Project 2: Technical Write Up

Introduction:

Point data that's been passively georeferenced using place names introduces ambiguity in the true location of its origin. A tweet tagged "Manchester" may be automatically geocoded to the geometric centroid Manchester's administrative boundary, rather than referenced to its true location. This approach generalises and collapses point data within a given area to a single point within, which creates false hotspots that distort patterns of distribution (Figure 1). Hence, these datasets are spatially ambiguous, as a points true locations are unknown, and visualisation can be misleading. An implementation of Huck et al.'s (2015) weighted redistribution algorithm redistributes this tweet data to create a map that could be more representative of spatial patterns, and highlight areas likely interested in the royal wedding to target advertising.

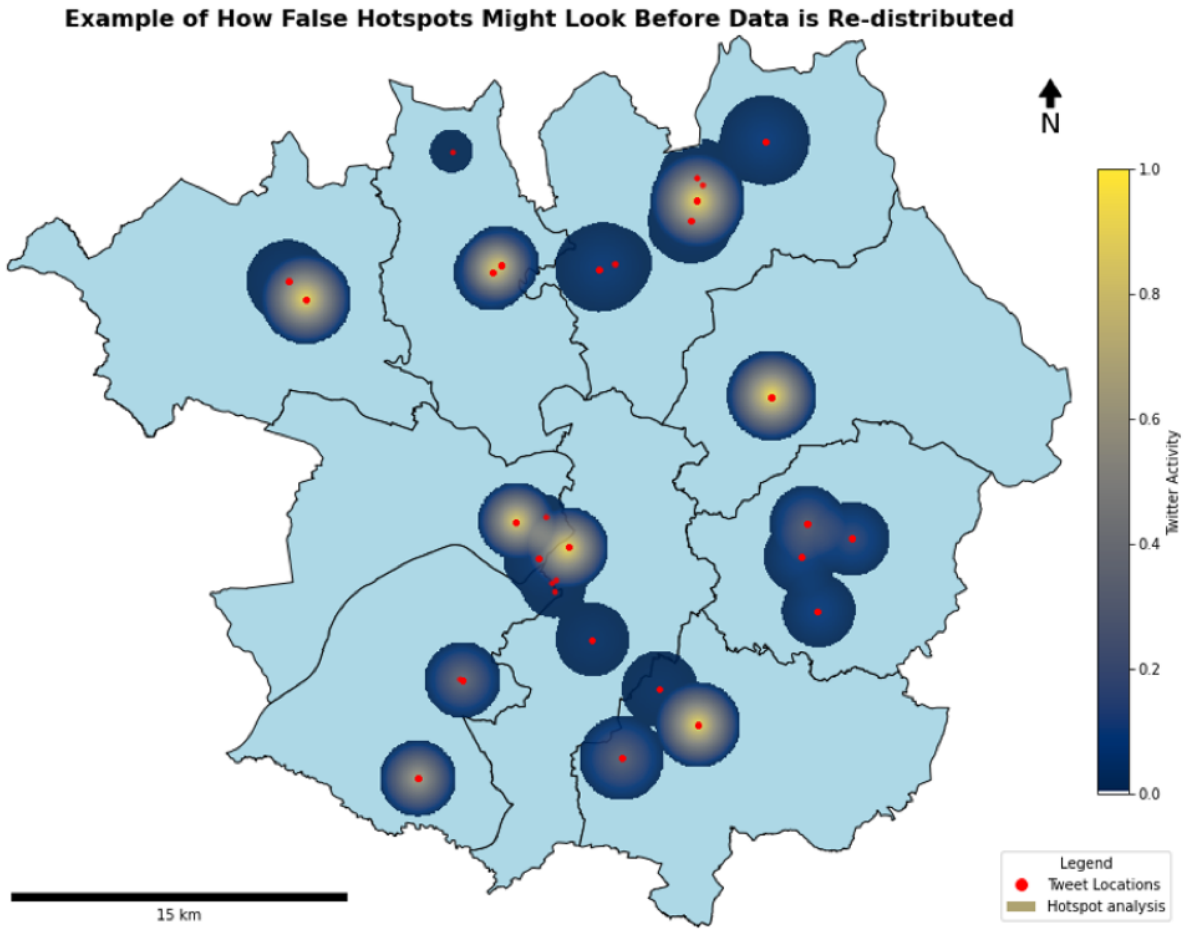

Figure 1: Example of what False Hotspots might look like when plotting spatially ambiguous data

