9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Learning urban areas from tourist data: a case study with spatially constrained clustering and Airbnb data

Eddie Rossi[1], Marco Agnolon[1], Bruno Zamengo[1], Francesco Silvestri[2]

1. Motion Analytica
2. University of Padova, *francesco.silvestri@unipd.it*

Keywords: clustering, tourism, data analytics, spatial partitioning, Airbnb dataset

Extended Abstract

Cities are usually partitioned into smaller areas: for instance, Copenhagen is split into 10 *bydele* (Indre By, Nørrebro, ...) and Venice into 6 *sestieri* (Cannaregio, San Marco, ...). These partitions have emerged for several reasons, like history (city expansion, defensive walls, ...), geography (canals, islands, ...), and administrative (postal service, merging of smaller towns, ...). Usually, different partition schemes might coexist in the same city; they might significantly differ and the choice is justified by a particular application: for instance, in the case of postal codes or census units. However, common organizations might not be relevant, or even misleading, for some urban analytics like tourism, environmental policy, and economics.

In this abstract, we provide a data-driven method to partition a city into areas according to criteria relevant to tourism analysis. More specifically, we use spatially constrained clustering on tourist data to partition a city: each area encloses a spatially contiguous territory with similar proprieties. Our approach proposes an initial partition of the city into hexagons using Uber's hexagonal hierarchical spatial index *H3* [1]. Then, we characterize each hexagon with a vector whose values are extracted from a suitable dataset relevant to tourism analysis. Finally, we apply agglomerative clustering to create areas with similar proprieties. As a case study, we use the Airbnb datasets provided by *Inside Airbnb*, a project that studies Airbnb's impact on residential communities [1]. The dataset contains information on Airbnb accommodations (e.g., private or shared rooms, apartments, houses) from several cities; for each accommodation, the dataset provides position, average price per bed, average review rate, availability during the year, and other information. For the presentation, we focus on Copenhagen and Venice[1]. The importance of big data in the analysis of cities has been highlighted in several works (e.g., [4, 3]). Recently, a data-oriented approach has been used to define the signature of a city [2]: this approach uses an initial partitioning into minimum units that capture building footprints and physical barriers (e.g., streets, railways, and water bodies) which are then aggregated with k-means; in comparison, our approach doesn't require an initial knowledge of the city structure and focuses on obtaining spatially constrained clusters.

Method For each city, we partition the entire space into minimum geographical units with the H3 index. We use the grid at level 8 for Copenhagen (resp., level 9 for Venice): each unit is then a hexagon with an apothem length of 461m (resp., 174m). Hexagons with less than 5 accomodations are removed. We associate to each hexagon a vector in $\mathbb{R}^d$, for a suitable value $d \geq 1$. Specifically, we consider two different vector definitions:

[1]For Copenhagen, we include the municipality of Frederiksberg, which is geographically contained within Copenhagen municipality, but are completely separated from an administrative point of view. For Venice, we focus on the main touristic part and do not include Mestre (the mainland part), Giudecca and other minor islands.

- *Price per bed.* We consider the distribution of the average price per bed of one night in the hexagon: we split the prices into 5 buckets and the $d = 5$-dimensional vector represents the percentage of the accommodations with a price per bed within the bucket range. Ranges are based on the 5-quantiles computed over the cities; thresholds are $0, 24.5, 375.0, 528.0, 740.0, 986.0, +\infty$ for Copenhagen, and $0, 43.4, 61.0, 84.5, 122.0, +\infty$ for Venice). We remove accommodations without price per bed. The final number of hexagons is 147 per Copenhagen and 57 for Venice.
- *Rate.* We use similar histograms based on 5-quantiles. The rate is in the range $[1,5]$, although highly concentrated in $[4,5]$. The thersholds are $1, 4.57, 4.71, 4.8, 4.91, 5$ for Copenhagen and $1, 4.5, 4.67, 4.76, 4.85, 5$ for Venice. We remove accommodations with less than 3 reviews. The final number of hexagons is 139 per Copenhagen and 56 for Venice.

Then we construct a graph representing hexagon connectivity: we construct an undirected and weighted graph $G = (V,E)$, where each node in V represents a hexagon; there exists an edge between two nodes if and only if the respective hexagons share an edge and, the weight is given by the Wasserstein distance between the respective vectors. By using Floyd-Warshall, we then construct the *compound dissimilarity matrix D* that gives the cost of the shortest path between any two pairs in V and it better captures the similarity among hexagons.

