Spatio-temporal analysis of socioeconomic neighborhoods: The Open Source Longitudinal Neighborhood Analysis Package (OSLNAP)

Sergio Rey[‡*], Elijah Knaap[‡], Su Han[‡], Levi Wolf[§], Wei Kang[‡]

https://youtu.be/VWMj_rNb0io

Abstract—The neighborhood effects literature represents a wide span of the social sciences broadly concerned with the influence of spatial context on social processes. From the study of segregation dynamics, the relationships between the built environment and health outcomes, to the impact of concentrated poverty on social efficacy, neighborhoods are a central construct in empirical work. From a dynamic lens, neighborhoods experience changes not only in their socioeconomic composition, but also in spatial extent; however, the literature has ignored the latter source of change. In this paper, we discuss the development of a novel, spatially explicit tool: the Open Source Longitudinal Neighborhood Analysis Package (OSLNAP) using the scientific Python ecosystem.

Index Terms—neighborhoods, GIS, clustering, dynamics

Introduction

For social scientists in a wide variety of disciplines, neighborhoods are central thematic topics, focal units of analysis, and first-class objects of inquiry. Despite their centrality in public health, sociology, geography, political science, economics, psychology, and urban planning, however, neighborhoods remain understudied. One of the reasons for that is because researchers lack appropriate analytical tools for understanding neighborhood evolution through time and space. Towards this goal we are developing the *open source longitudinal neighborhood analysis program* `(OSLNAP)`. We envisage OSLNAP as a toolkit for better, more open and reproducible science focused on neighborhoods and their sociospatial ecology. In this paper we first provide an overview of the main components of OSLNAP. Next, we present an illustration of selected OSLNAP functionality. We conclude the paper with a road map for future developments.

OSLNAP

Neighborhood analysis involves a multitude of analytic tasks, and different types of inquiry lead to different analytical pipelines in which distinct tasks are combined in sequence. OSLNAP is designed in a modular fashion to facilitate the composition of different pipelines for neighborhood analysis. Its functionality is available through several interfaces that include a web-based front end as well as a library for scripting in Jupyter notebooks or at the shell. As such, OSLNAP is intended to support different types of researchers and questions. For example, a sociologist interested in comparative segregation dynamics can use OSLNAP to derive time-consistent boundaries for a collection of US metropolitan areas from 1980-2010. Alternatively, public health epidemiologists can use the same boundaries to study the impact of neighborhood context on childhood obesity trends. Both of these types of studies might be characterized as "neighborhood effects" studies as neighborhood units serve as containers to study different socioeconomic processes.

An alternative group of studies falls under the "neighborhood dynamics" label. Here the interest is in the neighborhood units themselves and how their boundaries *and* internal socioeconomic composition evolve over time. Processes such as gentrification and the so called great inversion [Ehr12] where wealthy, higher educated, white populations are relocating into the center cities while growing numbers of minorities move to the suburbs both fundamentally restructure urban and suburban neighborhoods. OSLNAP is designed to support both neighborhood effects and neighborhood dynamics modes of inquiry.

Here we provide an overview of each of the main analytical components of OSLNAP before moving on to an illustration of how selections of the analytical functionality can be combined for particular use cases. OSLNAP's analytical components are organized into three core modules: [a] data layer; [b] neighborhood definition layer; [c] longitudinal analysis layer.

Data Layer

Like many quantitative analyses, one of the most important and challenging aspects of longitudinal neighborhood analysis is the development of a tidy and accurate dataset. When studying the socioeconomic makeup of neighborhoods over time, this challenge is compounded by the fact that the spatial units whose composition is under study often change size, shape, and configuration over time. The `harmonize` module provides social scientists with a set of simple and consistent tools for building transparent and reproducible spatiotemporal datasets. Further, the tools in `harmonize` allow researchers to investigate the implications of

* *Corresponding author: sergio.rey@ucr.edu*
‡ *Center for Geospatial Sciences, University of California, Riverside*
§ *School of Geographical Sciences, University of Bristol*