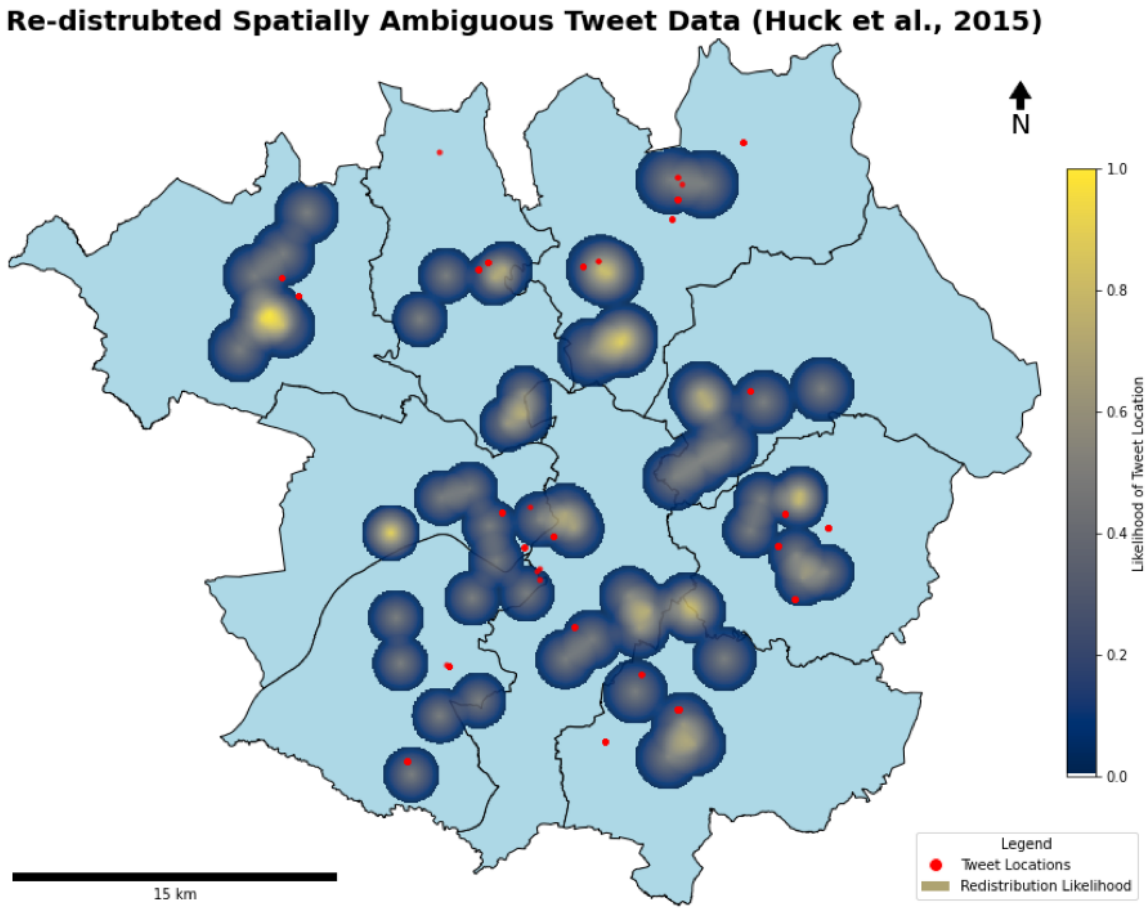

Figure 2: Likelihood distribution of where tweets about the royal wedding might originate from, or 'twitter activity'

