A Supervised Heuristic for a Balanced Approach to Regionalization

Tyler D. Hoffman[11], Taylor Oshan[21]

[1]Department of Geography, University of Maryland College Park

April 16, 2020

Summary

Regionalization refers to the design of areal zones by spatially aggregating smaller units into larger clusters. Algorithms to conduct regionalization typically require the desired number of clusters to be specified *a priori*, though a reasonable number is not always clear. Therefore, a heuristic is proposed to endogenously determine the number of clusters in a supervised setting (i.e., model-driven) by balancing the fit of a spatial model and the average area of clusters used as input. The heuristic is applied in a spatial interaction modeling context and a workflow is presented for integrating regionalization algorithms into larger spatial analysis frameworks.

KEYWORDS: regionalization, MAUP, spatial interaction, spatial clustering, scale

1. Introduction

City planning and sustainable urban growth have become inextricably linked with digital technology and cyberinfrastructure. Multiple cities around the world are now developing digital twins in efforts to improve the management of their urban resources and analyze dynamic processes, through large scale modeling and simulation. These efforts are supported by new streams of data generated from dispersed networks of monitoring devices. Initiatives like the *Array of Things*[3] and *LinkNYC*[4] are producing torrents of individual level spatiotemporal data with unprecedented resolutions and coverage, creating new opportunities to understand and influence cities. To ensure the effectiveness of urban planning strategies, it is imperative to understand the quality of data, create computationally efficient workflows, and navigate the complexities associated with system articulation and the inherent multiscale nature of many spatial processes.

One option that can balance several of the aforementioned challenges is to aggregate smaller spatial units into clusters forming larger spatial units. Ideally, this should be done in a way that maximizes data quality and maintains tractability for real-time decision-making. There are many ways to solve the regionalization (spatial clustering) problem, yet they predominantly require the user to exogenously specify a number of clusters. This is a natural limitation of unsupervised learning, where the analyst often must rely on their instinct to inform the proper choice of the number of clusters. However, it becomes possible to build in an endogenous criterion for optimality to help select an appropriate number of clusters by adopting a model-driven, supervised approach. This work develops a heuristic for determining an optimal number of clusters using a spatial interaction (SI) model to construct the supervised criterion.

Commonly used by academics and urban planners, SI models provide a longstanding framework for analyzing dimensions of aggregate human movement between origin and destination zones. Like many spatial modeling frameworks, SI models are sensitive to the scale of the zones used as input. While higher resolution data are opening new avenues of research, their increased detail often comes at the cost of computational complexity and higher uncertainty. As a result, this work also revisits the role of

[1] thoffman@umd.edu

[2] toshan@umd.edu

[3] https://arrayofthings.github.io/

[4] https://www.link.nyc/

regionalization within spatial analytical methods given recent advances in computing and the rise of big data.

2. Methodology

SI modeling provides a conceptual and technical foundation for explaining and predicting the flow of human movement, information flows, and international trade (Fotheringham and O'Kelly, 1989; Oshan, 2016). Importantly, SI models require data at the aggregate level in order to make accurate predictions about movement patterns between places. An ideal aggregation would ideally provide a set of regions that accurately represents the communities at the heart of important policy goals, which may not necessarily lie along administrative boundaries. If these regions are too large, the analysis will be too-coarse grained for meaningful policy; while if these regions are too small, the data may be too noisy and the analysis computationally intractable. Algorithms that accomplish this task of creating aggregate clusters of spatial units fall under the umbrella of regionalization.

More specifically, regionalization is the process of spatially "aggregating areas into homogeneous regions" subject to various constraints, such as contiguity, number, and scale of aggregated regions, or determining which areas need to be analyzed (Duque *et al.*, 2012). In the context of SI models, regionalization procedures could be used to define novel functional regions—e.g., travel to work areas or migration routes—that fulfill multiple social and economic criteria, potentially diverging from traditional boundaries such as census geographies. This yields a meso-scale community-driven approach to building meaningful regions as opposed to a purely top-down administrative approach or a bottom-up data-driven approach.

