AGILE: GIScience Series, 3, 61, 2022. https://doi.org/10.5194/agile-giss-3-61-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

A regionalization method filtering out small-scale spatial fluctuations

Lucas Spierenburg [1], Sander van Cranenburgh [2], and Oded Cats [1]

[1]Transport and Planning, Delft University of Technology, Delft, The Netherlands
[2]Transport and Logistics group, Department of Engineering Systems and Services, Delft University of Technology, Delft, The Netherlands

Correspondence: Lucas Spierenburg (l.j.spierenburg@tudelft.nl)

Abstract. Regionalization is the process of aggregating contiguous spatial units to form areas that are homogeneous with respect to one or a set of variables. It is useful when studying spatial phenomena or when designing region-based policies, as it allows to unravel the latent spatial structure of a dataset. However, this task is challenging when small-scale fluctuations in the data interfere with the phenomenon of interest. In such circumstances, regionalization techniques are prone to overfitting small-scale fluctuations, and producing erratic regions. This paper presents a regionalization method robust to small-scale variations that is particularly relevant when handling demographic data. Fluctuations are filtered out using a weighted spatial average before applying agglomerative clustering. The method is tested against a conventional agglomerative clustering approach on a fine-resolution demographic dataset, for a set of indicators quantifying: the ability to identify large-scale spatial patterns, the homogeneity of the regions produced, and the spatial regularity of these regions. These indicators have been computed for the two methods for a number of clusters ranging from 2 to 101, and results show that the proposed approach performs better than conventional agglomerative clustering more than 90% of the time at identifying large-scale patterns, and produces more regular regions 96% of the time.

Keywords. Regionalization, agglomerative clustering, regional science

1 Introduction

1.1 Background and scientific contribution

Regionalization is the process of aggregating contiguous spatial units to form areas that are homogeneous with respect to one or several variables. It is extensively used to identify a spatial structure in a dataset. For instance Östh et al. (2015) have determined the geographical extent of spatial segregation in cities; Ramachandra Rao and Srinivas (2008) have identified areas being vulnerable to flooding; and Sobolevsky et al. (2013) have delineated frontiers between areas with limited interactions using data on phone communication. Regionalization is also used as a pre-processing step when studying a spatial phenomenon to mitigate potential biases in the raw data. For instance, Spielman and Folch (2015) reduce uncertainty in a demographic dataset generated from a survey using regionalization, and Nakaya (2000) use it to address the modifiable areal unit problem resulting from an arbitrary zoning for mortality data.

Regionalization methods assess the variation between spatial units to delineate homogeneous regions (Duque et al., 2007). However, they are oblivious to the spatial scale of these variations, only the variations' magnitude is considered. This is an issue for the analyst when their objective is to identify a large-scale pattern in data disturbed by substantial small-scale interferences, as conventional regionalization methods would overfit small-scale fluctuations. This would lead to irregular regions with chaotic borders, which are impractical for region-based policymaking. This issue is particularly paramount when dealing with demographic data (Wolf et al., 2021). Such data are often available at a fine spatial resolution and are sometimes subject to inaccuracies, especially when generated from surveys. These two aspects usually introduce fluctuations in the data at a small spatial scale that do not in fact reflect a sharp demographic change. When delineating areas in which demographics are homogeneous – e.g. to study spatial segregation –, small-scale fluctuations should not disturb the regionalization process.

This paper presents the filtered-input agglomerative clustering, a regionalization method that is robust to small-scale spatial fluctuations, particularly suitable for demographic data. This method involves two steps. First, it filters out small-scale fluctuations in the data of interest, us-

ing a weighted average on the spatial units that is representative of their mutual proximity. Then, it applies an agglomerative clustering technique. We demonstrate the method by applying it to delineate neighborhoods with homogeneous demographics in a city, and comparing its output with the one from a conventional agglomerative clustering.

1.2 Case study

This work takes the city of Leiden (the Netherlands) as a case study. The objective is to draw neighborhoods that would be both homogeneous in terms of demographics, and spatially regular. The variables of interest are the shares of: people from Dutch descent, people with a non-western migration background, and people with a western migration background. As these three variables sum up to 1, only the first two are considered in the analysis, they express the full information contained in the data.