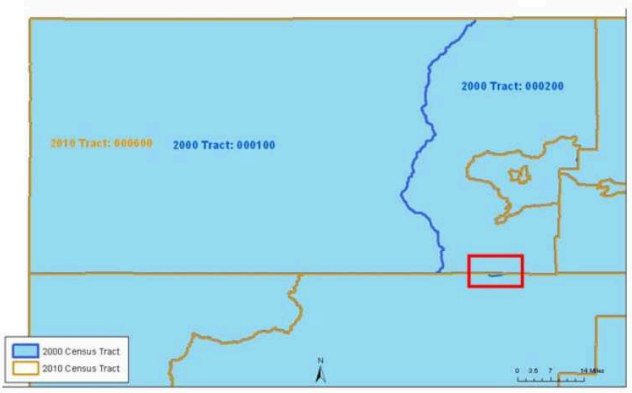

Fig. 1: *Enumeration Unit Changes [U.S10].*

alternative decisions in the data processing pipeline and how those decisions affect the results of their research.

Neighborhood demographic and socioeconomic data relevant to social scientists are typically collected via a household census or survey and aggregated to a geographic reporting unit such as a state, county or zip code which may be relatively stable. The boundaries of smaller geographies like census tracts, however, often are designed to encapsulate roughly the same number of people for the sake of comparability, which means that they are necessarily redrawn with each data release as population grows and fluctuates. Figure 1 illustrates the issues involved. Here two census tracts from 2000 have been merged to form a new tract in 2010. However, while one of the original tracts is completely contained in the new tracts, the second original tract is only partially contained in the new tract. In other words, since same physical location may fall within the boundary of different reporting units at different points in time, it is impossible to compare directly a single neighborhood with itself over time.

To facilitate temporal comparisons, research to date has proceeded by designating a "target" geographic unit or zone that is held constant over time, and allocating data from other zones using areal interpolation and other estimation techniques. This process is sometimes known as "boundary harmonization" [LSX16]. While "harmonized" data is used widely in neighborhood research, the harmonization process also has known shortcomings, since the areal interpolation of aggregate data is subject to the ecological fallacy–the geographic manifestation of which is known as the "Modifiable Areal Unit Problem" (MAUP) [Ope84]. Simply put, MAUP holds that areal interpolation introduces bias since the spatial distribution of variables in each of the overlapping zones is unknown. A number of alternative approaches have been suggested to reduce the amount of error by incorporating auxiliary data such as road networks, which help to uncover the "true" spatial distribution of underlying variables, but this remains an active area of research [Sch17], [SQ13], [Tap10], [Xie95].

In practice, these challenges mean that exceedingly few neighborhood researchers undertake harmonization routines in their own research, and those performing temporal analyses typically use exogenous, pre-harmonized boundaries from a commercial source such as the Neighborhood Change Database (NCDB) [Tat], or the freely available Longitudinal Tract Database (LTDB) [LXS14]. The developers of these products have published studies verifying the accuracy of their respective data, but those claims have gone untested because external researchers are unable to fully replicate the underlying methodology.

To overcome the issues outlined above, `OSLNAP` provides a suite of methods for conducting areal interpolation and boundary harmonization in the `harmonize` module. It leverages `geopandas` and `PySAL` for managing data and performing geospatial operations, and the PyData stack for attribute calculations [RA10]. The `harmonize` module allows a researcher to specify a set of input data (drawn from the space-time database described in the prior section), a set of target geographic units to remain constant over time, and an interpolation function that may be applied to each variable in the dataset independently. For instance, a researcher may decide to use different interpolation methods for housing prices than for the share of unemployed residents, than for total population; not only because the researcher may wish to treat rates and counts separately, but also because different auxiliary information might be applicable for different types of variables.

In a prototypical workflow, `harmonize` permits the end-user to carry out a number of tasks: [a] compile and query a spatiotemporal database using either local data or connections to public data services; [b] define the relevant variables to be harmonized and optionally apply a different (spatial and/or temporal) interpolation function to each; [c] harmonize temporal data to consistent spatial units by either selecting an existing native unit (e.g. zip codes in 2016), inputting a user-defined unit (e.g. a theoretical or newly proposed boundary), or developing new primitive units (e.g. the intersection of all polygons).

