Spatial Autocorrelation and Analytical Validity of Geomasked Health Microdata

Donut Masking vs. AGDM on a binary (rare) outcome

Mohamed Hassan

Seminar on Anonymization of Georeferenced Microdata

January 29, 2026

Roadmap (30-minute talk)

- Motivation: privacy vs analytical validity
- Data and geomasking methods
- Key concepts and terminology (spatial autocorrelation, neighbors, LISA)
- Methods and sensitivity checks
- Results: global, local, and similarity metrics
- Conclusions, limitations, and discussion

Motivation & Problem Setting

- Georeferenced microdata enable fine-grained health and exposure analyses.
- Exact coordinates are highly identifying $\rightarrow$ strong privacy risks.
- Geomasking protects privacy by perturbing locations, but can distort:
  - spatial weights / neighborhood structure,
  - global spatial autocorrelation,
  - local clusters and hotspot detection.
- Core trade-off: **privacy protection vs. analytical validity**.

Key Terms: Microdata and Georeferenced Microdata

- **Microdata:** one row = one individual (e.g., a person or patient).
- **Georeferenced microdata:** microdata with a location (point coordinate).
- Why it matters: location enables spatial analysis but increases re-identification risk.

Data

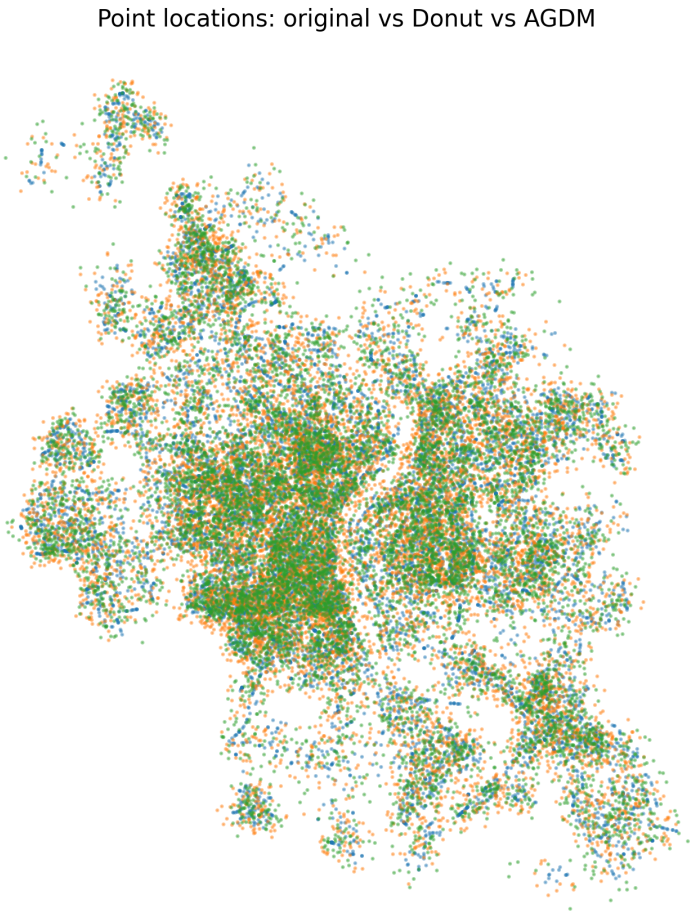
- Synthetic point microdata for Cologne, $n = 10{,}900$ individuals.
- Attributes per record: `id`, `district`, `gender`, `age`, `sleep_diso` (0/1), geometry.
- Outcome of interest: `sleep_diso` (binary; rare $\rightarrow$ weak clustering expected).
- Three point datasets (same attributes, different geometry):
  - original: `seminar_data/original_points.shp`
  - donut-masked: `seminar_data/moved_points_donut.shp`
  - AGDM-masked: `seminar_data/moved_points_agdm.shp`

Outcome of Interest: `sleep_diso`

- **Binary variable:** 1 = sleep disorder, 0 = no sleep disorder.
- **Rare outcome:** many zeros and few ones.
- Implication: spatial autocorrelation measures are expected to be *small* in magnitude.

