Spatial Autocorrelation and Analytical Validity of Geomasked Health Microdata

Seminar Report

Mohamed Hassan (ID: 272455)

January 30

Abstract

Georeferenced microdata enable fine-grained spatial health analyses but create privacy risks because exact point locations can facilitate re-identification. A common privacy protection approach is geomasking, which perturbs coordinates while keeping attributes unchanged. This report evaluates how two geomasking methods—Donut Masking and Adaptive Gaussian Distance Masking (AGDM)—affect spatial autocorrelation in a binary (rare) health outcome, `sleep_diso`, using synthetic point microdata for Cologne (n = 10,900). Spatial autocorrelation is assessed globally (Moran's I, Geary's C) and locally (Local Moran's I / LISA) using k-nearest neighbors weights (main: $k = 8$, sensitivity: $k \in \{6, 8, 10, 12\}$) with 999-permutation inference. Results indicate weak but statistically detectable positive spatial autocorrelation in the original data. Both masking methods attenuate spatial autocorrelation, with AGDM preserving local patterns slightly better than Donut according to local similarity metrics.

1 Introduction

1.1 Motivation: privacy vs analytical validity

Point-level health microdata with coordinates are highly valuable for studying neighborhood effects, environmental exposures, and localized clustering. However, exact coordinates are also highly identifying. Even if explicit identifiers are removed, individuals can potentially be re-identified through linkage to external information (e.g., address databases) or by unique combinations of location and demographics. Therefore, data providers often anonymize georeferenced microdata before release.

Geomasking is a common anonymization strategy that modifies point locations. While it reduces disclosure risk, it can also distort spatial relationships and potentially harm the analytical validity of spatial methods that rely on neighbor structure and distances. This report focuses on **analytical validity**: whether spatial autocorrelation patterns are preserved under geomasking.

1.2 Research question

How do geomasking methods (Donut vs AGDM) affect spatial autocorrelation in the binary health outcome `sleep_diso`, and which method preserves spatial autocorrelation patterns more closely to the original data?

2 Data

2.1 Dataset

The analysis uses synthetic point microdata for Cologne with $n = 10{,}900$ individuals. Each record includes:

- `id`, `district`, `gender`, `age`
- `sleep_diso`: binary health outcome (1 = sleep disorder, 0 = no sleep disorder)
- `geometry`: point coordinate

Three versions of the dataset are available as shapefiles (attributes identical; geometry differs):

- Original coordinates: `seminar_data/original_points.shp`
- Donut-masked coordinates: `seminar_data/moved_points_donut.shp`
- AGDM-masked coordinates: `seminar_data/moved_points_agdm.shp`

2.2 Outcome characteristics

`sleep_diso` is binary and described as rare (many zeros, few ones). For rare binary outcomes, spatial autocorrelation effect sizes (e.g., Moran's I) are often small in magnitude even when statistically significant. Therefore, interpretation emphasizes relative differences between original and masked datasets rather than expecting large global indices.

3 Geomasking methods (conceptual overview)

In both masking approaches, individual attributes remain unchanged and only point coordinates are displaced.

- **Donut Masking:** random displacement within a ring (bounded by a minimum and maximum radius), intended to avoid very small moves that reveal the original location and very large moves that severely distort geography.
- **AGDM:** an adaptive displacement approach intended to balance privacy protection and utility (the precise parameterization is taken as given by the provided masked dataset).

Figure 1 illustrates the displacement visually by overlaying original and masked point locations.