To ensure data anonymity, the data provider has rounded the share of groups in each zone to the closest 10%, and does not disclose data in spatial units where less than 5 inhabitants live. These two operations generate substantial small-scale fluctuations that are purely noise, which disturb a conventional agglomerative clustering method. This dataset is therefore relevant to test our proposed method.

2 Method

This work proposes a regionalization approach designed to capture large-scale spatial patterns, based on agglomerative clustering. We call this method filtered-input agglomerative clustering. An agglomerative clustering technique aggregates spatial units together into regions, where the objective is often to minimize the regions' variance for a given set of data. Small-scale fluctuations disturb the identification of large-scale patterns, as they steer the delineation of regions to fit local variations. We propose to filter them out using a weighted moving average at the input of the agglomerative clustering, where the weights depends on the walking time between spatial units, representing their mutual proximity. This approach is particularly relevant to delineate regions based on demographic data – e.g. to study spatial segregation – as individuals are not bound to their home location may move to nearby spatial units throughout the day. Subsection 2.1 describes the agglomerative clustering used, and subsection 2.2 summarizes the filtering step implemented at the input.

2.1 Agglomerative clustering

Clustering consists in partitioning objects into sets, called clusters, that are meaningful according to a certain criterion, e.g. homogeneity. In agglomerative clustering, all objects are considered as individual clusters in the initialization phase. Clusters are then merged iteratively, optimizing an objective function at each step. In this work, we merge spatial units together such that the within-clusters variance is minimized, using Ward's dissimilarity (Ward Jr., 1963). The cluster analysis is spatially constrained; merging operations leading to discontinuous clusters are banned using a connectivity matrix. In this matrix M, element m_{ij} is 1 if spatial unit i is adjacent to j, 0 otherwise. Figure 1 illustrates the agglomerative process in a city composed of 5 zones. In this example, spatial units *A* and *E* would form the most homogeneous cluster. However, they are not adjacent, as the connectivity matrix indicates. Therefore, units *A* and *B* are merged instead. The resulting increase in within-cluster variance is depicted in the dendrogram by the dissimilarity. This dendrogram summarizes the successive merging of clusters, and the corresponding increase in within-clusters variance.

2.2 Filtering step

To better capture large-scale spatial patterns in the data, we propose a method to filter out small-scale fluctuations using a weighted moving average (see equation 1). In this equation, $\overline{x_i}$ is the spatial average of variable x in spatial unit i, computed from the value of x in every spatial units j located less than 20-minutes away on foot. Variables x_j are weighted depending on the walking time t_{ij} between the centroids of spatial units i and j. The function $w(t_{ij})$ is called the travel impedance, it decreases when the walking time increases. The further away is a spatial unit, the lower is its impact on its neighbors. In the case study, since the variables of interest are the share of people from each social group in a spatial unit, these variables are also weighted based on the total population n_j in every spatial unit j. Therefore, the more populated is a spatial unit, the larger is its impact on its neighbors.

$$\overline{x_i} = \frac{\sum_j w(t_{ij}) \cdot n_j \cdot x_j}{\sum_j w(t_{ij}) \cdot n_j} \tag{1}$$

The travel impedance $w(t_{ij})$ in equation 2 models the extent to which individuals from unit j visit unit i. It is defined using the visitation law proposed by Schläpfer et al. (2021), in which the attractiveness of unit i is assumed to be constant (3600), as no data available would allow to estimate it. This constant is set to have a weighting coefficient of 1 when the walking time is 60 seconds. It does not affect the weighting average since it is present in both the numerator and denominator in equation 1. The walking time t_{ij} from the centroid of unit i to the one of unit j is computed using the street layout and a walking speed of 4.5 km/h. The centroid of each spatial unit is placed at the geometric center of the building footprint.

$$w(t_{ij}) = \begin{cases} 1 & \text{if } 0 \leq t_{ij}[s] < 60 \\ \frac{3600}{t_{ij}^2} & \text{if } 60 \leq t_{ij} < 1200 \\ 0 & \text{if } t_{ij} \geq 1200 \end{cases} \tag{2}$$