For a balanced approach to regionalization, an objective is proposed that seeks to choose the number of clusters that maximizes goodness-of-fit of the model (e.g., an SI model) and minimizes the amount of regionalization being performed as this leads to more information loss. The criterion devised here minimizes a combination of the standardized root-mean-square error (SRMSE) of the SI model as a proxy for goodness-of-fit and the average area of the units resulting from regionalization:

$$k_{crit} = \underset{k\,\in[4,N]}{\arg\min}\left[\frac{1}{M_S}SRMSE(k) + \frac{1}{M_A}A(k)\right] \tag{1}$$

where $SRMSE(k)$ and $A(k)$ are the SRMSE and average unit area of a regionalization with k clusters, N is the total number of original areal units, and M_S and M_A are normalizing constants that ensure the two quantities operate on the same numerical scale and are thus comparable in the minimization process. Since the criterion minimizes the sum, it identifies the number of clusters that leads to the least amount of both quantities, rather than solely preferring one or the other. Alternatively, parameters could be introduced, allowing one quantity to be favored over the other, though this is not pursued here. One could also add more terms to the objective or introduce constraints on the problem to incentivize other qualities (like compactness) in the resulting regionalization.

3. Results

To test the heuristic, county level census data from the American Community Survey (ACS) was used for the state of California to examine the integration of regionalization and SI models. As the optimization is 1-D, discrete, and bounded, it is done by brute force in these experiments. Three different regionalization algorithms were investigated (Ward linkage hierarchical clustering, SKATER regionalization (Assunção *et al.*, 2006), and regional k-means) to ensure the results are not an artifact of one particular algorithm. Notably, results from the three algorithms agreed fairly well despite indicating slightly different optimal k_{crit}values (16, 14, and 8, respectively). **Figure 1** shows the heuristic curves for the Ward linkage clustering, SKATER clustering, and regional k-means and highlights the optimal values of these curves. The Ward linkage regionalization for California is plotted in **Figure 2**.