Finally, we create the areas of the city by running an aggregate clustering (with average linkage) of the hexagons using the compound dissimilarity matrix. We observe that using the compound dissimilarity matrix (i.e., shortest paths) provides a better indicator of the distance between two hexagons, rather than edge weights since it considers the entire graph structure. Although the average linkage doesn't theoretically guarantee all hexagons in a cluster to be connected, we haven't observed this case in our experiments.

Results Figure 1 shows the results of our method using the price per bed for partitioning Copenhagen. The number of areas is the one maximizing the silhouette coefficient. The colors represent the average price per bed in that area, where dark red is the most expensive. For comparison, we also provide the map of Copenhagen districts: we note that the most central and popular parts (like Indre By, Nørrebro) have been clustered together, while the most external zones are in different clusters (interestingly, they recall Copenhagen bydele). We notice that clustering using the vector of price distribution allows us to better capture the variety of accommodations, rather than the average/median price within the hexagon. As an example, areas 10 and 11 have close median price; similarly for areas 12 and 13. However, area 11 has more mid-price accommodations than 12, while area 13 has more expensive accommodations than 12 (see Figure 2). We report the same analysis for Venice in Figure 3, where we observe that the most expensive part crosses two *sestieri* (San Marco, Dorsoduro). Finally, in Figure 4, we consider the opinion of users on accommodations. The rates are saturated in the range 4-5, with areas with the same median. For instance, areas 2, 3 and 4 have the same median, but differ in the rate distribution; in particular area 4 has more reviews with a lower rate (see Figure 5). The areas with high rates mostly consist of single hexagons containing a small number of accommodations.

References

[1] Inside Airbnb. `http://insideairbnb.com/`, 2023. [Online; accessed 10 February 2023].

[2] D. Arribas-Bel and M. Fleischmann. Spatial signatures - understanding (urban) spaces through form and function. *Habitat International*, 128:102641, 2022.

[3] M. Fleischmann, A. Feliciotti, and W. Kerr. Evolution of urban patterns: Urban morphology as an open reproducible data science. *Geographical Analysis*, 54(3):536–558, 2022.

[4] E. Glaeser, S. Kominers, M. Luca, and N. Naik. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137, 2018.

Figures

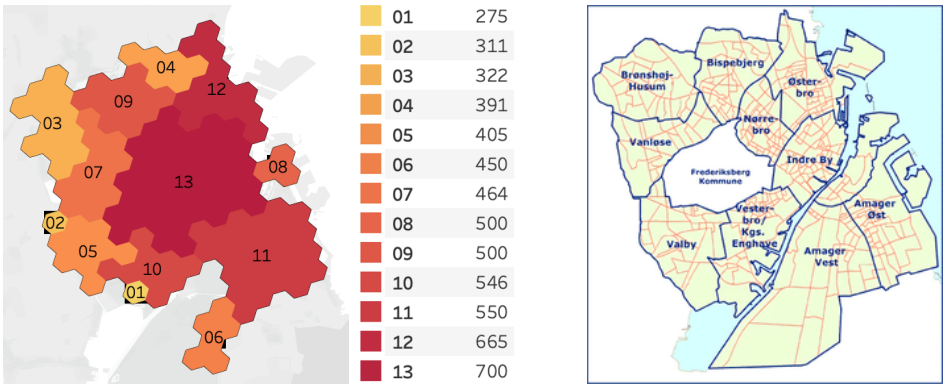

Figure 1: Results of the partitioning of Copenhagen in 13 areas using the price per bed. Colors provide the median price per bed within each area (prices in DKK). On the right, the *bydele* of Copenhagen (image from Wikipedia).

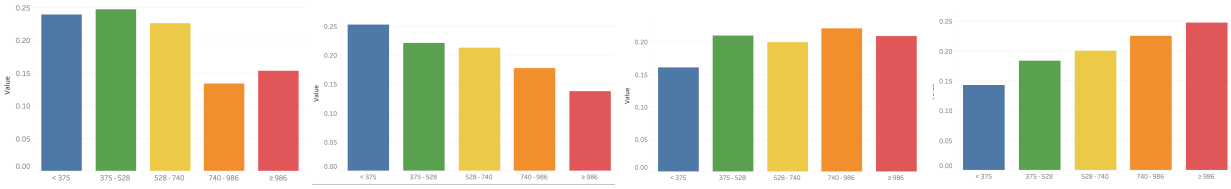

Figure 2: Differences in price/bed distribution for areas 10, 11, 12, and 13 for Copenhagen. Each bar reports the frequency in the respective quantile.

