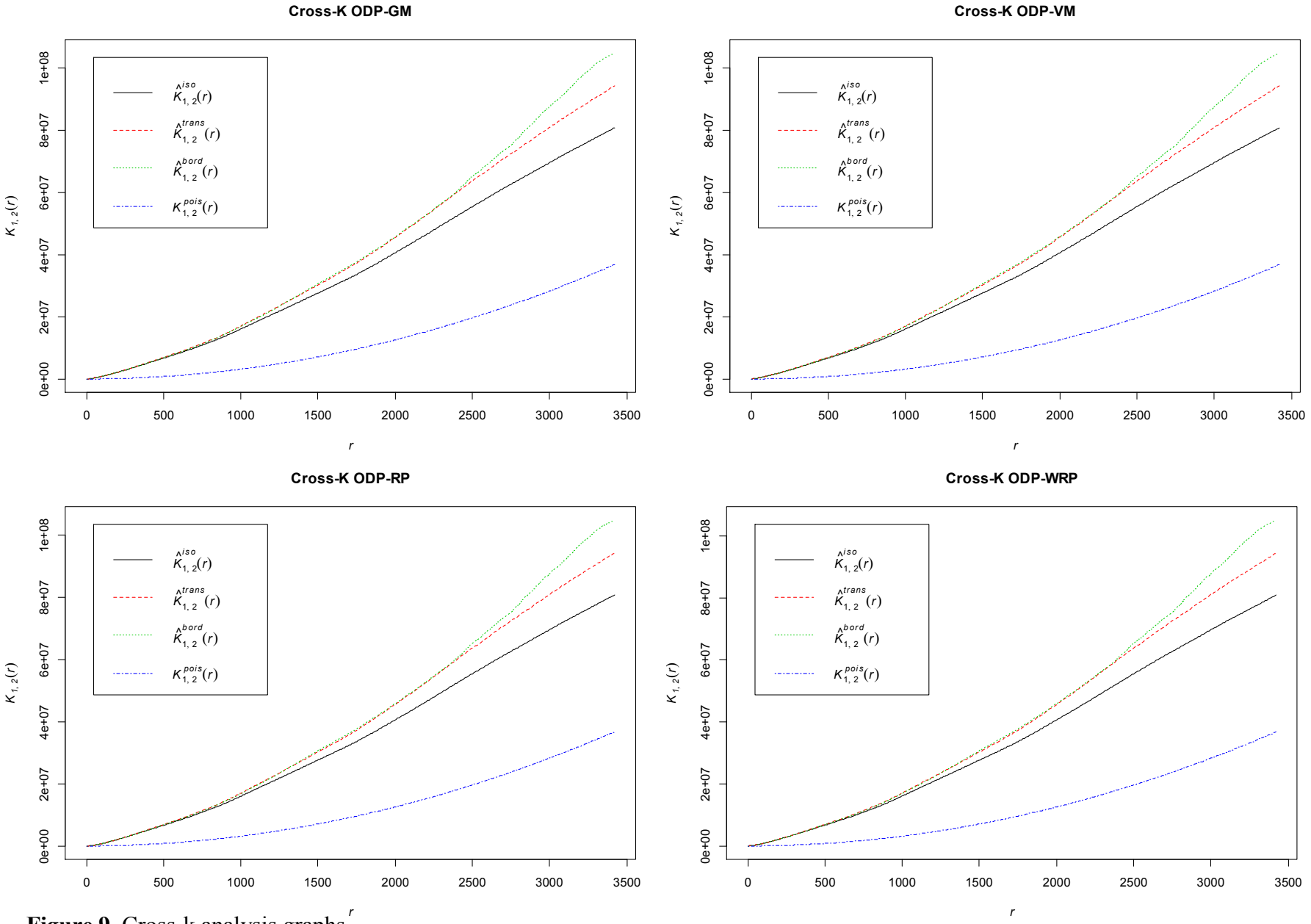


Figure 9. Cross-k analysis graphs

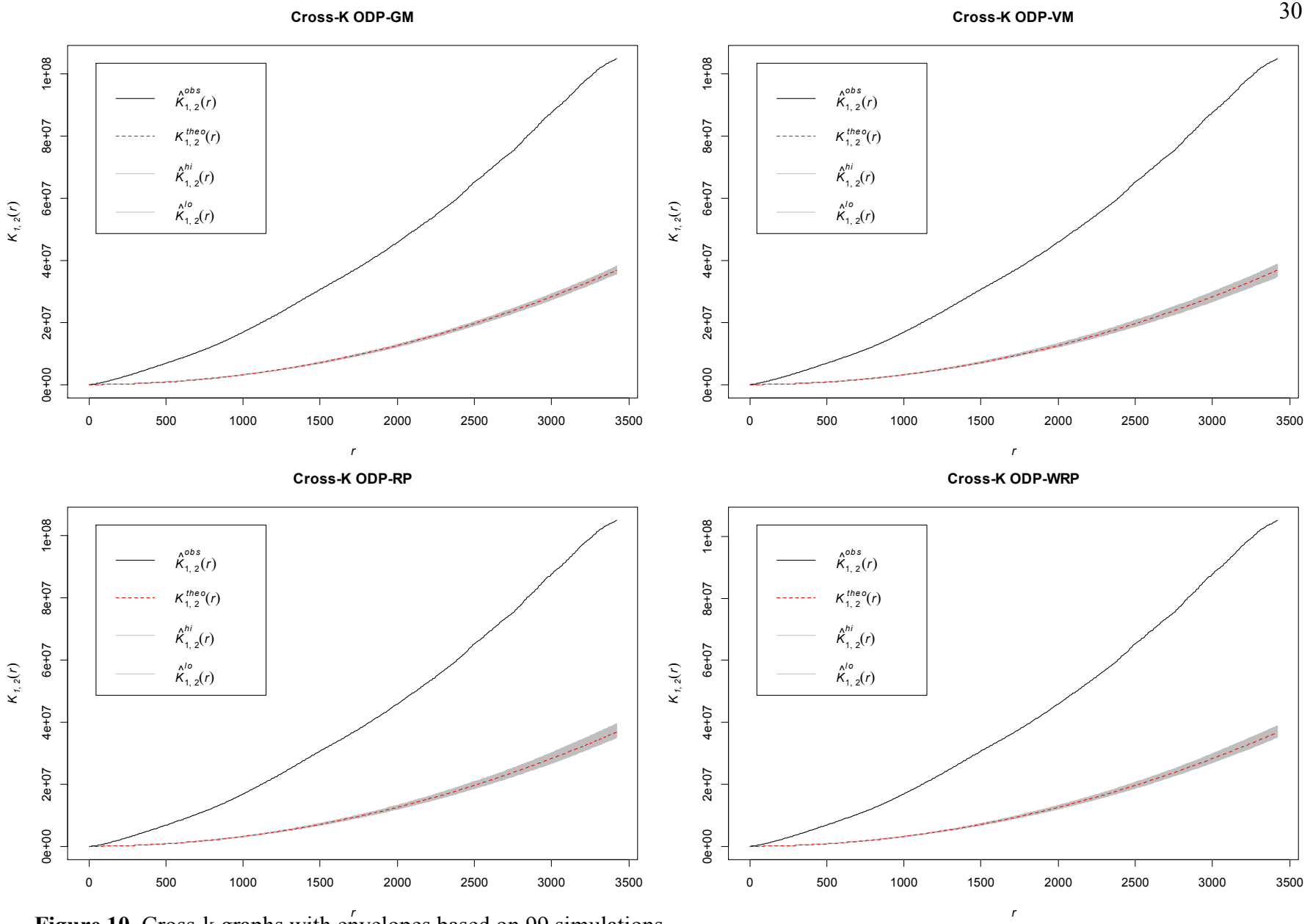


Figure 10. Cross-k graphs with envelopes based on 99 simulations

V. Conclusions

This study tested the performance of four obfuscation techniques in preserving the underlying spatial patterns of warm water energy consumption in the Hermagor District of Carinthia, Austria. Grid masking (GM), random perturbation (RP), and weighted random perturbation (WRP) all have previous documented uses in masking studies, but have not until this study been tested on an energy data set for all households in a single region. A major contribution of this study is the evaluation of Voronoi masking (VM) as an obfuscation technique. Between all the tests of underlying spatial pattern, VM outperforms the other obfuscation methods for preserving point distributions.

Similarity of Obfuscated Results

For exploratory statistics, the mean and median centers for VM are closer to the ODP means and medians compared to the other obfuscation techniques. This overall result is reaffirmed with a more complex examination of difference maps from a kernel density estimation of the original and masked data points. Voronoi masking and weighted random perturbation exhibit less variation from the kernel density rasters of the ODP. This is expected, since both of these methods are better tailored to the underlying spatial structure of data, moving points smaller distances where the density of points is higher. This better maintains patterns in concentrated areas as well as maintains the k-anonymity of the masked points.

One test where the other masking measures performed better than VM in matching the ODP results was with a global Moran's I, where the ODP pattern was clustered. VM did not reach significance for clustering, but the GM and RP data did. The semivariogram analysis did not provide much insight in this regard, since the trend of the data was primarily horizontal. In the nearest neighbor hierarchical cluster analysis, however, VM again best approximated the results

of the ODP, more closely matching the number of first-order clusters, their mean number of points, and their location. Grid masking and random perturbation fared more poorly in these regards. In the point similarity analysis, conducted using a cross-k analysis to test the dependence between the ODP and the obfuscated data, all the results exhibited significant spatial dependence. This remained true when 99 simulations were run to test significance. Given the almost identical nature of the cross-k plots among obfuscation methods, the overall similarity between all masked points and the ODP is confirmed. However, the results suggest that a different point similarity test would be best to uncover slight variations in the levels of dependence between the point structures.

