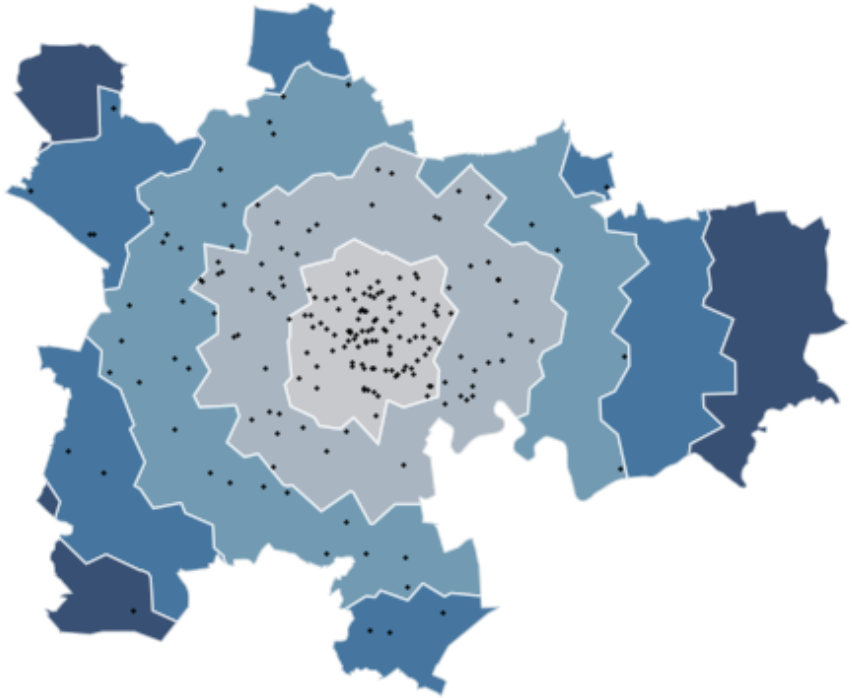
Centre for Research Training

Science Foundation Ireland For what's next

radiaπ



Randomly generated spatial datasets: A Python approach

Paddy Gorry & Peter Mooney

FOSS4G UK Local 2023

HOST INSTITUTIONS

University of Limerick Ollscoil Luimnigh

Maynooth University National University of Ireland Maynooth

UCD Dublin

FUNDED BY

Science Foundation Ireland For what's next

Skillnet Ireland

Access to data & teaching

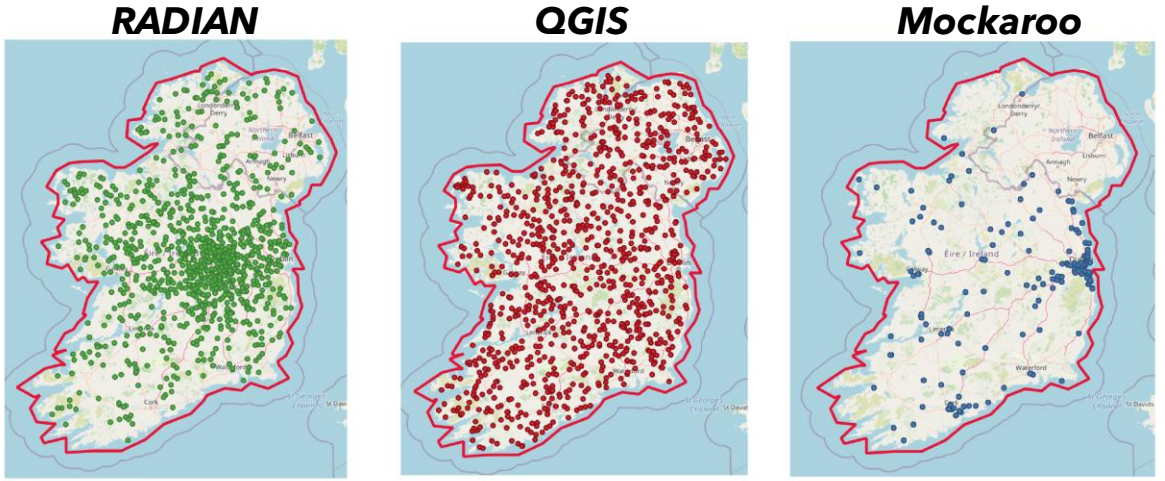
- Openly available spatial datasets (**OpenStreetMap**)
- Important in the teaching of spatial data analysis (Jarvis, 2011):
  - Importing/exporting datasets
  - Data **cleaning**
  - Data **visualization**
  - Data **analysis** (spatial or otherwise)
- Generating **custom datasets** is useful for teaching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

OpenStreetMap
The Free Wiki World Map

mockaroo

Introducing…**RADIAN**

- **RA**n**D**om spat**I**al d**A**ta ge**n**erator
  - Created as part of my **MSc thesis** (supervised by Peter Mooney)
- Produce **synthetic geographic datasets** that:
  - Appear "**realistic**"
  - Available in GeoJSON & SQL formats
  - Can be generated **quickly**!
- Exercises that are **interactive**, **interesting**, and **relevant** help encourage student engagement (Donker et al, 2022)

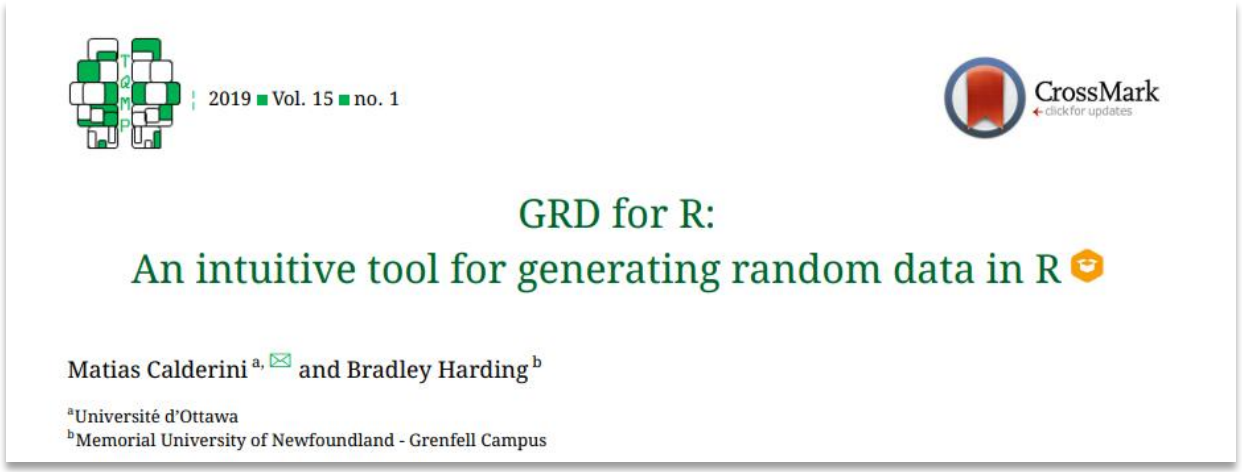


Equal-area Voronoi

Variable-area Voronoi - Equal Points

Variable-area Voronoi - Points by Area

Related Work: Synthetic Spatial Data

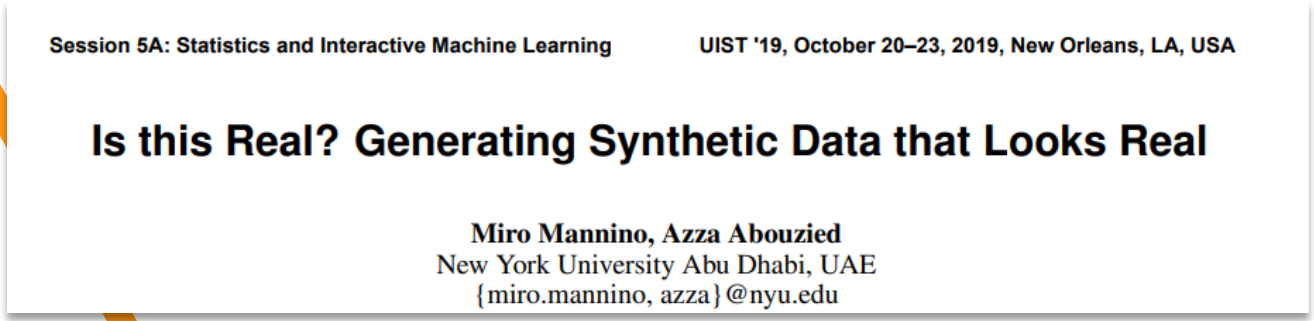


2019 ■ Vol. 15 ■ no. 1

GRD for R:
An intuitive tool for generating random data in R

Matias Calderini [a], ✉ and Bradley Harding [b]