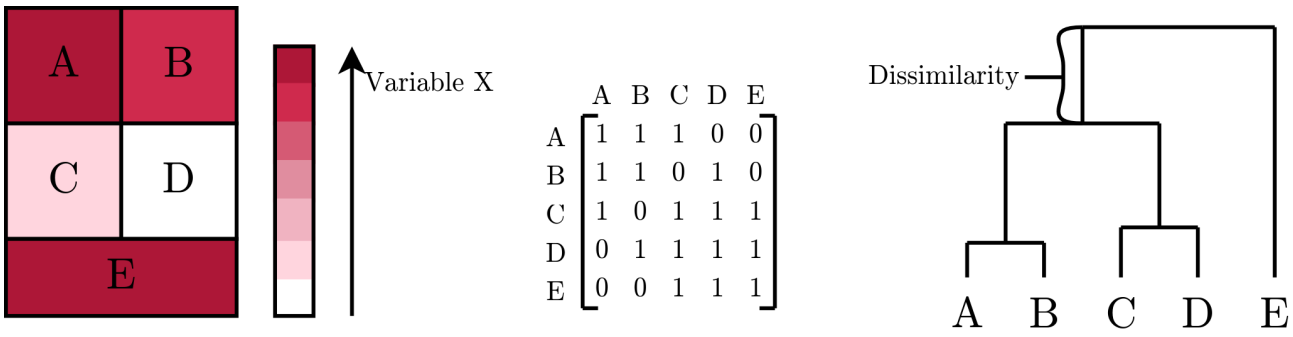

Figure 1. An example of a dendrogram (right) obtained after applying agglomerative clustering on a city composed of 5 zones (left) on a single variable X, using the connectivity matrix corresponding to the city's topology (middle).

Finally, before applying agglomerative clustering on the filtered data, we standardize the variables of interest.

2.3 Data and Software Availability

2.3.1 Data

The data used in this work are: demographic data, street data, and data on the building footprint. Open demographic data are collected from Netherlands Statistics, CBS (2017), for year 2017. The dataset provides the number of inhabitants, and the proportion of: people from Dutch descent, people with a western migration background, and people with a non-western migration background (more info at (CBS, 2017)), at the postcode level (units of around $100 \times 100\ m^2$).

The street layout and the building footprint are extracted from OpenStreetMap, using the python library osmnx for the streets and the QuickOSM plugin for the buildings (Boeing, 2017; Trimaille and Charzat, 2022).

2.3.2 Hardware and software

The results were produced using a laptop with an Intel® Core™ i5-10210U CPU, 32GiB of RAM, and an Intel® UHD Graphics GPU. The operating system is Ubuntu 21.10, 64-bit. The computation of the results were performed in less than an hour.

This work uses Python 3.9.7 for the computation of the results with the following libraries: osmnx, pandas, geopandas scipy, sklearn, networkx. QGIS is used to visualize the results, and to extract the building layout.

The computational workflow supporting this publication is executed via a set of 5 scripts published under license CC-BY-4.0 at https://doi.org/10.17605/OSF.IO/JM96F.

3 Results

In this section, we compare the performance of our filtered-input agglomerative clustering with a conventional agglomerative clustering. The three criteria evaluated are: the ability to identify large-scale patterns; the homogeneity of the regions produced, and their geometrical regularity. The geometrical irregularity of a region is defined here by the extent to which the region's borders are erratic, and the extent to which it is spatially spread. The two methods are evaluated in this paper with a number of 8 clusters. The results presented here hold for the other values tested ranging from 2 to 101. To assess the ability to detect large-scale spatial patterns, we compare the two methods' outputs with a conventional agglomerative clustering applied on the 5-digits postcodes (instead of the 6-digits ones), corresponding to a coarser resolution. At this resolution, the data available are the same and fluctuate less, allowing to better identify large-scale patterns. These data come from an independent dataset from the data provider, they have not been created by the authors from the aggregation of the 6-digits postcodes data.