Performance of Grid Masking as Representing Statistical Units

An objective of this study was to determine how well masking would fare for privacy and pattern preservation compared to the currently implemented technique of aggregation to 125-meter statistical units. The grid masking used in this study, which snapped the ODP to the centroids of 125-meter grid cells, best approximates this aggregation technique and serves as its proxy in this analysis. The results for GM present convincing evidence that this aggregation is more disruptive of spatial patterns than alternative masking techniques. GM demonstrated greater departure from the ODP kernel density patterns, as shown in the KDE difference maps for warm water energy consumption. In the nearest neighbor cluster analysis, grid masking resulted in fewer clusters detected, and the cluster ellipses tended to be offset from the ODP ellipses. Performing a cluster analysis based on such an aggregated pattern is more likely to lead to inaccurate results that could impact decision support. If cluster analysis is a key part of analysis, this study recommends an obfuscation technique that is more tailored to underlying data structure, such as VM or WRP.

Privacy Preservation

Another key part of this research was determining how well each obfuscation technique preserved privacy as measured by k-anonymity. This measure relies on the distance to the *k*th nearest neighbor. As this study included all residences in the study area, this is a truer measure of k-anonymity than would be measured based on a data set that only included a percentage of the population. VM and GM both lowered the average distance to the 1st nearest neighbor. This is because both techniques tend to snap points to each other, placing them at identical locations. This preserves privacy by making a viewer of the data less likely to deductively infer which household a given data point originates from. The greater the number of nearby neighbors, the less power is available for such inferences. As the number of neighbors increased to 10 and 20, VM continued to outperform the other masking techniques, even GM, for lowest average distance to the *k*th nearest neighbor.

A next step for research on Voronoi masking is an evaluation of how it could possibly be reversed to reveal identities. While it outperforms the other obfuscation methods for the metrics of spatial pattern and privacy tested in this study, an obfuscation method is only valuable it cannot be reversed and deciphered. The advantage for privacy in random perturbation and its weighted cousin is that the randomization makes the resulting pattern challenging to reverse engineer to find actual identities for data records. The pattern in VM is not random, and is starkly based on underlying spatial structure. More research is needed in reverse engineering of masking techniques and the ability to infer identities.

References

Allshouse, W.B., Fitch, M.K., Hampton, K.H., Gesink, D.C., Doherty, I.A., Leone, P.A., Serre, M.L., Miller, W.C. 2010. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. Geocarto International 25(6): 443-452.

AbdelMalik, P., M. N. K. Boulos, and R. Jones. 2008. The perceived impact of location privacy: A web-based survey of public health perspectives and requirements in the UK and Canada. *Bmc Public Health* 8.

Curtis, A., Mills, J.W., Agustin, L., and Cockburn, M. 2011. Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems* 35(1): 57-64.

Duckham, M. and L. Kulik. 2007. Location privacy and location-aware computing. In *Dynamic & mobile GIS: Investigating change in space and time*, ed. J. Drummond, R. Billen, E. Joao, and D. Forrest. Boca Raton, FL: CRC Press 34-51.

Elwood, S., and A. Leszczynski. 2011. Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum* 42 (1):6-15.

Hampton, K.H., Fitch, M.K., Allshouse, W.B., Doherty, I.A., Gesink, D.C., Leone, P.A., Serre, M.L., Miller, W.C. 2010. Mapping health data: improved privacy protection with donut method geomasking. *American Journal of Epidemiology* 172(9): 1062-1069.

Kar, B., R.C. Crowsey, and J.J. Zale. 2013. The Myth of location privacy in the United States: surveyed attitude versus current practices. The Professional Geographer 65(1): 47-64.

Kounadi, O. and Leitner, M. 2014. Why does geoprivacy matter? The scientific publications of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics* 9(4): 34-45.

Krumm, J. 2007. Inference attacks on location tracks. *5th International Conference, Proceedings, PERVASIVE 2007* Toronto, Canada, May 13-16, 2007. 127-143.

Kwan, M. P.; Casas, I.; Schmitz, B. C. 2004. Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks?. *Cartographica* 39: 15-28.

Leitner, M., and A. Curtis. 2004. Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives* 49: 22-39.

Luo, L., McLafferty, S., Wang, F. 2010. Analyzing spatial aggregation error in statistical models of late-stage cancer risk: a Monte Carlo simulation approach. *International Journal of Health* Geographics 9 (51): 1-14.

Nouwt, S. 2008. Reasonable expectations of geo-privacy? SCRIPTed 375 5(2) http://www.law.ed.ac.uk/ahrc/script-ed/vol5-2/nouwt.pdf

Paulus, G., Kosar, B., Erlacher, C., Anders, K.H. 2014. Energy efficient communities – development of a WebGIS portal for managing local energy data. American Association of Geographers Annual Meeting. Tampa, FL.

Shi, X., Alford-Teaster, J., and Onega, T. 2009. Kernel density estimation with geographically masked points. *Proceedings of the 17th International Conference on Geoinformatics* August 2009.

STATCube - Statistical Database of Statistics Austria. 2014. http://statcube.at/superwebguest/

Sweeney, L. 2002. k-Anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5): 557-570.

Vicente, C.R., Freni, D., Bettini, C. and C.S. Jensen. 2011. Location-related privacy in geo-social networks. *Internet Computing* IEEE 20-27.

Zandbergen, P.A. 2014. Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Advances in Medicine* 2014: 1-14.