Neighborhood Identification

Neighborhoods are complex social and spatial environments with multiple interacting individuals, markets, and processes. Despite decades of research it remains difficult to quantify neighborhood context, and certainly no single variable is capable of capturing the entirety of a neighborhood's essential essence. For this reason, several traditions of urban research focus on the application of multivariate clustering algorithms to develop neighborhood typologies. Such typologies are sometimes viewed as more holistic descriptions of neighborhoods because they account for multiple characteristics simultaneously [Gal01].

One notable tradition from this perspective called "geodemographics", is used to derive prototypical neighborhoods whose residents are similar along a variety of socioeconomic and demographic attributes [FG89], [SS14]. Geodemographics have been applied widely in marketing [FE05], education [SL09], and health research [PGL+11] among a wide variety of additional fields. The geodemographic approach has also been criticized, however, for failing to model geographic space formally. In other words, the geodemographic approach ignores spatial autocorrelation, or the "first law of geography"–that the attributes of neighboring zones are likely to be similar.

Another tradition in urban research, known as "regionalization" has thus been focused on the development of multivariate clustering algorithms that account for spatial dependence explicitly. To date, however, these traditions have rarely crossed in the literature, limiting the utility each approach might have toward applications in new fields. In the `cluster` module, we implement both clustering approaches to (a) foster greater collaboration among weakly connected components in the field of geographic information science, and (b) to allow neighborhood researchers to investigate the performance of multiple different clustering

solutions in their work and evaluate the implications of including space as a formal component in their clustering models.

In OSLNAP, the `cluster` module leverages the scientific python ecosystem, building from scikit-learn [PVG+11], geopandas [Geo18], and PySAL [Rey15]. Using input from the Data Layer, the `cluster` module allows researchers to develop neighborhood typologies based on either attribute similarity (the geodemographic approach) or attribute similarity with incorporated spatial dependence (the regionalization approach). Given a space-time data set, the `cluster` module permits three different treatments of time when defining neighborhoods. The first focuses on the case where only a single cross-section is available, and the clustering is carried out to define neighborhoods for that one point in time. In the second case, multiple waves or periods of observations are available and the clustering is repeated for each time slice of observations. This can be a useful approach if researchers are interested in the durability and permanence of certain kinds of neighborhoods. If similar types reappear in multiple cross sections (e.g. if the k-means algorithm places the k-centers in approximately similar locations each time period), then it may be inferred that the metropolitan dynamics are somewhat stable, at least at the macro level, since new kinds of neighborhoods do not appear to be evolving and old, established neighborhood types remain prominent. The drawback of this approach is the type of a single neighborhood cannot be compared between two different time periods because the types are independent in each period.

In the third approach, clusters are defined from all observations in all time periods. The universe of potential neighborhood types is held constant over time, the neighborhood types are consistent across time periods, and researchers can examine how particular neighborhoods get classified into different neighborhood types as their composition transitions through different time periods. While comparatively rare in the research, this latter approach allows a richer examination of socio-spatial dynamics. By providing tools to drastically simplify the data manipulation and analysis pipeline, we aim to facilitate greater exploration of urban dynamics that will help catalyze more of this research.

To facilitate this work, the `cluster` module provides wrappers for several common clustering algorithms from `scikit-learn` that can be applied . Beyond these, however, it also provides wrappers for several *spatial* clustering algorithms from `PySAL`, in addition to a number of state-of-the art algorithms that have recently been developed [Wol18].

In a prototypical workflow, `cluster` permits the end-user to: [a] query the (tidy) space-time dataset created via the `harmonize` module; [b] define the neighborhood attributes and time periods and on which to develop a typology; [c] run one or more clustering algorithms on the space-time dataset to derive neighborhood cluster membership. Clustering may be applied cross-sectionally or on the pooled time-series, and clustering may incorporate spatial dependence, in which case `cluster` provides options for users to parameterize a spatial contiguity matrix. Clustering results may be reviewed quickly via the built-in `plot()` method, or interactively by leveraging the planned `geovisualization` module.