3.1 Qualitative analysis

The performance of the filtered-input agglomerative clustering is first assessed qualitatively, analyzing the shape of the regions produced on a map. Figure 2 depicts the delineation of homogeneous neighborhoods provided by the conventional agglomerative clustering on the 6-digits postcodes (map A), the 5-digits postcodes (map C), and the filtered-input agglomerative clustering on the 6-digits postcodes (map B). In these maps, the postcodes' color represent their demographic trends. One can see that data for the 6-digits postcodes (A and B) fluctuate more than for the 5-digits postcodes (C).

The conventional and the filtered-input agglomerative clustering identify the same regions, but the delineation is different (see A and B in Figure 2). First, region 1 in

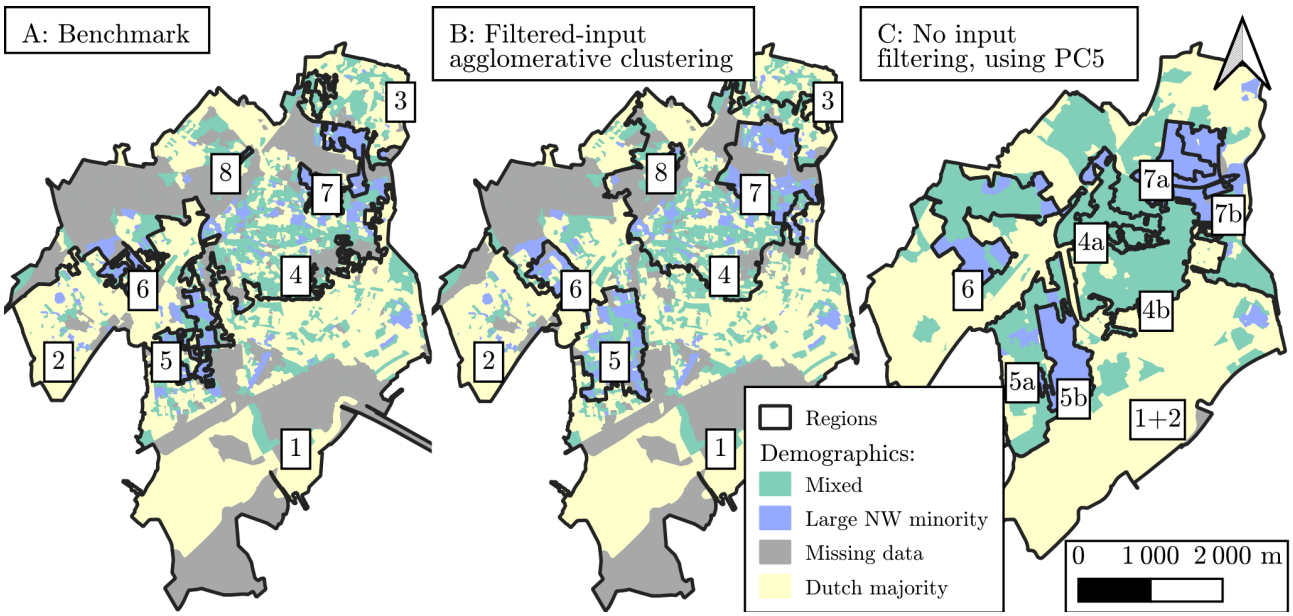

Figure 2. Spatial aggregation of postcodes into 8 homogeneous areas in Leiden, based on demographic data. The postcodes are colored as follows: yellow for postcodes in which the proportion of individuals from Dutch descent exceeds the city's average (70%), blue if the proportion of individuals with a non-western migration background exceeds 30% (the city's average is 16%), and green otherwise. The black contours delineate the regions' borders drawn by the different methods tested. In map A, we cluster the 6-digits postcodes using the benchmark method. In map B, we cluster the 6-digits postcodes using the filtered-input agglomerative clustering. In map C, we cluster the 5-digits postcodes using conventional agglomerative clustering.

the South, region 2 in the West, and region 3 in the North are concentrating individuals from Dutch descent. Second, region 4 in the center is mixed. Third, regions 5, 6, 7 and 8 are regions in which the non-western minority is segregated. Overall, the regions delineated by the filtered-input agglomerative clustering are less homogeneous, but more spatially regular (see region 7).