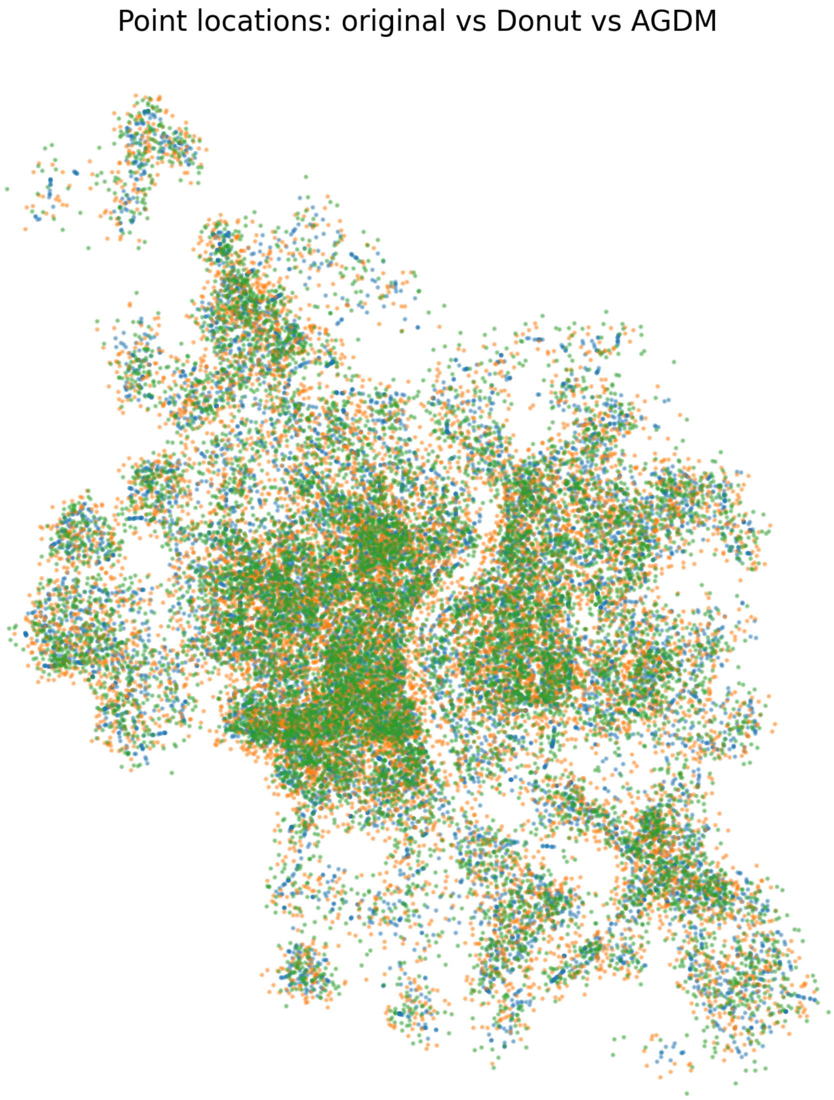

Figure 1: Overlay of point locations for original, Donut-masked, and AGDM-masked geometries.

4 Methods

4.1 Spatial weights

Spatial autocorrelation statistics require a definition of "neighbors." For point microdata, k-nearest neighbors (kNN) weights are used. kNN has two practical advantages: (i) it ensures every point has neighbors (avoiding isolates), and (ii) it is less sensitive to varying point density than a fixed distance band.

The main analysis uses $k = 8$ nearest neighbors with **row-standardized** weights. Sensitivity is checked for $k \in \{6, 8, 10, 12\}$.

4.2 Global spatial autocorrelation

Two global measures are computed for `sleep_diso`:

- **Global Moran's I:** summarizes overall similarity among neighbors (positive values indicate clustering of similar values).
- **Geary's C:** emphasizes neighbor differences (values below 1 indicate positive autocorrelation; values closer to 1 indicate weaker autocorrelation).

Significance is assessed using **permutation tests** with 999 permutations for each dataset.

4.3 Local spatial autocorrelation (LISA)

Local Moran's I (LISA) is computed with 999-permutation inference for each point, producing:

- a local statistic per point,
- a significance decision,
- and a cluster/outlier label for significant points: High–High (HH), Low–Low (LL), High–Low (HL), Low–High (LH), or Not significant.