Longitudinal Analysis

Having identified the neighborhood types for all units of analysis over the whole time span, researchers might be interested in how they evolve over time. The third core module of `OSLNAP`'s analytical components, `change`, provides a suite of functionality towards this end. Traditional longitudinal analysis in neighborhood contexts focuses solely on changes in residential socioeconomic composition, while we and others have argued that changes in geographic footprints are also substantively interesting [RAF+11]. Therefore, this component draws upon recent methodological developments from spatial inequality dynamics and implements two broad sets of spatially explicit analytics to provide deeper insights into the evolution of socioeconomic processes and the interaction between these processes and geographic structure.

Both sets of analytics operate on time series of neighborhood types; they each take as input a set of spatial units of analysis (e.g. census tracts) that have been assigned a categorical variable for each point in time (e.g. the output of the `cluster` module). They differ, however, in how the time series are modeled and analyzed. The first set centers on *transition analysis*, which treats each time series as stochastically generated from time point to time point. It is in the same spirit of the first-order Markov Chain analysis where a (k,k) transition matrix is formed by counting transitions across all the k neighborhood types between any two consecutive time points for all spatial units. One drawback of this approach is that it treats all the time series as being independent of one another and following an identical transition mechanism. The spatial Markov approach was proposed by [Rey01] to interrogate potential spatial interactions by conditioning transition matrices on neighboring context while the spatial regime Markov approach allows several transition matrices to be formed for different spatial regimes which are constituted by contiguous spatial units. Both approaches together with inferences have been implemented in Python Spatial Analysis Library (PySAL) [Rey15] and Geospatial Distribution Dynamics (giddy) package [gid18]. The `change` module considers these packages as dependencies and wraps relevant classes and functions to make them consistent and efficient for longitudinal neighborhood analysis.

The other set of spatially explicit approach to neighborhood dynamics is concerned with *sequence analysis* which treats each time series of neighborhood types as a whole, in contrast to *transition analysis*. The core of *sequence analysis* is the similarity measure between a pair of sequences. Various aspects of a neighborhood sequence such as the order in which successive neighborhood types appears, the year(s) in which a specific neighborhood type appears, and the duration of a neighborhood type could be the focus of the similarity measure. Choosing which aspect or aspects to focus on should be driven by the research question at hand and the interpretation should proceed with caution [SR16]. A major approach of *sequence analysis*, the optimal matching (OM) algorithm, which was originally used for matching protein and DNA sequences [AT00], has been adopted to measure the similarity between neighborhood sequences in metropolitan areas such as Los Angeles and Chicago [Del16], [Del17]. It generally works by finding the minimum cost for transforming one sequence to another using a combination of operations including substitution, insertion, deletion and transposition. The similarity matrix is then used as the input for another round of clustering to derive a typology of neighborhood trajectory to produce several sequences of neighborhood types typically happening in a particular order [Del16].

In a prototypical workflow, the `change` module permits the end user to explore the nature of neighborhood change from a dynamic, holistic or combined holistic & dynamic perspective. From a dynamic perspective, *transition analysis* can be used to apply a first-order Markov chain model to look at probabilities

of transitioning between neighborhood types over time. It also supports the use of a spatial Markov chains model to interrogate the role of spatial interactions in shaping neighborhood dynamics or the application of a spatial regime Markov chains model to explore spatially heterogeneous neighborhood dynamics. From a holistic perspective, *sequence analysis* involves the application of the OM algorithm with classic cost functions for substitution, insertion, deletion and transposition, or those explicitly taking account of potential spatial dependence and spatial heterogeneity. Finally, a combined holistic & dynamic perspective is gained by feeding the output from *transiton analysis*, which is the empical transition probability matrix, or spatially dependent transition probability matrices into *sequence analysis* to help set operation costs.

Empirical Illustration

In the following sections we demonstrate the utility of `OSLNAP` by presenting the results of several initial analyses conducted with the package. We begin with a series of cluster analyses, which are then used to analyze neighborhood dynamics. Typically, workflows of this variety would require extensive data collection, munging and recombination; with `OSLNAP`, however, we accomplish the same in just a few lines of code. Using the Los Angeles metropolitan area as our example, we present three neighborhood typologies, each of which leverages the same set of demographic and socioeconomic variables, albeit with different clustering algorithms. The results show similarities across the three methods but also several marked differences. This diversity of results can be viewed as either nuisance or flexibility, depending on the research question at hand, and highlights the need for research tools that facilitate rapid creation and exploration of different neighborhood clustering solutions. For each example, we prepare a cluster analysis for the Los Angeles metropolitan region using data at the census tract level. We visualize each clustering solution on a map, describe the resulting neighborhood types, and examine the changing spatial structure over time. For each of the examples, we cluster on the following variables: race categories (percent white, percent black, percent Asian, percent Hispanic), educational attainment (share of residents with a college degree or greater) and socioeconomic status (median income, median home value, percent of residents in poverty).