The agglomerative clustering at the 5-digits postcodes level (map C) identifies regions that are similar to the other two methods, but the overlap is not perfect: some regions identified by the two previous approaches are either merged (regions 1 and 2), or split into sub-regions (regions 4, 7 and 5). From this figure, we cannot clearly evaluate which output of the conventional or the filtered-input agglomerative clustering overlaps the most with the regions of the third map.

To conclude, the filtered-input agglomerative clustering delineate more regular regions, at a cost of fitting less the raw data, and the three maps does not allow to determine which of the two methods performs the best in identifying large-scale spatial pattern.

3.2 Quantitative analysis

This subsection compares the conventional and the filtered-input agglomerative clustering quantitatively using indicators on: the ability to detect large-scale spatial pattern, the homogeneity of the regions produced, and their geometrical regularity. The first two columns of table 1 display the indicators considered when the number of clusters is set to 8. The authors have also computed the indicators for a number of clusters ranging from 2 to 101. Last column in table 1 indicates the number of times where the filtered-input agglomerative clustering outperformed the benchmark, among the 100 values tested for the number of clusters.

The method's ability to identify large-scale spatial pattern is measured from the overlap between its output and the one of the agglomerative clustering performed on the 5-digits postcodes (map C in figure 2). To that end, we compute the adjusted Rand, the Fowlkes-Mallow, and the adjusted mutual information indexes. These indicators are the most widely used to compare a clustering method's output to a set of clusters considered as the ground-truth. These indicators are upper-bounded to 1, corresponding to a perfect overlap between two sets of clusters. When the number of clusters is set to 8, the filtered-input agglomerative clustering successfully outperforms the benchmark according to the adjusted Rand and to the adjusted mutual information indexes, but not to the Fowlkes-Mallows index. However, the filtered-input agglomerative clustering performs better than the benchmark in all three indexes more than 90 % of the time.

The homogeneity of the clusters is evaluated from the ratio between the between-clusters sum of squares (BCSS, see equation 4), and the total sum of squares (TSS, see equation 3), on the raw data. In these equations, X_{ki} and Y_{ki} are two variables of interest, where i is one of the n_k objects in cluster k, and $\overline{X_k}$ represent the average of X in cluster k. The larger is the ratio $BCSS/TSS$, the more varia-

Table 1. Indicators on the similarity with the 5-digits clustering, homogeneity, and geometrical regularity of the regions produced, using the conventional and the filtered-input agglomerative clustering. The first two columns provide the indicators' value when the number of clusters is set to 8. The indicators have been compared for a number of clusters varying from 2 to 101, and the last column indicates the number of times where the approach proposed outperformed the benchmark.

Indicators		Convent.	Filtered input	Success rate
Similarity with the 5-digits clustering	Adjusted Rand index	0.314	0.341	0.98
	Fowlkes Mallows index	0.514	0.509	0.93
	Adjusted mutual information index	0.402	0.504	1
Homogeneity	BCSS/TSS	0.435	0.340	0
Geometric indicators	Total perimeter	173 km	120 km	1
	Average distance from center	1.42 km	1.35 km	0.96

tion in the data is explained by the variation between clusters. The filtered-input agglomerative clustering approach produce regions with comparable, yet lower homogeneity than the ones produced by the benchmark, regardless of the number of clusters. One can expect such a result: the agglomerative clustering minimizes the within-cluster variance (WCSS, see equations 5 and 6). The benchmark method minimizes the within-cluster variance for the raw data, while the approach proposed in this paper minimizes the within-cluster variance for the filtered data. It is therefore normal that the benchmark fits better the raw data. However, one should be aware that the better performance of the benchmark approach might hide the overfitting of small-scale fluctuations.