[a] Université d'Ottawa
[b] Memorial University of Newfoundland - Grenfell Campus



Session 5A: Statistics and Interactive Machine Learning UIST '19, October 20–23, 2019, New Orleans, LA, USA

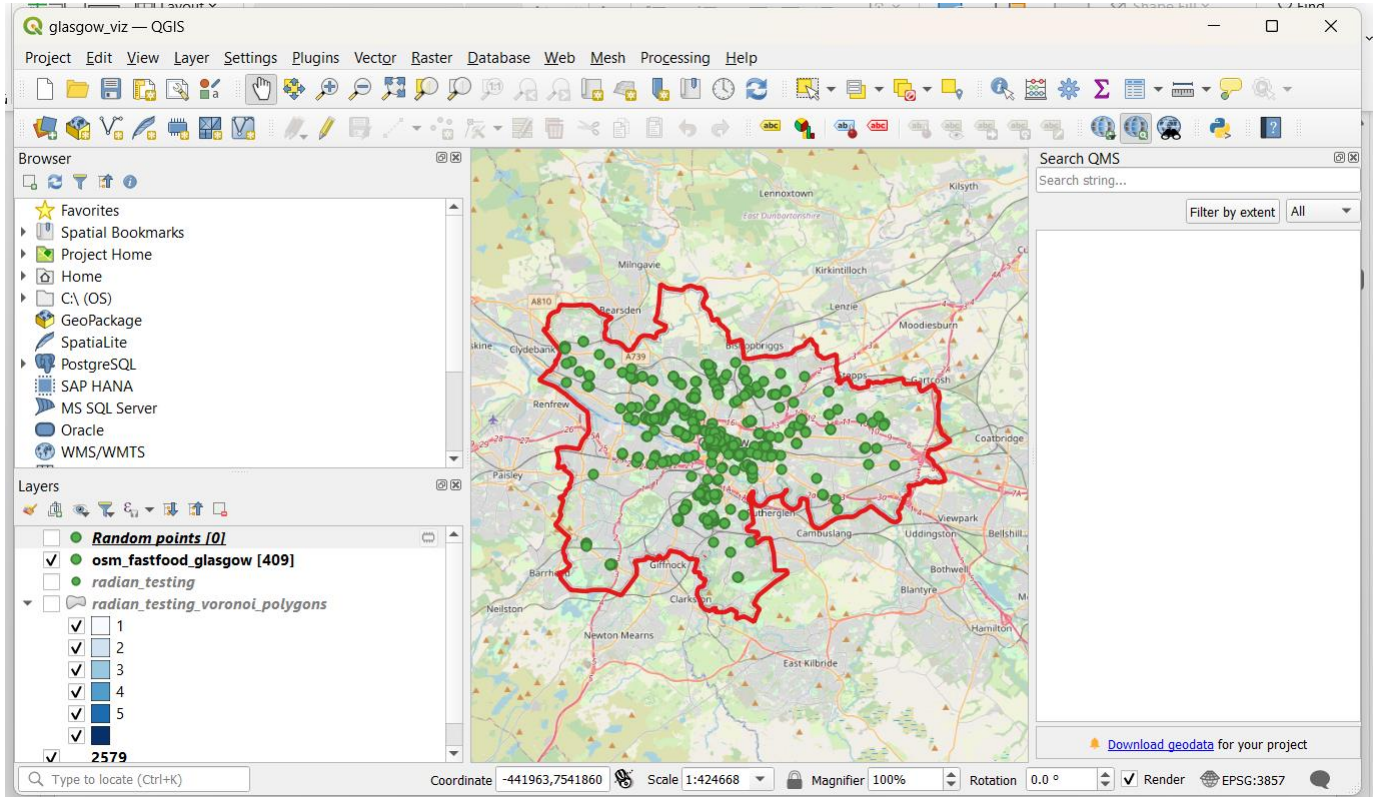
Is this Real? Generating Synthetic Data that Looks Real

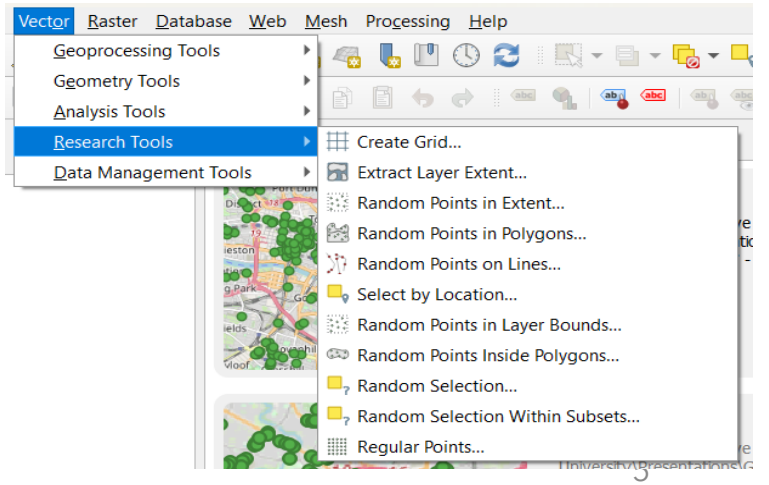
Miro Mannino, Azza Abouzied
New York University Abu Dhabi, UAE
{miro.mannino, azza}@nyu.edu