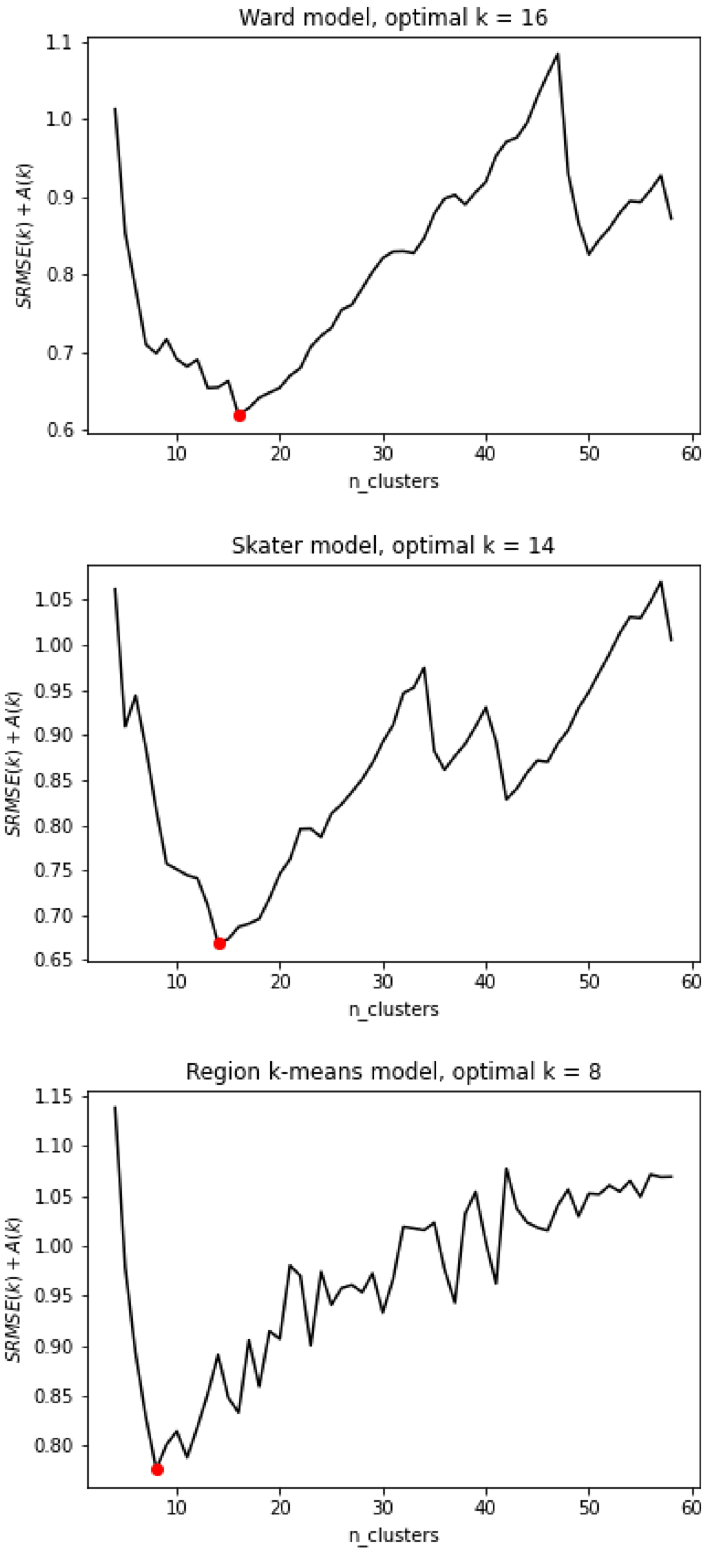

Figure 1 Example of optimal clustering selection plot for Ward linkage, SKATER, and regional k-means spatial clustering on California ACS data. The red dot is the optimal number of clusters suggested by the proposed heuristic k_{crit}. Goodness-of-fit results were averaged over 25 trials to smooth the effects of randomness.

