Randomly generated spatial datasets: A Python approach

Paddy Gorry & Peter Mooney

FOSS4G UK Local 2023
Access to data & teaching

- Openly available spatial datasets (*OpenStreetMap*)
- Important in the teaching of spatial data analysis (Jarvis, 2011):
  - Importing/exporting datasets
  - Data *cleaning*
  - Data *visualization*
  - Data *analysis* (spatial or otherwise)
- Generating *custom datasets* is useful for teaching any (or all) of the above topics
Introducing... RADIAN

- RA\text{n}D\text{o}m spat\text{I}al d\text{A}ta generator
  - Created as part of my MSc thesis (supervised by Peter Mooney)
- Produce \textbf{synthetic geographic datasets} that:
  - Appear “\textbf{realistic}”
  - Available in GeoJSON & SQL formats
  - Can be generated \textbf{quickly}!
- Exercises that are \textbf{interactive}, \textbf{interesting}, and \textbf{relevant} help encourage student engagement (Donker et al, 2022)
Related Work: Synthetic Spatial Data

• The GRD package in SPSS and R allows generation of synthetic datasets (Calderini, 2019)

• Mannino & Abouzied described a software to generate “real-looking synthetic data” by specifying the properties of the dataset (Mannino, 2019)

• These tools can’t generate spatial datasets!
QGIS

- Quantum GIS
- An Open-Source GIS visualization and analysis tool
- Can generate random points in polygons and other shapes
Generates a set of uniformly distributed points within a polygon
- These points are given a single primary key attribute

- **Pros**: Speed, ease-of-use, variety of formats for export & visualization
- **Cons**: Points distributed unrealistically, no options for any additional metadata
Generating points “realistically”

What makes geographic data look “real”? 

Both datasets contain 409 points: one is real, one fake
Packages

• Some of the main Python packages used for this project:

  GeoPandas
  Shapely
  scikit-learn

• Random, numpy, Matplotlib, etc.
Random points in a Polygon

• Given some region described by a polygon
• Determine the bounding box:
  • The max/min latitude and longitude (x & y) values that contains the polygon
• Generate a 2D point using the max/min x & y values as the parameters
• Remove the points which lie within the bounding box but outside of the desired polygon
A closer look at “fake” data

Without context or real data for comparison how can we tell this data is fake?

How popular is a takeaway in the river?
A closer look at “fake” data

- **Context** is important in identifying if data appears realistic
- The **uniform distribution** here could in fact be realistic depending on **what data it is emulating**
- This will also be influenced by the **type of region** within which the points are being generated (country vs. city)
A radial distribution

• Note the difference in **density** as you travel further from the centre

• We generate **an equal number** of points in each buffer region

• The smaller central regions (A) will have a much **higher density** than the outer regions (E)

• **This results in a “radial” distribution** that mimics real geographic data
Voronoi Polygons

- The partitioning of a plane with **n points** into convex polygons such that:
  - Each polygon contains exactly one generating point
  - Every point in a given polygon is closer to its generating point than to any other.
- Burrough et al, 2015
Voronoi-based Buffers

• Circular buffers are too uniform – hence we use **Voronoi-based buffers**

• **K-Means clustering** is used to create a set of **centroids** to create **Voronoi regions**

• These regions are classed by their **distance to the centroid** of the source polygon
• Generation can be split into **primary** and **secondary**
  • Primary occurring at the level of the **source polygon** (moving or original centroid)
  • Secondary occurring at the local-level in **secondary Voronoi regions** (equal or variable area)

• The **ratio** between primary and secondary points is set in the parameters
Incorporating additional metadata to points

- Radian is controlled by a `.JSON` parameter file.
- This allows for generation of additional variables:
  - Random strings, integers, and timestamps
  - Custom variables given a list of strings and a set of weights
- Datasets can be tailored to the topic being taught or assessed.
A synthetic dataset of 400 Fast-Food restaurants in Glasgow

An example of some metadata attached to the points
Use cases for RADIAN

• Generate unseen data for **CS621: Spatial Databases**
• Allow students to generate their own data for assignments and projects
• Used to generate examples for teaching materials to increase engagement
• **ChatGPT** has shown the utility of being able to generate unseen synthetic data quickly
Future of RADIAN

• **Current work**
  • Use of raster grids and triangulation – more efficient generation
  • Improvements to current RADIAN implementation

• **Future work**
  • AI/ML solutions for generation of synthetic spatial data (GeoPointGan)
  • **Placement Project:** Synthetic satellite imagery

---

GeoPointGan: Synthetic Spatial Data with Local Label Differential Privacy

Teddy Cunningham
University of Warwick
Coventry, UK
teddy.cunningham@warwick.ac.uk

Hongkai Wen
University of Warwick
Coventry, UK
hongkai.wen@warwick.ac.uk

Konstantin Klemmer
University of Warwick
Coventry, UK
k.klemmer@warwick.ac.uk

Hakan Ferhatosmanoglu
University of Warwick
Coventry, UK
hakan.f@warwick.ac.uk
Questions & Queries: patrick.gorry.2015@mumail.ie

Thank you!