- 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!**

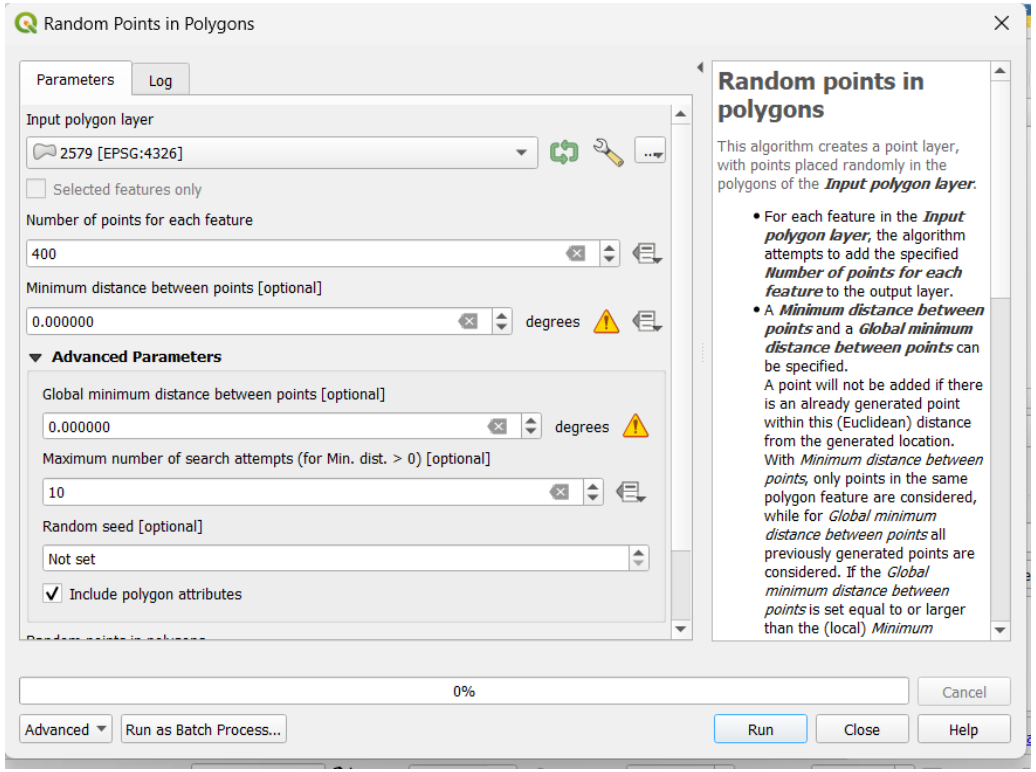
QGIS

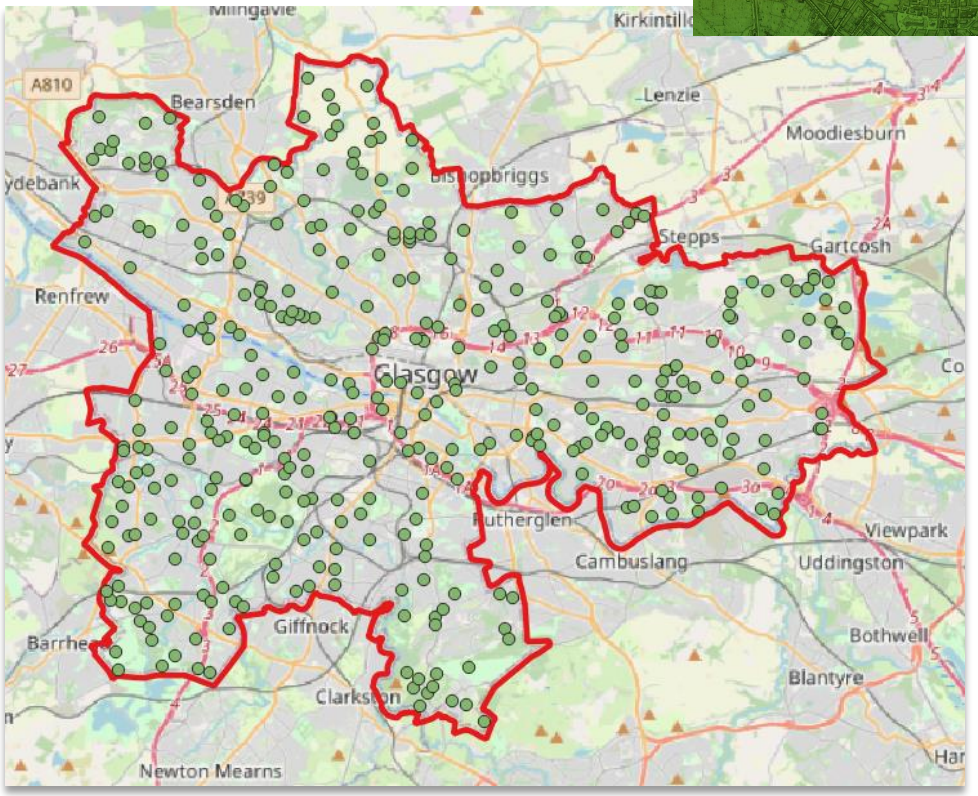
- Quantum GIS
  - An Open-Source GIS **visualization** and **analysis** tool
  - Can **generate random points** in polygons and other shapes





QGIS & Random Point Generation



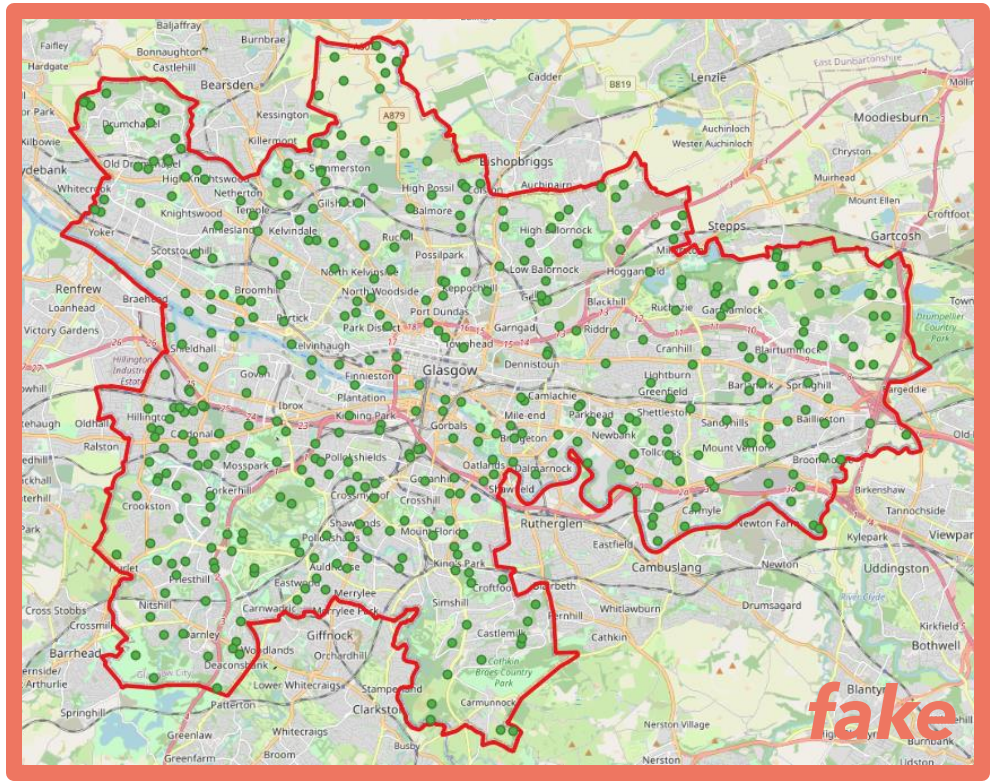


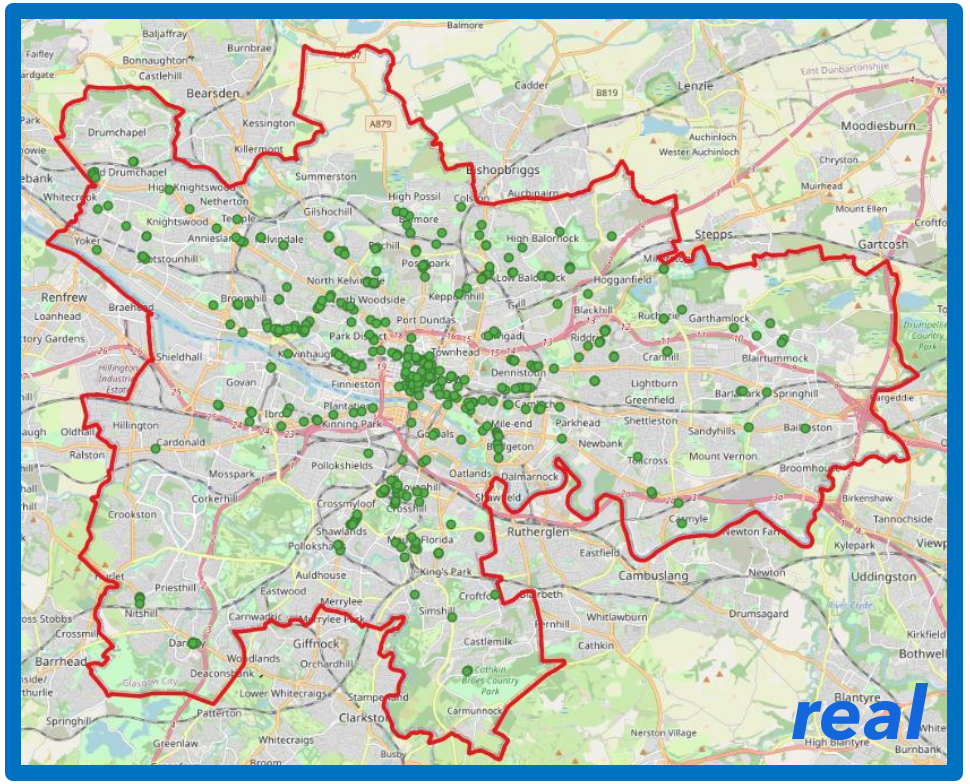
- Generates a set of **uniformly distributed** points within a polygon
- These points are given a single **primary key** attribute