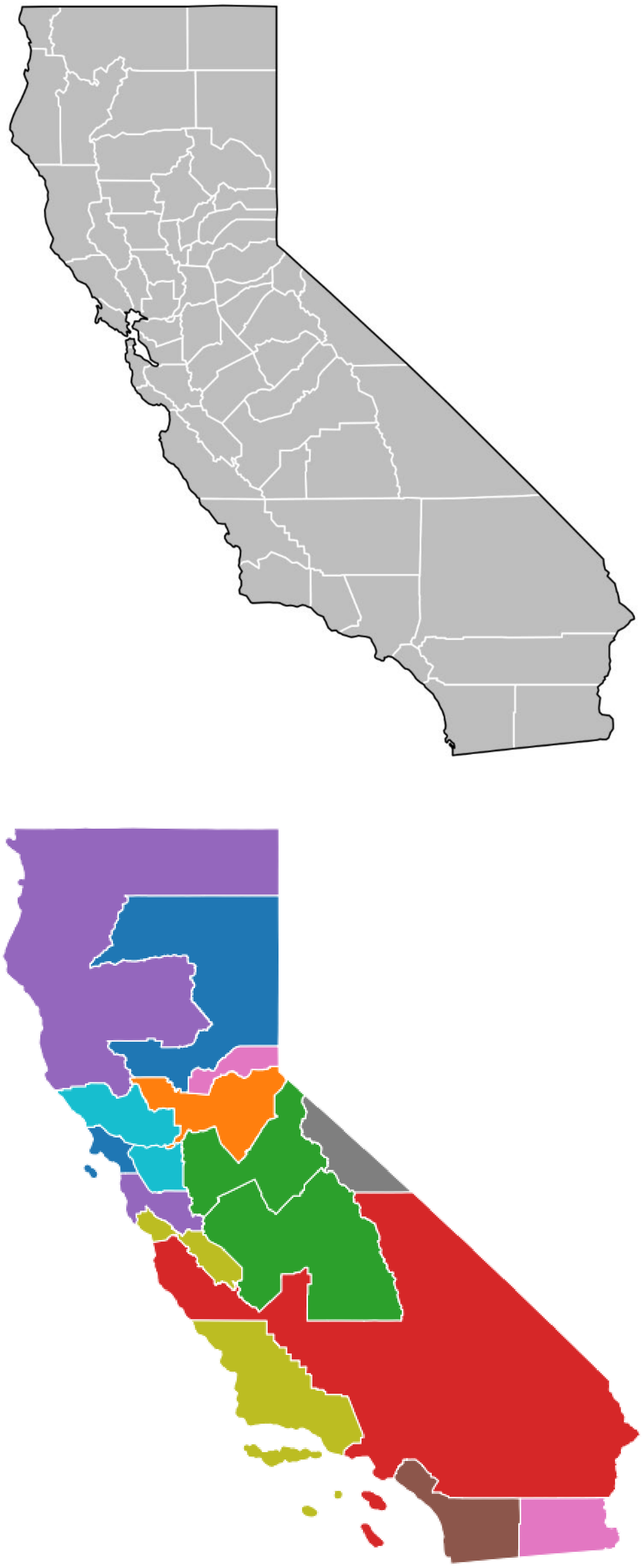

Figure 2 Top: map of California counties pre-aggregation. Bottom: optimal clustering given by Ward linkage spatial clustering using $k_{crit} = 16$ clusters. In both images, spatial units are not part of the same cluster if they are separated by a white border.

4. Conclusions and Future Directions

This work motivates the meaningful aggregation of higher resolution data and the fusion of traditional and non-traditional data sources. In the realm of SI modeling, for example, this will promote novel functional regions and facilitate new conceptualizations of human mobility. Importantly, concerns of data privacy and quality are less relevant at this new meso-scale aggregation. This has vast implications in the realm of real-time urban planning where the ability to extract meaningful, model-driven regions means planners can better predict how urban systems are impacted by epidemics, natural disasters, and infrastructure failures.

Moreover, it can be preferable to use regions obtained from this procedure instead of traditional spatial units for several reasons. Primarily, our method makes fewer assumptions about the regions, as they are designed with the model in mind. As a result, introducing this regionalization step into a modeling workflow can enhance the model's meaning and interpretation. Additionally, aggregated zones produce less computational overhead as there are fewer spatial units which must be accounted for.

This experiment provides a blueprint for better integrating regionalization and spatial analysis. As such, can be replicated with different datasets, models, and geographic contexts. In the future, this work may contribute towards several important problems in spatial data science. First, by introducing an endogenous specification of the optimal number of clusters for regionalization this research avenue is promising for making progress on the modifiable areal unit problem (MAUP)—the issue that spatial analysis results are dependent on the way in which they are aggregated—with the current work particularly focused in the context of movement data and SI (Openshaw, 1976; Openshaw, 1983). Second, by designing theory-driven regions based on multiple criteria, it may be possible to limit the number of alternative zonation schemes that favor particular groups (i.e., gerrymandering). Ultimately, this work suggests the possibility of a theoretically optimal regionalization scheme given a model and specific explanatory variables.

References

Assunção, R.M., Neves, M.C., Câmara, G., & Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7): 797-811.

Duque, J.C., Anselin, L. & Rey, S.J. (2012). The max-p regions problem. *Journal of Regional Science*, 52: 397-419.

Duque, J.C., Ramos, R., Suriñach, J. (2007). Supervised regionalization methods: a survey. *International Regional Science Review*, 30(3): 195-220.

Farmer, C.J.Q. & Fotheringham, A.S. (2011). Network-based functional regions. *Environment and Planning A*, 43: 2723-2741.

Fotheringham, A. S., & O'Kelly, M. E. (1989). *Spatial Interaction Models: Formulations and Applications*. Kluwer Academic Publishers.

Noronha, V.T. & Goodchild, M.F. (1992). Modeling interregional interaction: implications for defining functional regions. *Annals of the Association of American Geographers*, 82(1): 86-102.

Openshaw, S. (1976). Optimal zoning systems for spatial interaction models. *Environment and Planning A*, 9: 169-184.

Openshaw, S. (1983). *The modifiable areal unit problem*. Norwick: Geo Books.

Oshan, T. M. (2016). A primer for working with the Spatial Interaction modeling (SpInt) module in the python spatial analysis library (PySAL). REGION, 3(2): 11.

Wei, R., Rey, S. & Knaap, E. (2020): Efficient regionalization for spatially explicit neighborhood delineation. International Journal of Geographical Information Science.

Biographies

Tyler D. Hoffman is an undergraduate Math major at the University of Maryland, College Park. He will be attending graduate school in Geography at Arizona State University to study cutting-edge algorithms for spatial data science and to promulgate data science methods to policymakers who can best use them.

Taylor M. Oshan is an Assistant Professor in the Center for Geospatial Information Science within the Department of Geographical Sciences at the University of Maryland, College Park. His research interests are centered on developing and applying spatial analysis methods, as well as building open source tools.