Agglomerative Ward

We begin with a simple example identifying six clusters via the agglomerative Ward method. Following the geodemographic approach, we aim to find groups of neighborhoods that are similar in terms of their residential composition, regardless of whether those neighborhoods are physically proximate. Initialized with the demographic and socioeconomic variables listed earlier, the Ward method identifies three clusters that are predominantly white on average but which differ with respect to socioeconomic status. The other three clusters, meanwhile, tend to be predominantly minority neighborhoods but are differentiated mainly by the dominant racial group (black versus Hispanic/Latino) rather than by class. The results, while unsurprising to most urban scholars, highlight the continued segregation by race and class that characterize American cities. For purposes of illustration, we give each neighborhood type a stylized moniker that attempts to summarize succinctly its composition (again, a common practice in the geodemographic literature). To be clear, these labels are oversimplifications of the socioeconomic context within each type, but they help facilitate rapid consumption of the information nonetheless. The resulting clusters are presented in Figure 2.

- Type 0. racially concentrated (black and Hispanic) poverty
- Type 1. minority working class
- Type 2. integrated middle class
- Type 3. white upper class
- Type 4. racially concentrated (Hispanic) poverty
- Type 5. white working class

When the neighborhood types are mapped, geographic patterns are immediately apparent, despite the fact that space is not considered formally during the clustering process. These visualizations reveal what is known as "the first law of geography"–that near things tend to be more similar than distant things (stated otherwise, that geographic data tend to be spatially autocorrelated) [Tob70]. Even though we do not include the spatial configuration as part of the modeling process, the results show obvious patterns, where neighborhood types tend to cluster together in euclidian space. The clusters for neighborhoods type zero and four are particularly compact and persistent over time (both types characterized by racially concentrated poverty), helping to shed light on the persistence of racial and spatial inequality. With these types of visualizations in hand, researchers are equipped not only with analytical tools to understand how neighborhood composition can affect the lives of its residents (a research tradition known as neighborhood effects), but also how neighborhood identities can transform (or remain stagnant) over time and space. Beyond the simple diagnostics plots presented above, `OSLNAP` also includes an interactive visualization interface that allows users to interrogate the results of their analyses in a dynamic web-based environment where interactive charts and maps automatically readjust according to user selections.

Affinity Propagation

Affinity propagation is a newer clustering algorithm with implementations in scikit-learn that is capable of determining the number of clusters endogenously (subject to a few tuning parameters). Initialized with the default settings, `OSLNAP` discovers 14 neighborhood types in the Los Angeles region; in a way, this increases the resolution of the analysis beyond the Ward example, since increasing the number of clusters means neighborhoods are more tightly defined with lower variance in their constituent variables. On the other hand, increasing the number of neighborhood types also increase the difficulty of interpretation since the each type will be, by definition, less differentiable from the others. In the proceeding section, we discuss how researchers can exploit this variability in neighborhood identification to yield different types of dynamic analyses. Again, we find it useful to present stylized labels to describe each neighborhood type:

- Type 0. white working class
- Type 1. white extreme wealth
- Type 2. black working class
- Type 3. Hispanic poverty
- Type 4. integrated poverty
- Type 5. Asian middle class
- Type 6. white upper-middle class
- Type 7. integrated Hispanic middle class
- Type 8. extreme racially concentrated poverty
- Type 9. integrated extreme poverty