- **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

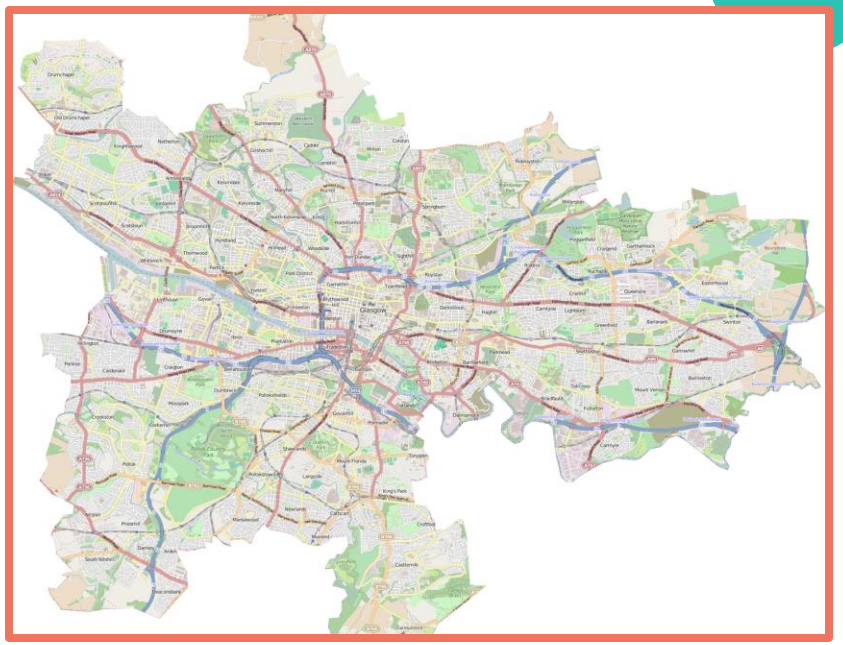
Packages

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

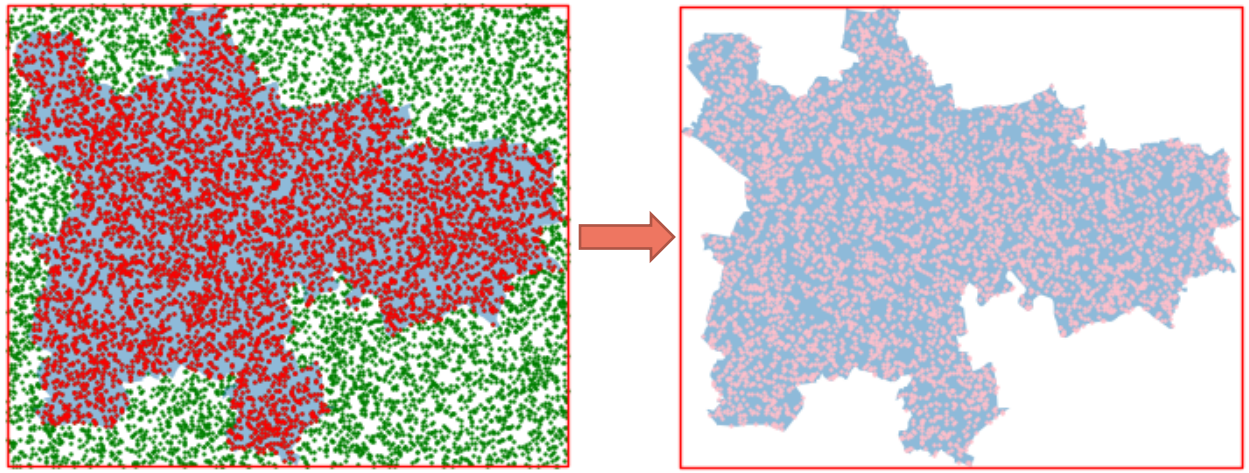
- Random, numpy, Matplotlib, etc.

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**

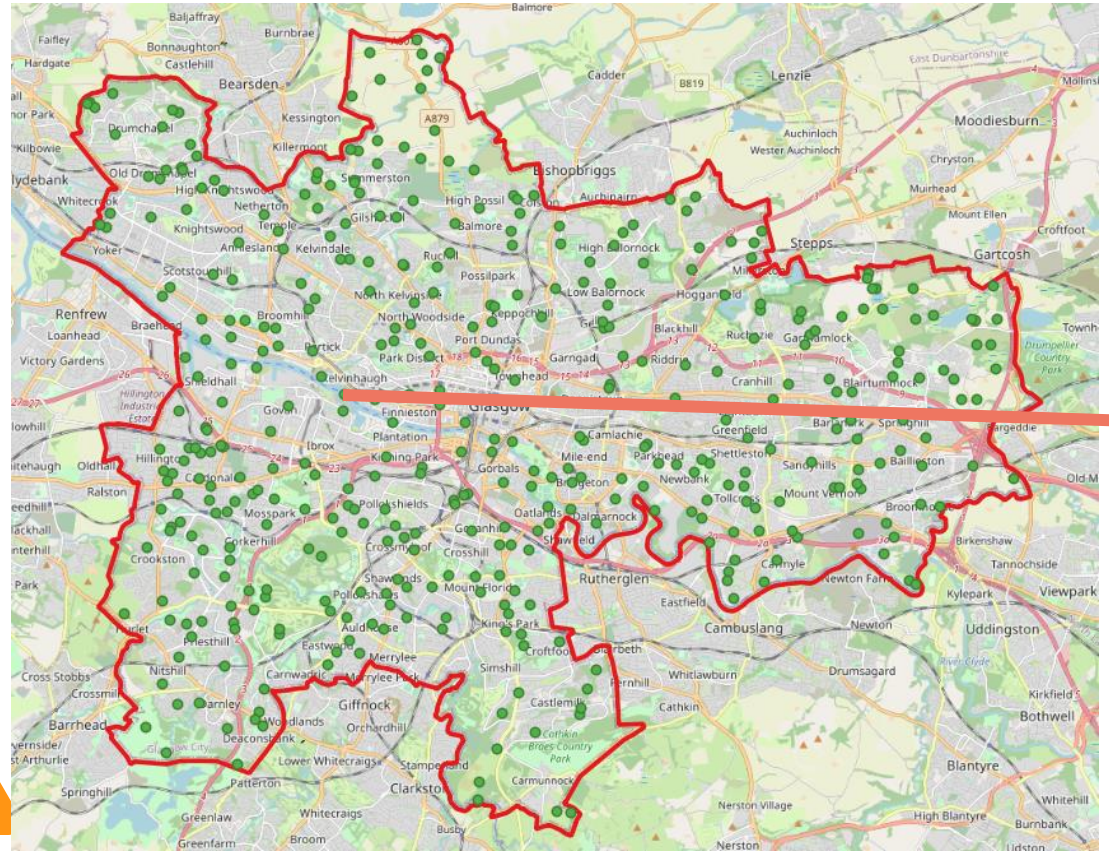


```
def point_in_poly(poly):
    minX, minY, maxX, maxY = poly.bounds
    new_point = shapely.Point([random.uniform(minX, maxX), random.uniform(minY, maxY)])
    return new_point
```