Methodology:

Core algorithm:

For each administrative area, the helper function filter_points() retrieves tweet point data using GeoPandas' sindex, which leverages an R-tree to narrow down a subset of points intersecting the area's bounding box in O(log n) time. A precise within() check then confirms points inside the geometry. The distribution radius, derived from Huck et al. (2015), is calculated once per admin area for efficiency, as it remains constant for all points in the area. An inner loop iterates through the subset of points, generating seeds and updating the output raster which is returned and visualised by visualise_results()

Generate seed():

The generate_seed() function efficiently batch generates an array of random x, y coordinates within the bounding. It adjusts the number of coordinates generated by subtracting viable seed count from the sample, avoiding redundancy by excess generation.

List comprehension using a contains() check, filters seeds within the admin geometry. NumPy arrays use vectorized operations to efficiently evaluate whole arrays simultaneously, a scalable solution. The process repeats until enough valid seeds are found, breaking the while loop. The get_weight() function retrieves seed values from the weighting surface, and a lambda function identifies the greatest weight seed from the list, streamlining implementation.

Calculate_distribution():

The calculate_distribution() function operates in image space for consistency and faster calculations. A lambda function defines a reusable Euclidean distance calculation, and a column stack of skimage.draw's disk() function efficiently identifies cells around the seed for iteration, avoiding conditional checks to skip cells outside the radius within a bounding rectangle. Each cell is assigned a normalized value from the Euclidean decay equation (Huck et al., 2015). A try-except block handles out-of-bounds errors, avoiding inefficient conditional checks for each cell and only requiring additional memory when necessary.

CRS choice:

Data is projected to the British National Grid (BNG) ESPG, which has suitable precision for Greater Manchester and doesn't require re-projection of the weighting layer.

Analysis:

Generalisation:

The representation of Greater Manchester is at a coarse scale, with L4 boundaries, such as villages, aggregated into single boroughs. Additionally, Wigan isn't included.

Raster data weight is uniform across a cell, limiting its ability to represent fine-scale changes. The 100m resolution of the weighting surface smooths variations, potentially obscuring detailed patterns in population density or points of interest that finer resolutions might reveal.

However, this is suitable as the task is to identify broader areas for advertisements. The algorithm relies on variation in the weighting layer, and 100m resolution strikes a balance by capturing meaningful variation while avoiding excessive noise that could obscure useful street and neighbourhood level trends.