4.4 Similarity evaluation for local results

To compare original vs masked local results, three similarity metrics are computed:

- **Spearman correlation** between Local Moran's I values (original vs masked),
- **Jaccard overlap** of significance patterns (significant vs not significant),
- **Cluster agreement rate** for points where at least one dataset is significant.

5 Results

5.1 Sensitivity of global Moran's I to k

Figure 2 shows Moran's I for multiple k values. The main purpose is to assess whether the relative differences between original and masked datasets are stable when k changes.

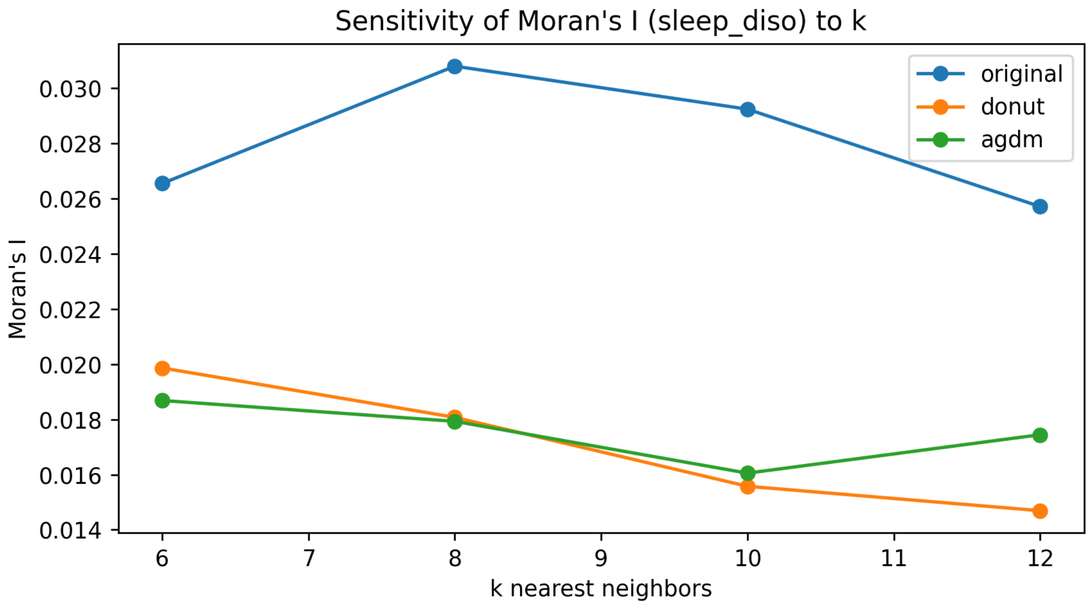

Figure 2: Sensitivity analysis: Global Moran's I for `sleep_diso` across $k \in \{6, 8, 10, 12\}$ for original, Donut, and AGDM datasets.

5.2 Global autocorrelation

Table 1 summarizes global Moran's I and Geary's C (with permutation p-values) for the main specification ($k = 8$).

Table 1: Global spatial autocorrelation for `sleep_diso` (kNN, $k = 8$, 999 permutations).

Dataset	Moran's I	p	Geary's C	p
Original	0.0308	0.001	0.9697	0.001
Donut	0.0181	0.001	0.9861	0.037
AGDM	0.0179	0.001	0.9768	0.001

Interpretation. All three datasets show statistically significant positive spatial autocorrelation, but effect sizes are small, consistent with a rare binary outcome. Both masking methods attenuate global autocorrelation relative to the original (Moran's I decreases; Geary's C moves closer to 1). AGDM is closer to the original than Donut on Geary's C, suggesting slightly better preservation of global spatial structure on that measure.

5.3 Local autocorrelation (LISA maps)

Figure 3 compares LISA cluster maps for the original and masked datasets. The visual goal is to assess whether significant clusters/outliers appear in similar locations and retain similar cluster labels after masking.