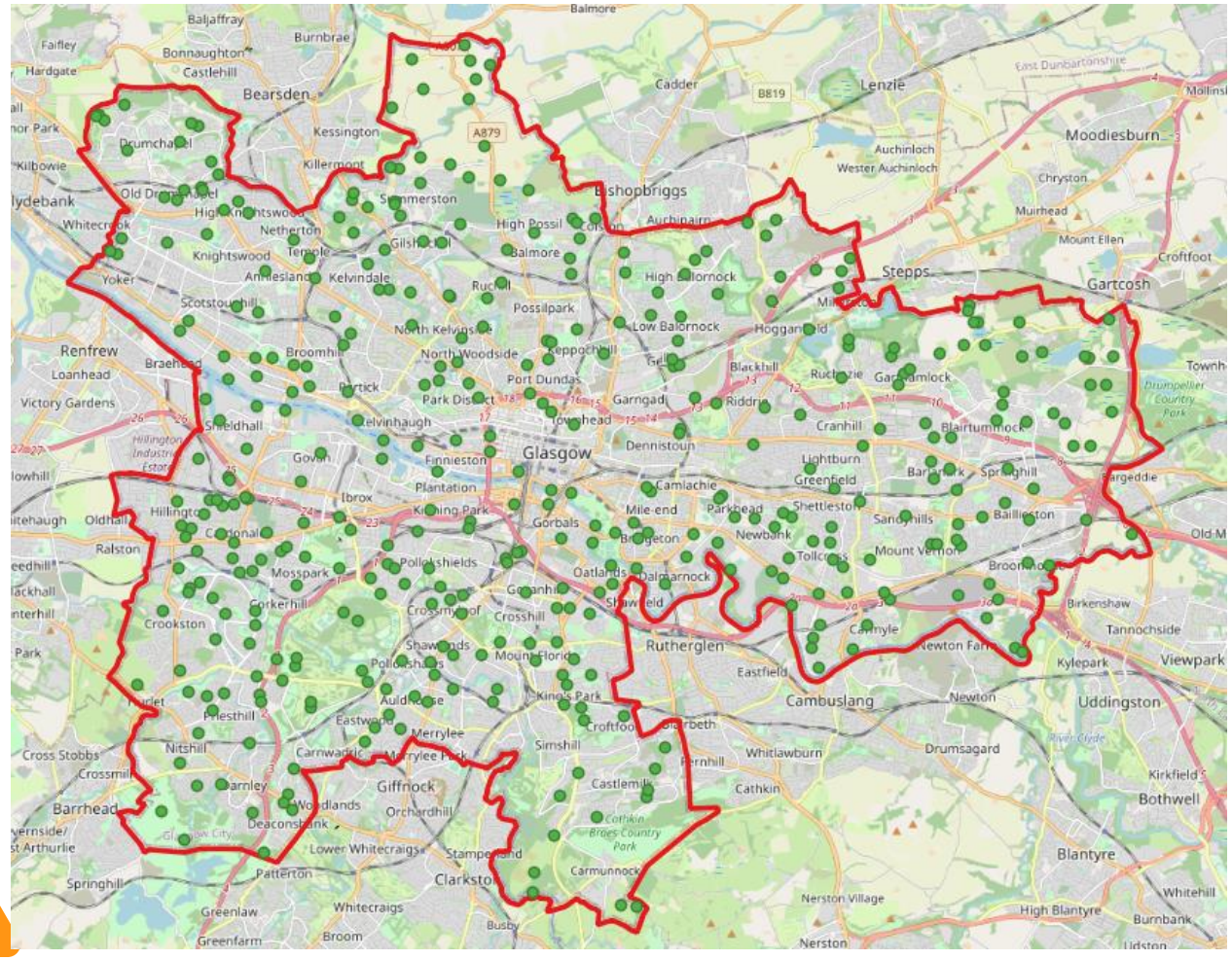
A closer look at “fake” data

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



How popular is a takeaway in the river?

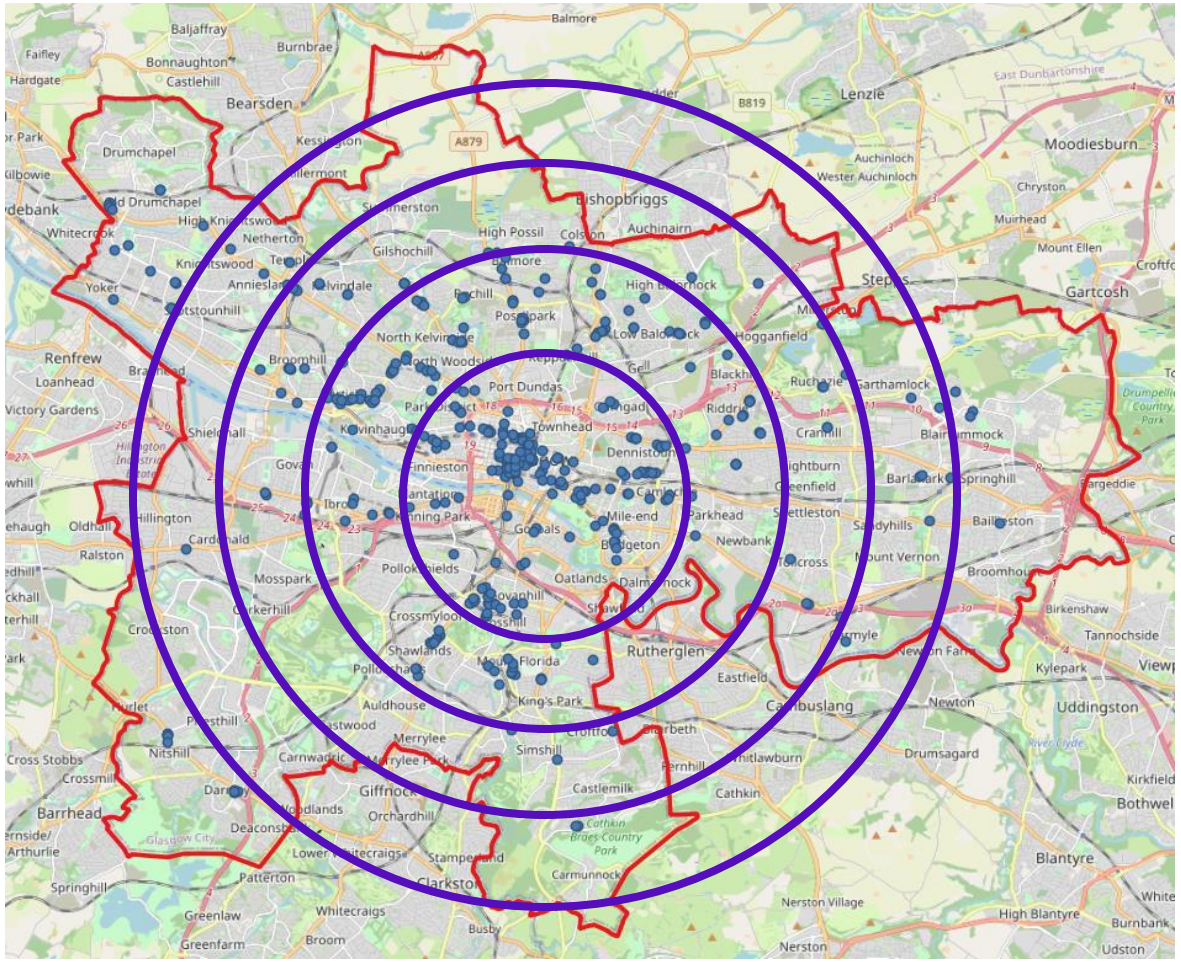
A closer look at “fake” data



- **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)