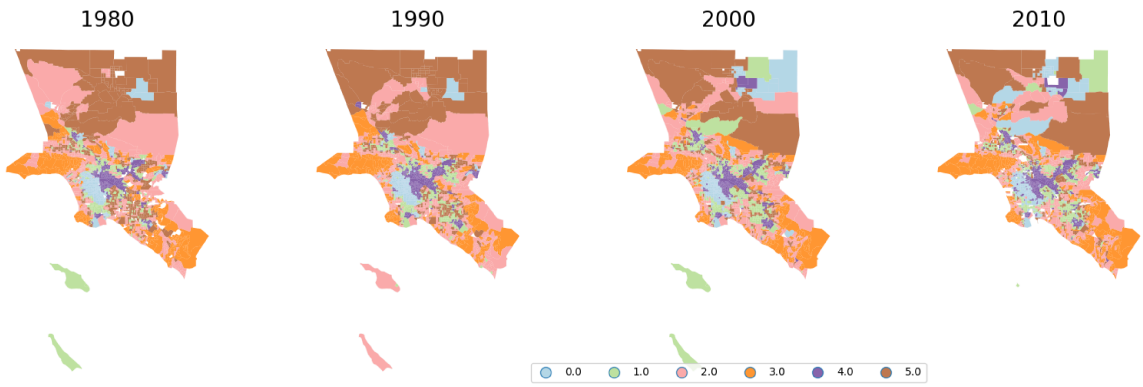

Fig. 2: *Neighborhood Types in LA using Ward Clustering.*

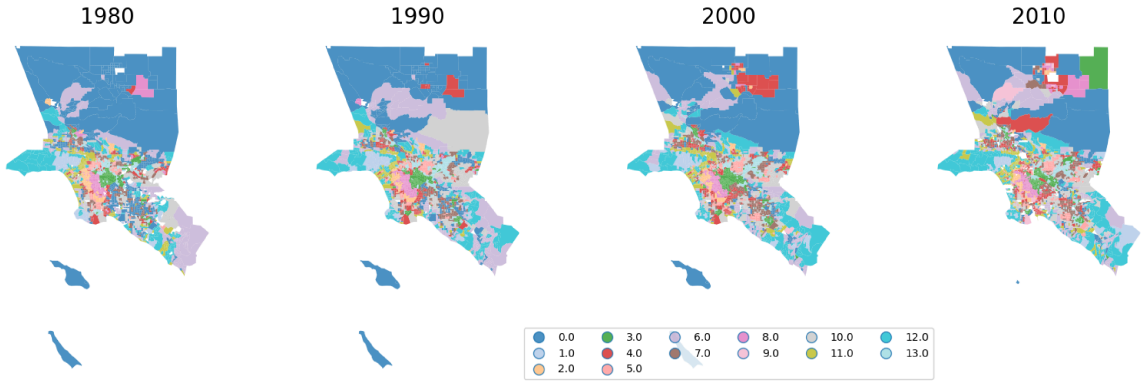

Fig. 3: *Neighborhood Types in LA using Affinity Propagation.*

- Type 10. Asian upper middle class
- Type 11. integrated white middle class
- Type 12. white elite
- Type 13. Hispanic middle class

Despite having more than double the number of neighborhood types in the Ward example, many of the spatial patterns remain when using affinity propagation clustering, including concentrated racial poverty in South Central LA, concentrated affluence along much of the coastline, black and Hispanic enclaves in the core of the city, and white working class strongholds in more rural areas to the north of the region. Comparing these two examples makes clear that some of the sociodemographic patterns in the LA region are quite stable, and are somewhat robust to the clustering method or number of clusters. Conversely, by increasing the number of clusters in the model, researchers can explore a much richer mosaic of social patterns and their evolution over time, such as the continued diversification of the I-5 corridor along the southern portion of the region.

SKATER

Breaking from the geodemographic approach, the third example leverages SKATER, a spatially-constrained clustering algorithm that finds groups of neighborhoods that are similar in composition, but groups them together if and only if they also satisfy the criteria for a particular geographic relationship [Wol18]. As such, the family of clustering algorithms that incorporate spatial constraints (from the tradition known as "regionalization") must be applied cross-sectionally, and yield an independent set of clusters for each time period, as shown in Figure 4. The clusters, thus, depend not only on the composition of the census units, but also their spatial configuration and connectivity structure at any given time.