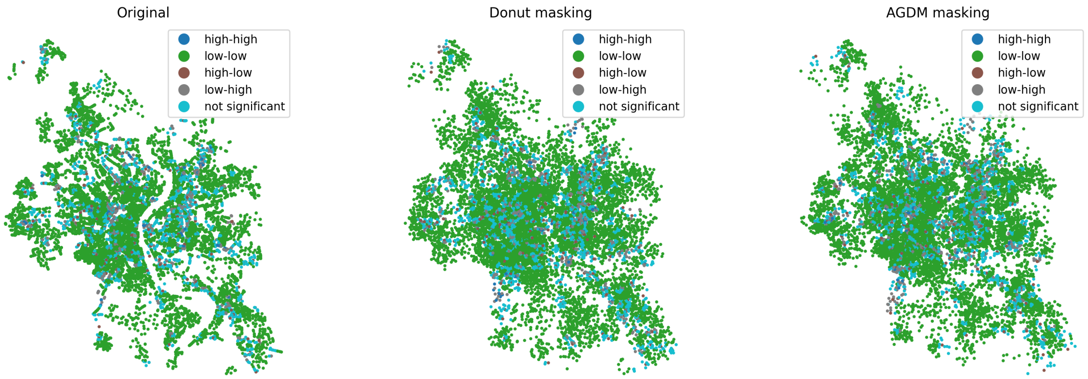

Figure 3: LISA cluster comparison (original vs Donut vs AGDM) for `sleep_diso`.

5.4 Quantitative similarity of local results

Table 2 reports similarity metrics for Donut vs original and AGDM vs original.

Table 2: Local similarity metrics (masked vs original) for LISA results on `sleep_diso`.

Masked method	Spearman ρ (Local I)	Jaccard (significance)	Cluster agreement
Donut	0.2258	0.7225	0.6422
AGDM	0.3225	0.7406	0.6776

Interpretation. AGDM shows higher similarity than Donut on all three metrics, indicating slightly better preservation of local spatial autocorrelation patterns. Nonetheless, absolute Spearman correlations are modest, consistent with the fact that local statistics are sensitive to changes in neighborhood structure introduced by geomasking.

6 Discussion

6.1 Analytical validity implications

For global summaries, masked datasets still show statistically detectable positive autocorrelation, but with reduced magnitude. For local mapping, geomasking can change which points are significant and how cluster types are labeled, which matters for hotspot/outlier interpretation. In this evaluation, AGDM preserves local patterns slightly better than Donut.

6.2 Why effect sizes are small

Because `sleep_diso` is binary and rare, most neighborhoods contain many zeros, and clustering signals are expected to be weak. With a large sample size, permutation tests can still return small p-values even for small effect sizes. Therefore, conclusions emphasize relative changes across datasets rather than absolute magnitudes.

7 Limitations and future work

- **Dependence on spatial weights:** results can vary with k and with alternative weight definitions (distance bands, distance decay). Sensitivity was checked for selected k values, but further robustness checks are possible.
- **Binary rare outcome:** local clusters may be fragile; interpretation should be cautious.
- **Synthetic data:** findings demonstrate methodological effects, but validation on real datasets and outcomes is needed.
- **Privacy not quantified:** this report evaluates utility/validity; a complete privacy–utility assessment would also include disclosure risk metrics.

8 Conclusion

Geomasking alters spatial relationships and can attenuate both global and local spatial autocorrelation patterns. In the Cologne synthetic dataset, `sleep_diso` exhibits weak but statistically detectable positive autocorrelation in the original data. Both Donut and AGDM reduce measured autocorrelation, but AGDM preserves local patterns slightly better than Donut according to similarity metrics and global Geary's C. For researchers working with anonymized georeferenced microdata, masking method choice can meaningfully affect analytical validity, especially for local cluster mapping.