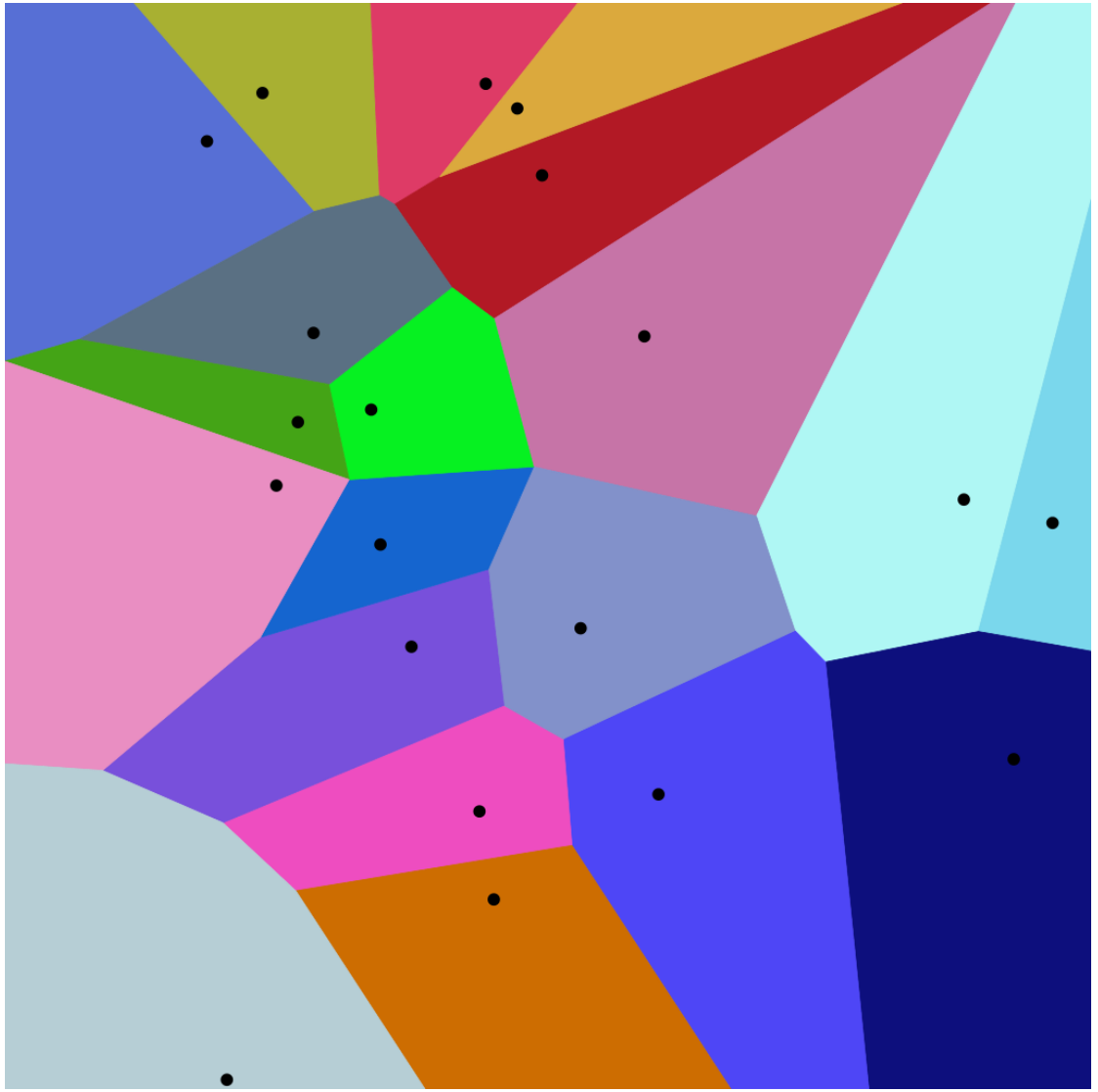
A radial distribution

- 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

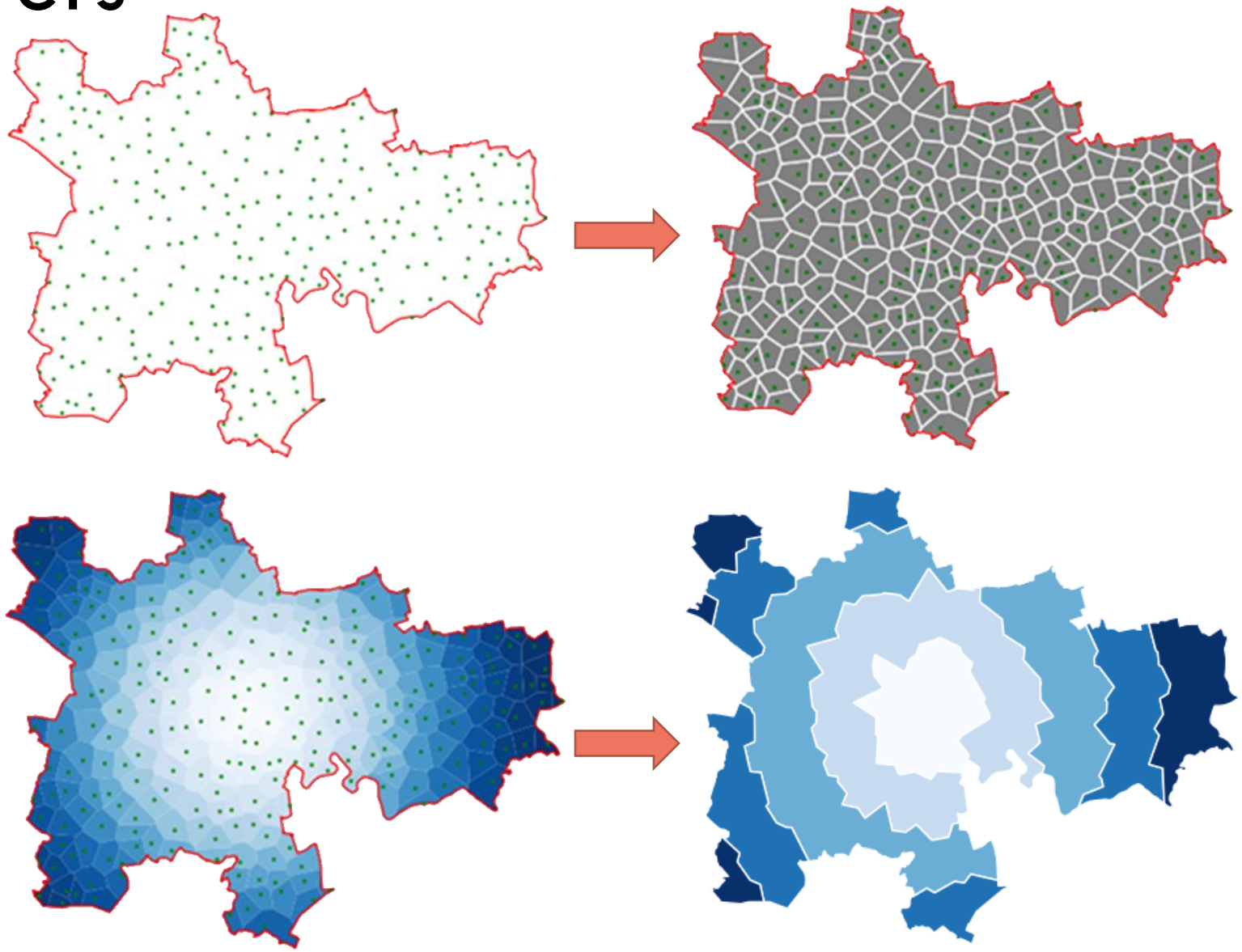
- The partitioning of a plane with **n points** into convex polygons such that:
  - **Each polygon contains exactly one generating point**
  - Every **point in a given polygon is closer to its generating point than to any other.**
  - Burrough et al, 2015