Despite the fact that clusters are independent from one year to the next (and thus, we lack appropriate space in this text for describing the SKATER results for each year) comparing the results over time nonetheless yield some interesting insights. Regardless of the changing spatial and demographic structure of the Los Angeles region, some of the neighborhood boundaries identified are remarkably stable, such as the area of concentrated affluence in Beverly Hills and its nearby communities that jut out to the region's West. Conversely, there is considerable change among the predominantly minority communities in the center of the region, whose boundaries appear to be evolving considerably over time. In these places, a researcher might use the output from SKATER to conduct an analysis to determine the ways in which the empirical neighborhood boundaries derived from SKATER conform to residents' perceptions of such boundaries, their evolution over time, and their social re-definition as developed by different residential groups [Wol18]. Irrespective of its

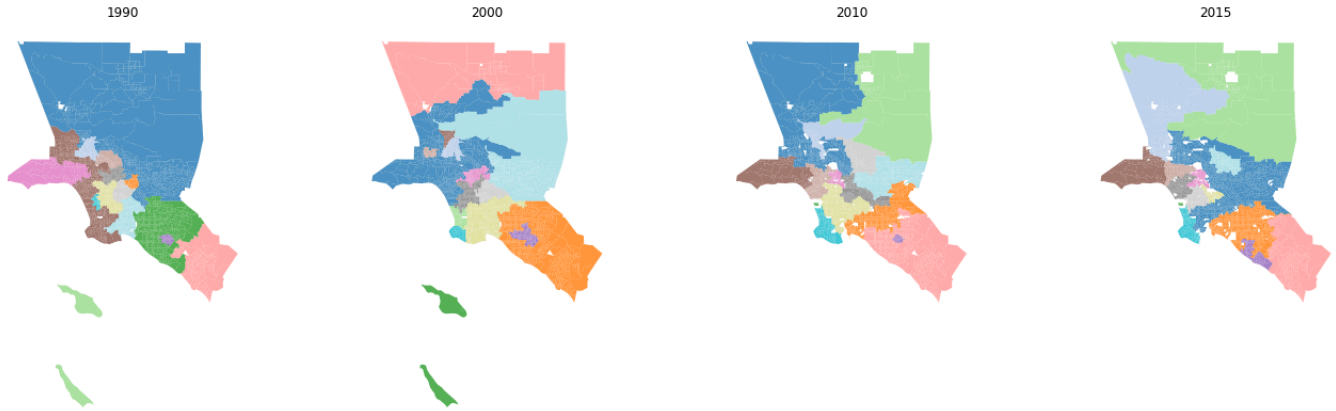

Fig. 4: *Neighborhood Types in LA using SKATER.*

particular use, the regionalization approach presents neighborhood researchers with another critical tool for understanding the bi-directional relationship between people and places.

In each of the sample analyses presented above, we use `OSLNAP` to derive a set of neighborhood clusters or types that can be used to analyze the demographic makeup of places over time. In some cases, these maps can serve as foundations for descriptive analyses or be analyzed as research projects in their own right. In other cases, in which social processes rather than the demographic makeup of communities are the focus of study, the neighborhood types derived here can be used as input to dynamic analyses of neighborhood change and evolution, particularly as they relate to phenomena such as gentrification and displacement. In the following sections, we demonstrate how the neighborhood typologies generated by `OSLNAP`'s `cluster` module can be used as input to the `change` module to explore the neighborhood evolution.

Transition Analysis to Neighborhood Change

The `change` module can provide insights into the nature of neighborhood change in the Los Angeles metropolitan area. We utilize the neighborhood types for all census tracts of the Los Angeles metropolitan area across four census years identified by selected clustering algorithms in the former section as the input for the `change` module. Among the three clustering algorithms, SKATER was applied to each cross section of census tracts independently yielding clusters which are not directly comparable over time. Thus, we focus only on the six neighborhood types identified by the agglomerative Ward method (Fig. 2) and the fourteen neighborhood types identified by the affinity propagation method (Fig. 3).