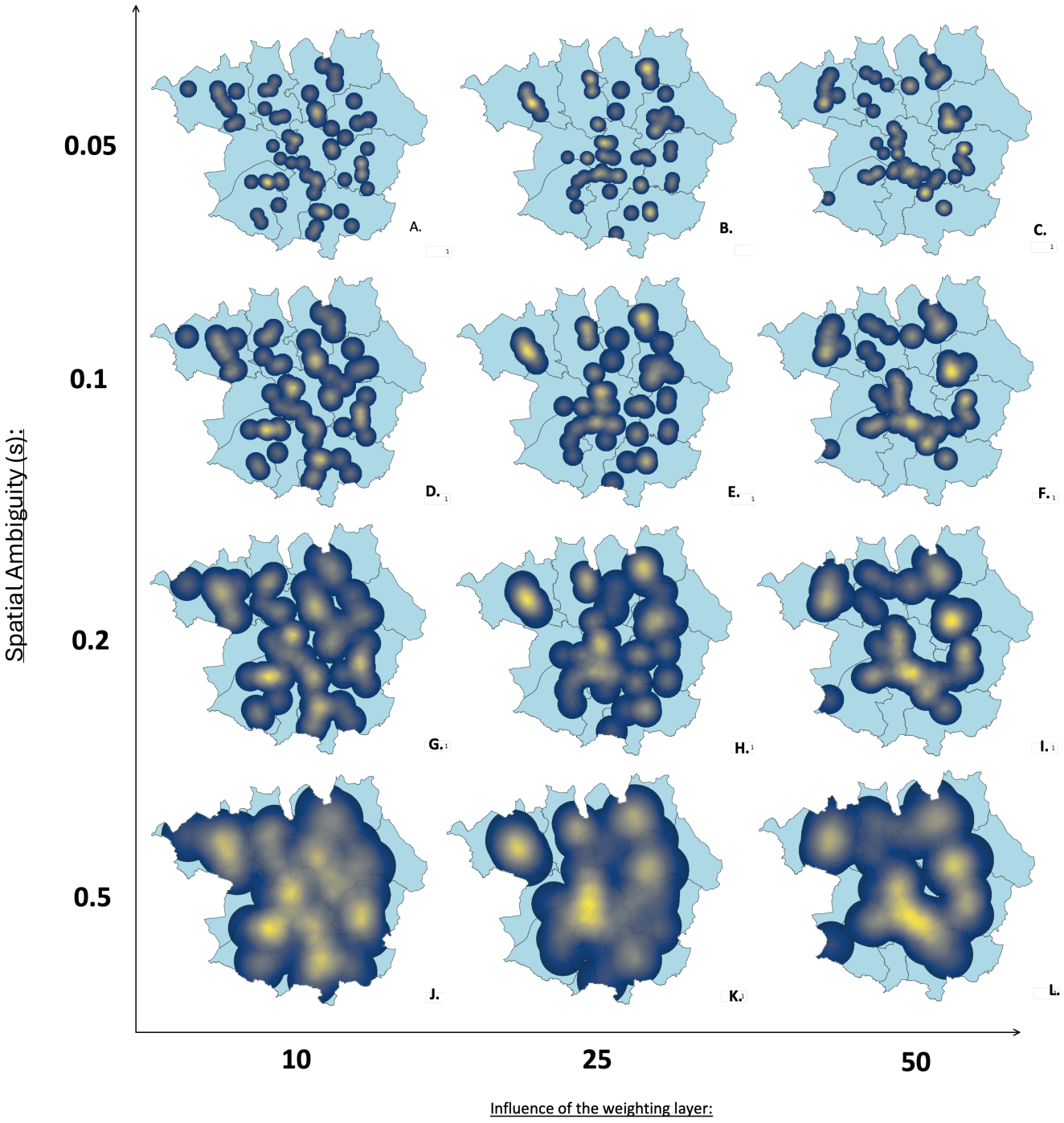
Spatial Ambiguity (s):
0.05
0.1
0.2
0.5
A.
B.
C.
D.
E.
F.
G.
H.
I.
J.
K.
L.
10
25
50
Influence of the weighting layer:

Figure 3:

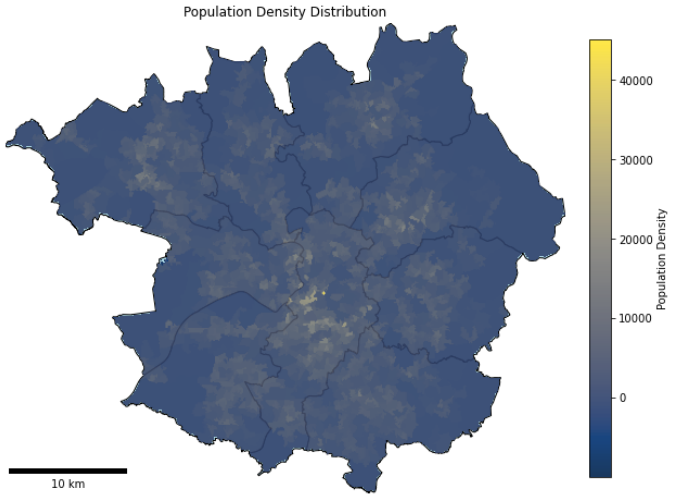

Figure 4: Weighting Layer

Value Selection:

Spatial ambiguity (*s*) represents the level of uncertainty about the precise location of a data point within its administrative area. Higher s values reflect a greater spatial ambiguity (0.5-1.0), visualised as broader likelihood distributions which might be suitable for local radio advertising, that is unsure of where interested listeners might be. Lower S values (0.1, 0.2) produce smaller, localised distributions that are useful for precise advertising like billboards but requires greater confidence in the re-distribution.