Geomasking Overview: What Changes vs. What Stays

- **Changes:** point geometry (coordinates).
- **Unchanged:** all attributes (`sleep_diso` labels remain attached to records).
- Two methods compared:
  - Donut masking (random displacement with minimum/maximum radius).
  - AGDM (adaptive approach aiming to balance privacy/utility).



Geomasking: Quick Definitions

- **Geomasking:** deliberately shifting point locations to protect privacy.
- **Donut masking:** random displacement within a ring (min and max radius).
- **AGDM:** adaptive displacement aiming to balance privacy and utility.
- In all cases: attributes stay the same, only coordinates change.

Research Question & Evaluation Goal

- **RQ:** How do Donut vs. AGDM affect spatial autocorrelation in `sleep_diso`?
- **Goal:** Identify which masking method preserves spatial autocorrelation patterns most closely to the original data.
- Evaluate both:
  - **Global** dependence (overall clustering/dispersion),
  - **Local** dependence (where clusters/outliers appear).

Concept: Spatial Autocorrelation (Plain Language)

- **Autocorrelation:** how similar a variable is to itself across space.
- **Positive spatial autocorrelation:** similar values cluster together.
- **Negative spatial autocorrelation:** high values near low values.
- **No autocorrelation:** spatial pattern looks random.

Method Overview

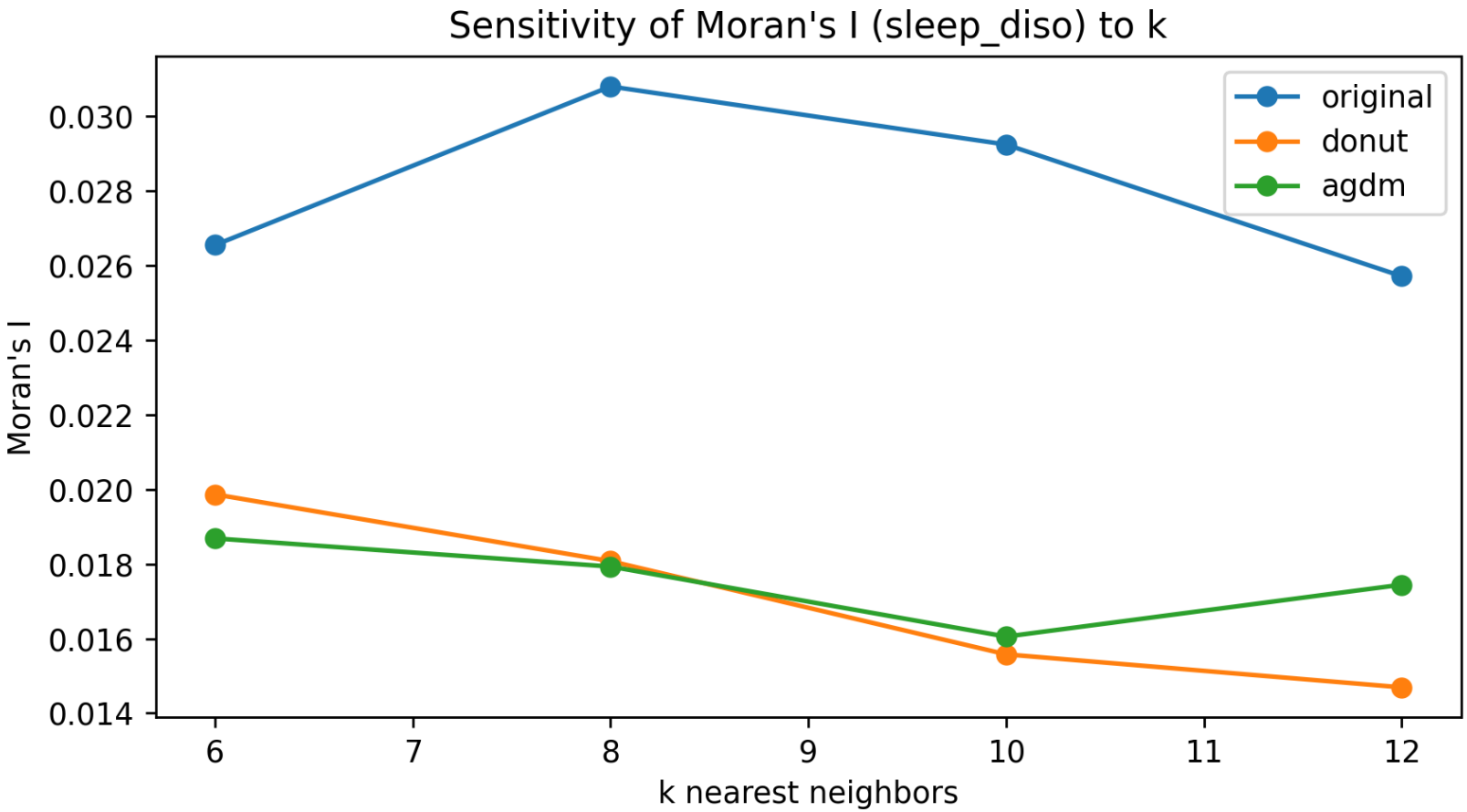
- Spatial weights: k-nearest neighbors (kNN), row-standardized.
- Main specification: $k = 8$; sensitivity: $k \in \{6, 8, 10, 12\}$.
- **Global** spatial autocorrelation (999 permutations):
  - Global Moran's I (positive I indicates clustering).
  - Geary's C ($C < 1$ indicates positive autocorrelation).
- **Local** spatial autocorrelation (LISA; 999 permutations):
  - Local Moran's I + cluster types: High–High, Low–Low, High–Low, Low–High, Not significant.