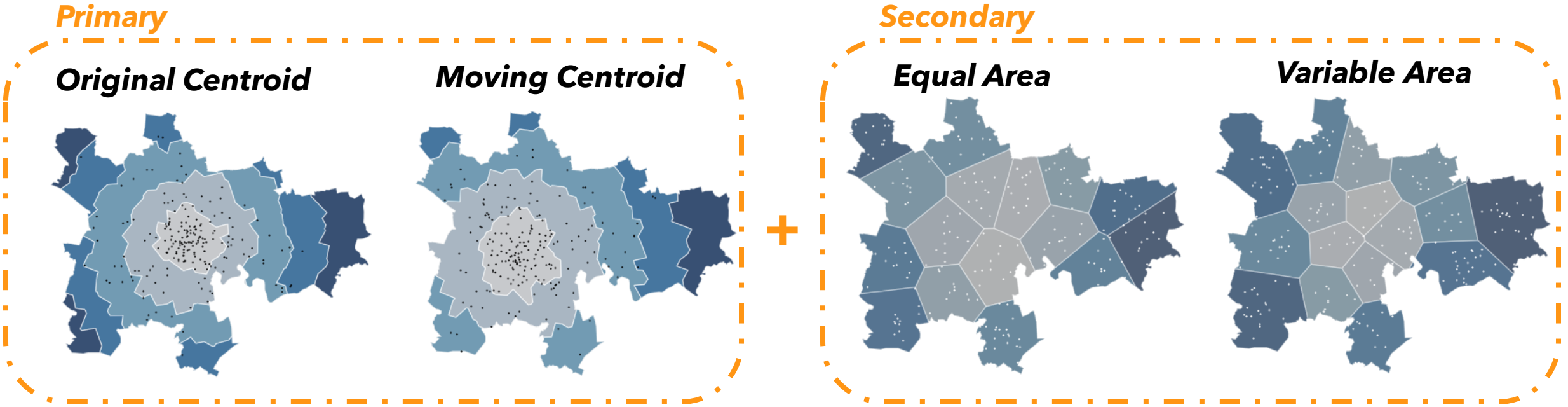
N=20

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



Primary & Secondary Generation

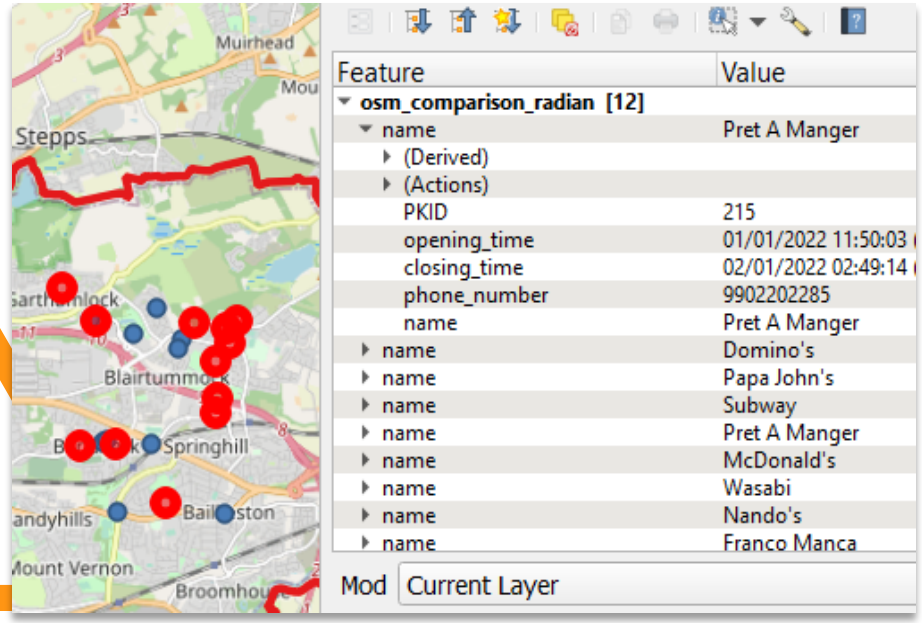


- Generation can be split into **primary** and **secondary**
  - Primary occurring at the level of the **source polygon** (moving or original centroid)
  - Secondary occurring at the local-level in **secondary Voronoi regions** (equal or variable area
- The **ratio** between primary and secondary points is set in the parameters

Incorporating additional metadata to points

```
"default_vars" : true,
"rand_var_types" : ["ts", "ts", "int"],
"rand_var_names" : ["opening_time", "closing_time", "phone_number"],
"rand_var_params" : [["2022-01-01 11:00:00", "2022-01-01 15:00:00"],["2022-01-01 22:00:00", "2022-01-02 03:00:00"],[1000000000,9999999999]],
```

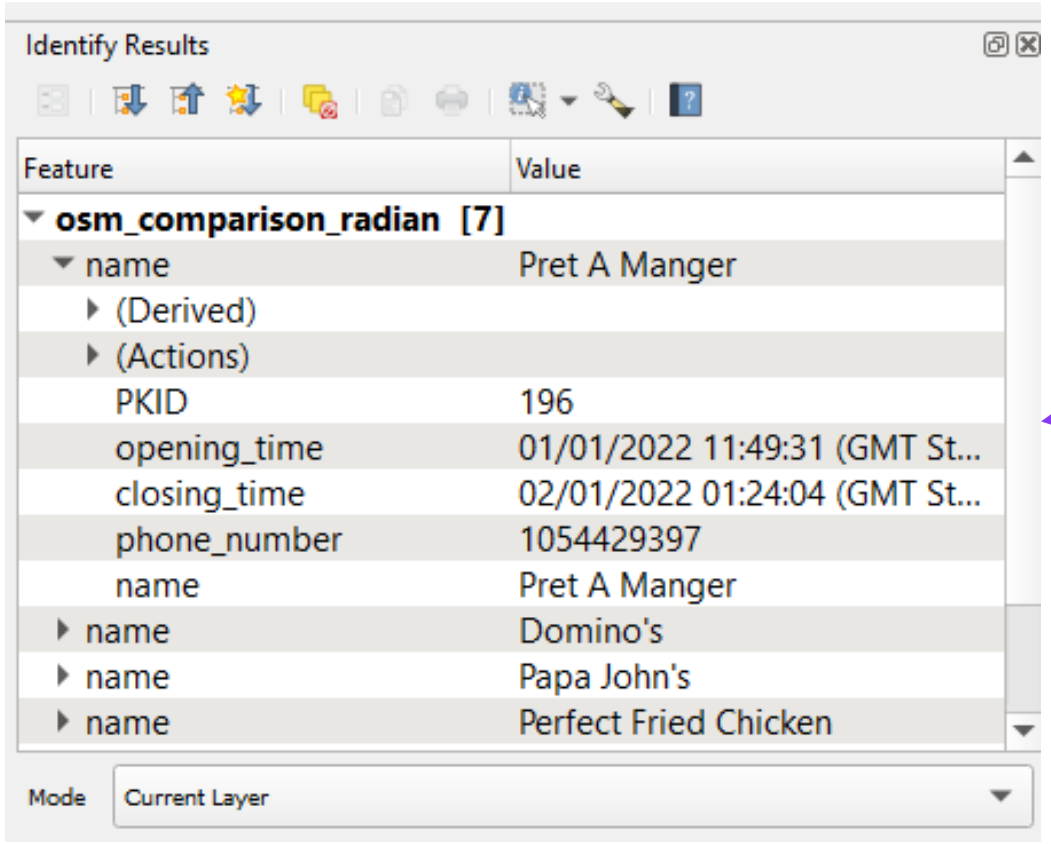
```
"extra_var":false,
"extra_var_types" : ["str", "str"],
"extra_var_name":["var_string", "string_equal_dist"],
"extra_var_file":["restaurant.csv", "restaurant_no_weights.csv"],
```



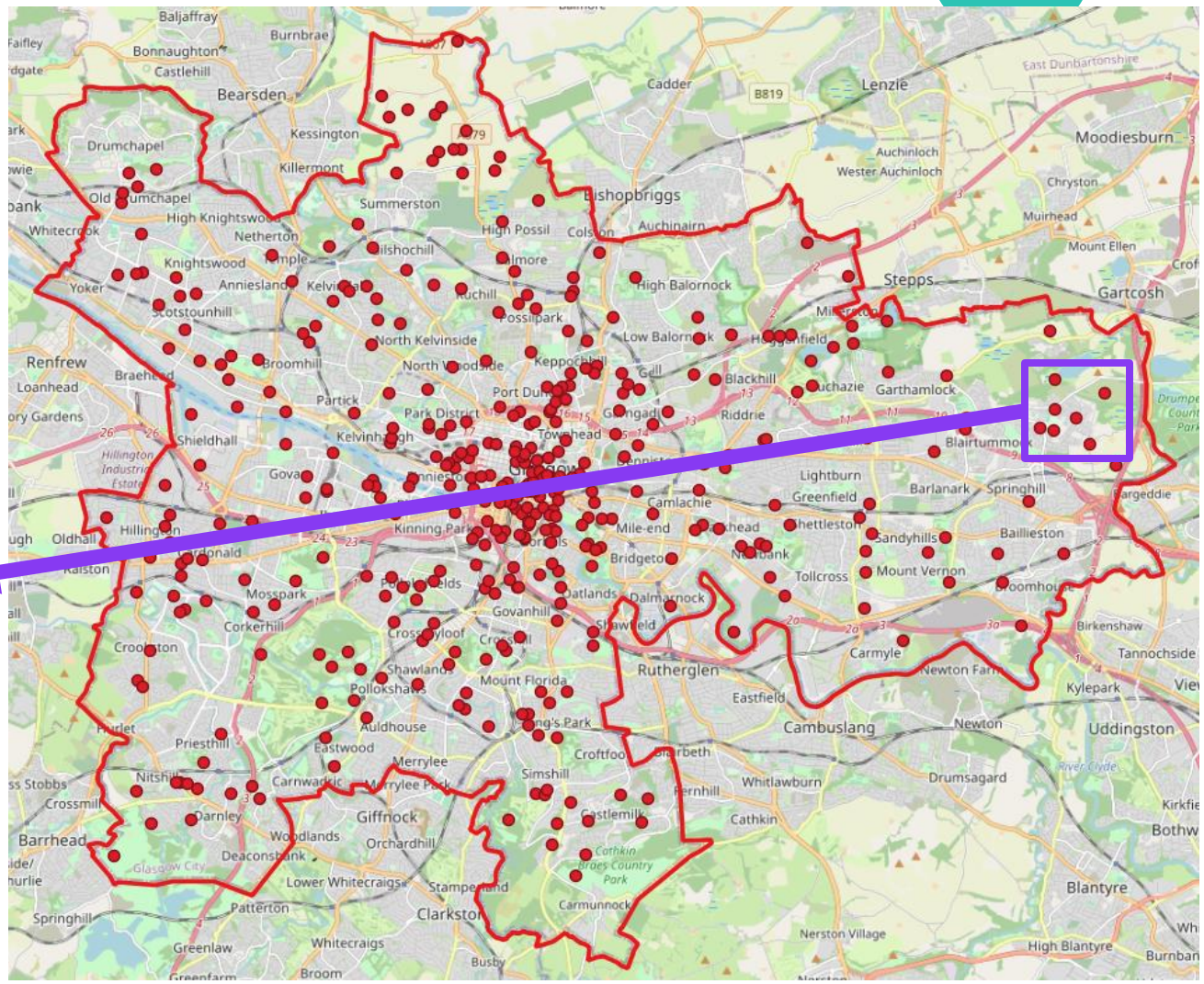
	A	B
1	string	weight
2	Pret A Manger	234
3	Subway	169
4	McDonald's	167
5	KFC	112
6	Nando's	108
7	Greggs	106
8	Domino's	95
9	PizzaExpress	93
10	itsu	58