Likelihood values do not provide absolute predictions of tweet locations and are unsuitable for detailed quantitative analysis. Instead, they represent relative likelihoods, enabling basic comparisons between areas where higher values might suggest greater interest that influenced by subjective parameters. They are best suited for qualitative analysis and require consideration within broader context to assess the significance and value of clusters and distribution results.

Increasing the sample number (**w**) increases the likelihood of seeds being generated in - and selecting - high-weight locations, aligning redistribution pattern closer to the weighting layer resulting in greater clustering (figure 3). This places more confidence in the weighting layer's ability to determine tweet location. An advantage of the algorithm is redistributing points within administrative areas, responding to local ambiguity rather than redistributing points across Greater Manchester, which could lead to a false hotspot at the dense city center.

Population density is relevant context for weighting, as tweets are more likely to originate from densely populated areas. However, the goal is to target advertising in areas interested in the royal wedding, not solely high-density locations. While population density overlaps with location likelihood, older demographics may be more inclined to watch the wedding (Statista, 2024) and live in less dense suburban neighbourhoods. Including a demographic weighting layer could better reflect origin likelihood.

I selected an ‘s’ value of 0.3, and sample number of 25 to produce results that are granular enough for highly targeted advertisement but acknowledges a balance between stochasticity and weighting influence to still target audiences in lower density areas.

Limitations:

The tweet dataset's small size (1023 tweets) means each likelihood distribution carries significant weight, making outputs highly sensitive to stochastic variations. Hence, results are consistently inconsistent, relying on hand selection of maps based on subjective interpretation that introduces conformation bias in determining what might be useful and representative of interest distribution.

The algorithm faces limitations due to discrepancies between the shapefile's eight Greater Manchester districts and the dataset's 18 Level 3 administrative areas, causing issues in scale of generalisation.

Not all data points are collapsed to the same centroid of their labelled administrative boundary. Some L3 areas label multiple unique points, suggesting different go-referencing methods, and reflective of the different scales of generalisation between the GM dataset, and 18 admin areas referenced in the point data

```
# Group by 'l3' and count unique values in 'geometry' for each group
unique_locations_per_l3 = point_data.groupby('l3')['geometry'].nunique()

# Print the result
print(unique_locations_per_l3)
```

```
l3
Altrincham            1
Ashton under Lyne     1
Ashton-under-Lyne     1
Bolton                2
Bury                  3
Cheadle               2
Dukinfield            1
Heywood               2
Hyde                  1
Littleborough         1
Manchester            5
Manchester City       1
Oldham                1
Rochdale              4
Sale                  3
Salford               3
Stalybridge           2
Stockport             3
```

For example, tweet data georeferenced and collapsed within a smaller area like Cheadle may be redistributed across the larger administrative polygon Stockport. This can result in tweets being redistributed far outside their georeferenced boundaries,

which introduces imprecision and misrepresentation. Redistribution should occur within the administrative area a point was georeferenced to.

Citations:

Published by D. Clark. (2024) *Support for the Monarchy Britain 2024, by age*, *Statista*. Available at: https://www.statista.com/statistics/863893/support-for-the-monarchy-in-britain-by-age/ (Accessed: 19 December 2024).

Huck, J., Whyatt, D. and Coulton, P. (2015) 'Visualizing patterns in spatially ambiguous point data', *Journal of Spatial Information Science* [Preprint], (10). doi:10.5311/josis.2015.10.211.