Concepts: Neighbors and Spatial Weights

- Spatial analysis needs a rule for "who is near whom."
- **k-nearest neighbors (kNN):** each point uses its k closest points.
- **Row-standardized weights:** each point's neighbors sum to 1.
- Why it matters: changing neighbors changes autocorrelation results.

Spatial Weights Choice & Sensitivity

- Why kNN for point microdata:
  - ensures connectedness (no isolated observations),
  - robust to varying point density,
  - directly reflects "local neighborhood" for LISA.
- Main analysis uses $k = 8$ (balance: local detail vs. stability).
- Sensitivity check: stability of Moran's I across $k \in \{6, 8, 10, 12\}$.



Global vs Local Measures (Definitions)

- **Global measures:** one number for overall clustering (Moran's I, Geary's C).
- **Local measures (LISA):** identify clusters and outliers at specific locations.
- Both are needed: global summaries can hide local changes.

Global Results: Moran's I and Geary's C (sleep_diso)

- Moran's I values are small (rare binary outcome) but statistically significant (999 permutations).
- Both masking methods reduce global autocorrelation relative to the original.
- AGDM appears closer to original on Geary's C (closer to 1 means weaker autocorrelation; original shows stronger positive autocorrelation).

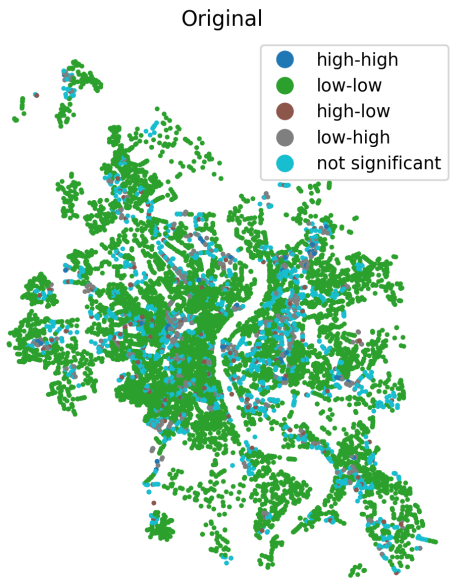
Dataset	Moran's I	p	Geary's C	p
Original	0.0308	0.001	0.9697	0.001
Donut	0.0181	0.001	0.9861	0.037
AGDM	0.0179	0.001	0.9768	0.001