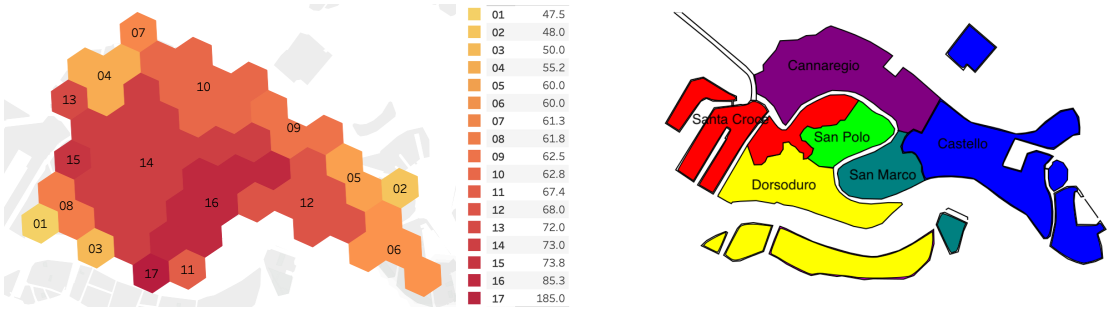

Figure 3: Results of the partitioning of Venice in 17 areas using the price per bed. Colors provide the median price per bed in each area (price in Euro). On the right, the *sestieri* of Venice (image from Wikipedia).

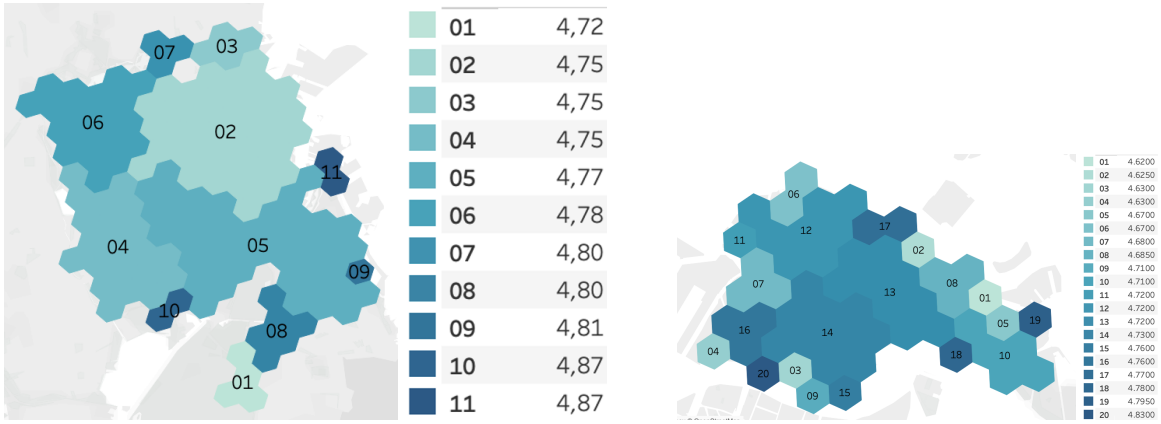

Figure 4: Results of the clustering in 23 and 9 areas of Copenhagen (left) and Venice (right) using the average rate of each accommodation. Colors provide the median rate per area (dark blue is the highest rate).

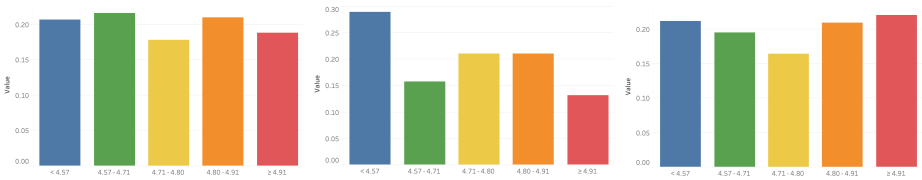

Figure 5: Differences in rate distribution for areas 2, 3, and 4 for Copenhagen. Each bar reports the frequency in the respective quantile.