- 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

RADIAN Example



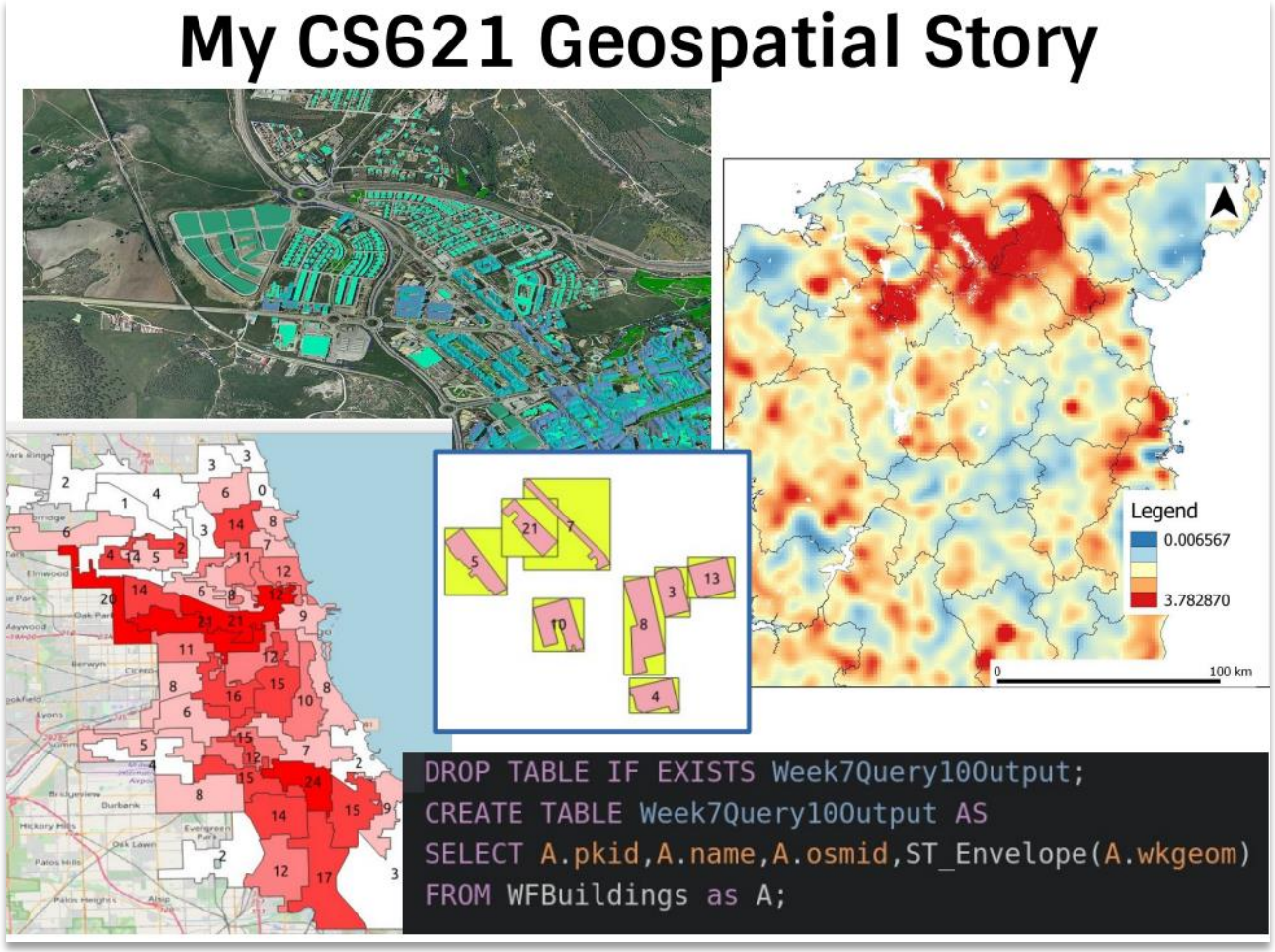
An example of some metadata attached to the points



A synthetic dataset of 400 Fast-Food restaurants in Glasgow

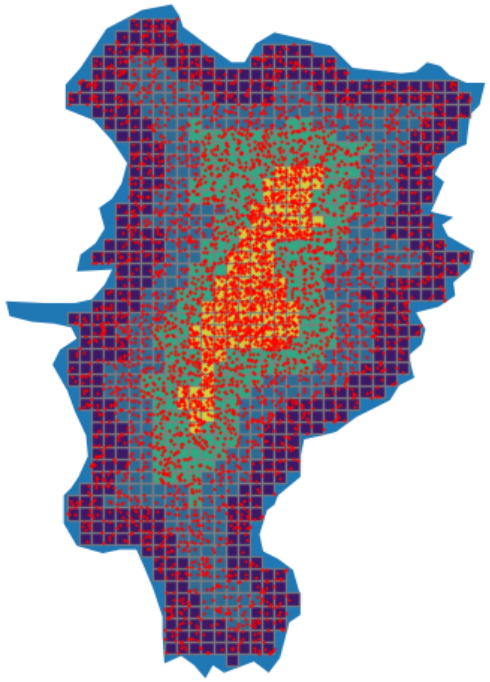
Use cases for RADIAN

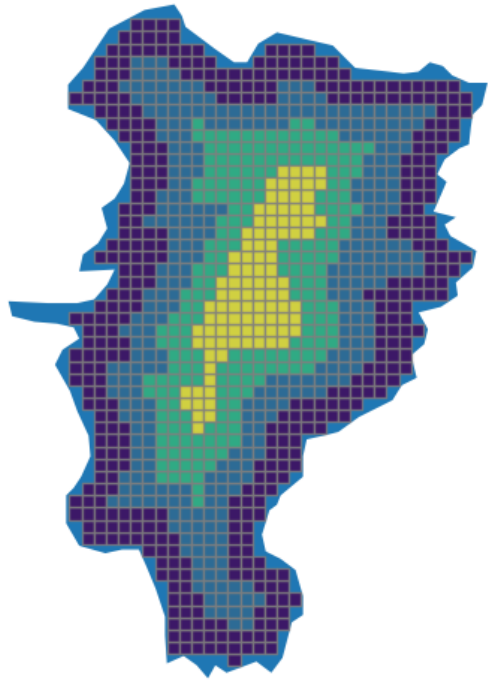
- Generate unseen data for ***CS621: Spatial Databases***
- Allow students to generate their own data for **assignments and projects**
- Used to generate examples for teaching materials to **increase engagement**
- **ChatGPT** has shown the utility of being able to generate unseen synthetic data quickly



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

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

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

Hakan Ferhatosmanoglu
University of Warwick
Coventry, UK
hakan.f@warwick.ac.uk

radiaπ

Centre for Research Training

Science Foundation Ireland SFI For what's next

GISRUK 2023 Paper

Questions & Queries:

patrick.gorry.2015@mumail.ie

Thank you!

HOST INSTITUTIONS

UNIVERSITY OF LIMERICK OLLSCOIL LUIMNIGH

Maynooth University National University of Ireland Maynooth

UCD DUBLIN

FUNDED BY

Science Foundation Ireland SFI For what's next

Skillnet IRELAND