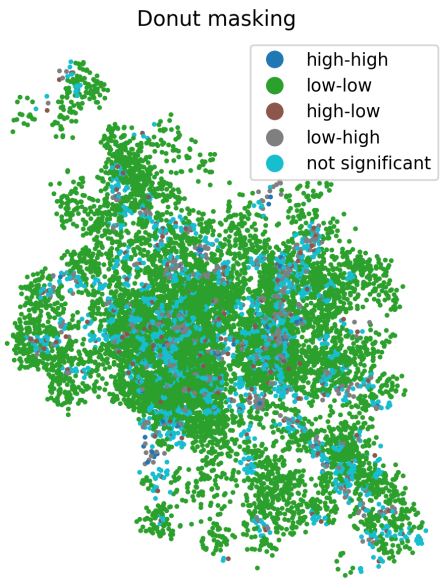
How to Interpret the Global Results

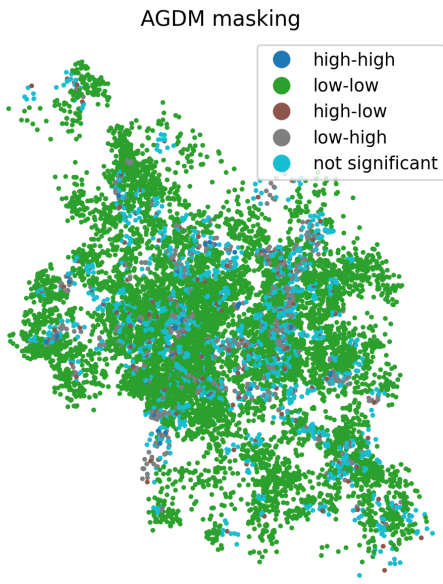
- **Moran's I:** small positive values indicate weak clustering.
- **Geary's C:** values closer to 1 indicate weaker autocorrelation.
- Comparison logic:
  - If masked values move toward 0 (Moran) or 1 (Geary), clustering is weakened.
  - The closer to original, the better the preservation of spatial structure.

Local Results: LISA Cluster Maps

- LISA identifies localized clustering/outliers in `sleep_diso` relative to neighbors.
- Cluster types:
  - High–High / Low–Low: local clustering,
  - High–Low / Low–High: spatial outliers,
  - Not significant: no evidence at chosen α (permutation-based).
- Visual comparison: do significant areas and cluster types persist after masking?







LISA Cluster Labels (Quick Definitions)

- **High–High (HH):** high value surrounded by high values (hotspot).
- **Low–Low (LL):** low value surrounded by low values (coldspot).
- **High–Low (HL):** high value surrounded by low values (outlier).
- **Low–High (LH):** low value surrounded by high values (outlier).
- **Not significant:** no local pattern detected.

Quantitative Similarity: Local Autocorrelation Preservation

- Compare original vs masked:
  - Spearman correlation of Local Moran's I values,
  - Jaccard overlap of significance (significant vs not),
  - Cluster agreement rate where at least one dataset is significant.
- Higher values indicate better preservation of local patterns.

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

How to Read the Similarity Metrics

- **Spearman correlation:** do points rank similarly in local Moran's I?
- **Jaccard overlap:** are the same points significant?
- **Cluster agreement:** do significant points keep the same label?
- Higher values = closer to the original spatial pattern.

Conclusion

- Both masking methods attenuate spatial autocorrelation in `sleep_diso`.
- **AGDM preserves spatial dependence slightly better overall**:
  - closer global Geary's C to original,
  - higher local similarity (Spearman, Jaccard, cluster agreement).
- Practical implication: for analyses relying on local clustering/hotspots, masking choice matters; AGDM may offer better analytical validity under similar privacy constraints.

Limitations & Discussion Questions

- Dependence on spatial weights choice (kNN, choice of k); sensitivity helps but does not eliminate subjectivity.
- Binary, rare outcome: small Moran's I expected; statistical significance can reflect large n.
- Synthetic data: results may differ for real population distributions and true exposure mechanisms.
- No explicit privacy-utility curve here (utility assessed; privacy strength not quantified in this deck).

Discussion prompts

- How should analysts choose k (or weights) when only masked data are available?
- Would stronger displacement (larger radii) break local clusters entirely, and how to detect that?
- Which utility targets matter most: global indices, hotspot stability, regression coefficients, or model fit?

Summary: Main Takeaways

- Geomasking protects privacy but changes neighborhood structure.
- Spatial autocorrelation is weak (rare outcome) but still measurable.
- Both Donut and AGDM reduce global and local autocorrelation.
- AGDM preserves spatial patterns slightly better than Donut.
- Analysts should choose masking methods carefully when local patterns matter.