$$TSS = \sum_{k}^{K} \sum_{i}^{n_k} (X_{ki} - \overline{X})^2 + (Y_{ki} - \overline{Y})^2 \tag{3}$$

$$BCSS = \sum_{k}^{K} n_k \cdot \left[(\overline{X_{k.}} - \overline{X})^2 + (\overline{Y_{k.}} - \overline{Y})^2 \right] \tag{4}$$

$$WCSS = \sum_{k}^{K} \sum_{i}^{n_k} (X_{ki} - \overline{X_{k.}})^2 + (\overline{Y_{ki}} - \overline{Y_{k.}})^2 \tag{5}$$

$$TSS = BCSS + WCSS \tag{6}$$

Spatial irregularity is measured from the total perimeter of all zones, and the regions' average distance from their respective centers. Both indicators are smaller for the approach proposed in this study, meaning that the regions delineated are more spatially regular (this can also be visually inspected in figure 2). This result also applies for most of the values tested for the number of clusters.

To conclude, the filtered-input agglomerative clustering detects better large-scale patterns when compared to the benchmark, draws more regular but less homogeneous regions. However, the higher homogeneity of the regions produced by the benchmark might result from the overfitting of the small-scale fluctuations.

4 Conclusion

Identifying large-scale spatial patterns is challenging when the data present small-scale fluctuations for conventional regionalization methods. For instance, the conventional agglomerative clustering used as a benchmark in figure 2 produces regions with chaotic shapes and erratic borders. Such behavior is explained by the fitting of small-scale fluctuations. To mitigate this issue, this work proposes to filter these fluctuations using a weighted spatial average before applying agglomerative clustering. As the weighted spatial average depends on the walking time between zones, the method is suited for data involving people; e.g. demographics, election results, or criminality rate. The method proposed is particularly relevant for delineating homogeneous regions based on demographic data, as such data are often available at a fine resolution and might be subject to small-scale fluctuations. The results show that the filtered-input agglomerative clustering successfully identifies large-scale spatial pattern, and produces spatially regular regions that are practical to use by policymakers.

References

Boeing, G.: OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, Computers, Environment and Urban Systems, 65, 126–139, https://doi.org/10.1016/j.compenvurbsys.2017.05.004, 2017.

CBS: Kerncijfers per postcode, https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode, licensed under CC-BY-3.0-NL, 2017.

Duque, J. C., Ramos, R., and Suriñach, J.: Supervised regionalization methods: A survey, International Regional Science Review, 30, 195–220, https://doi.org/10.1177/0160017607301605, 2007.

Nakaya, T.: An Information Statistical Approach to the Modifiable Areal Unit Problem in Incidence Rate Maps, Environment and Planning A: Economy and Space, 32, 91–109, https://doi.org/10.1068/a31145, 2000.

Ramachandra Rao, A. and Srinivas, V.: Regionalization of Watersheds, 2008.

Schläpfer, M., Dong, L., O'Keeffe, K., Santi, P., Szell, M., Salat, H., Anklesaria, S., Vazifeh, M., Ratti, C., and West, G. B.: The universal visitation law of human mobility, Nature, 593, 522–527, https://doi.org/10.1038/s41586-021-03480-9, 2021.

Sobolevsky, S., Szell, M., Campari, R., Couronné, T., Smoreda, Z., and Ratti, C.: Delineating geographical regions with networks of human interactions in an extensive set of countries, PLoS ONE, 8, https://doi.org/10.1371/journal.pone.0081707, 2013.

Spielman, S. E. and Folch, D. C.: Reducing uncertainty in the American Community Survey through data-driven regionalization, PLoS ONE, 10, https://doi.org/10.1371/journal.pone.0115626, 2015.

Trimaille, E. and Charzat, M.: QuickOSM, a QGIS plugin, https://docs.3liz.org/, 2022.

Ward Jr., J. H.: Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58, 236–244, https://doi.org/10.1080/01621459.1963.10500845, 1963.

Wolf, L. J., Knaap, E., and Rey, S.: Geosilhouettes: Geographical measures of cluster fit, Environment and Planning B: Urban Analytics and City Science, 48, 521–539, https://doi.org/10.1177/2399808319875752, 2021.

Östh, J., Clark, W. A., and Malmberg, B.: Measuring the scale of segregation using k-nearest neighbor aggregates, Geographical Analysis, 47, 34–49, https://doi.org/10.1111/gean.12053, 2015.