We start with the aspatial transition analysis which pools all the time series of neighborhood types and counts how many transitions between any pair of neighborhood types across immediate consecutive census years $(t, t+10)$ (or $(t, t+5)$ for 2010-2015) which are further organized into a (k,k) transition count matrix $\boldsymbol{N}$. Adopting the maximum likelihood estimator for the first-order Markov transition probability as shown in Equation (1), a (k,k) transition probability matrix can thus be constructed providing the insights in the underlying dynamics of neighborhood change. The $(6,6)$ and the $(14,14)$ transition probability matrices for Ward and affinity propagation clusters are estimated and visualized in

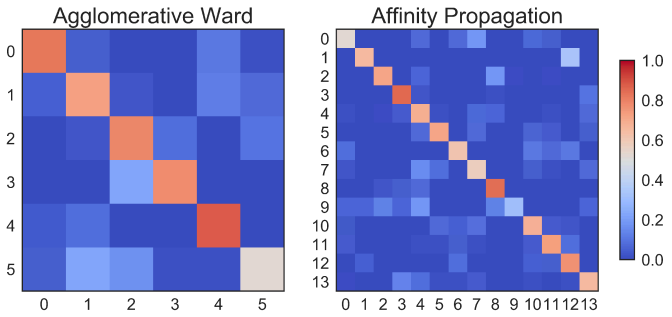

Fig. 5: *Markov transition probability matrix for Ward and Affinity Propagation clusters.*

Fig. 5 where the color in grid (i, j) represents the probability of transitioning from neighborhood type i to j in the next census year. It is obvious that both transition probability matrices are characterized by large diagonal entries, indicating a certain level of neighborhood stability for the focal four census years. This is especially true for the Ward neighborhood type 4 which is characterized by racially concentrated (Hispanic) poverty. The probability of staying at this type is 0.876 meaning that there is only 12.4% chance of changing to other neighborhood types once the census tract enters into type 4.

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{q=1}^{k} n_{iq}}, \quad \text{where} \quad i, j \in \mathbb{S} = \{1, 2, \cdots, k\} \tag{1}$$

Moving from the aspatial transition analysis, we interrogate potential spatial interactions among neighborhood dynamics using the spatial Markov chain approach. More specifically, we hypothesize that the transition probability for any focal census tract is not constant, but rather dependent on the spatial context, that is, the most common neighborhood type of contiguous tracts, the so-called spatial lag. Therefore, k exhaustive and mutually exclusive subsamples are constructed based on the spatial lag at t, from which k (k,k) transition probability matrices are estimated based on Equation (1). Fig. 6 displays the spatial Markov transition probability matrices for Ward neighborhood types. It should be noted that the interpretation with these conditional transition probabilities should proceed with caution as the increased number of parameters to be estimated here could lead to large standard errors for some estimates. For example, the $(0,0)$ entry in the

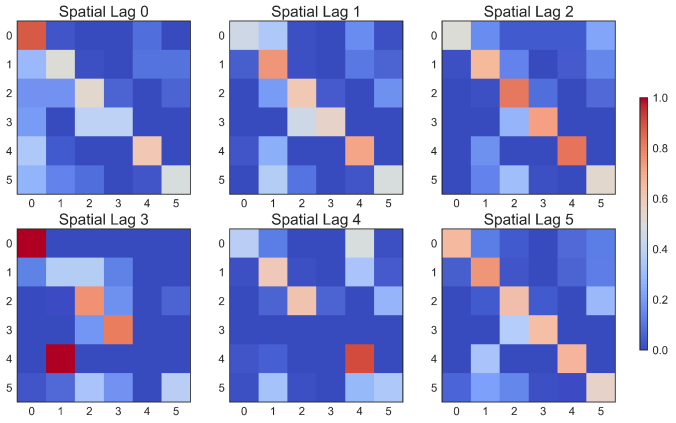

Fig. 6: *Spatial Markov transition probability matrices for Ward clusters.*

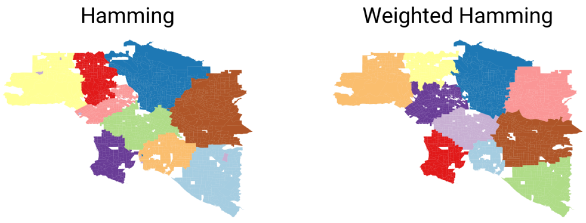

Fig. 7: *Neighborhoods with similar spatial-social histories since 1980*

subplot of Spatial Lag 3 is 1. The tendency of interpreting the 100 percent to be tracts "perfectly stuck at" Ward neighborhood type 0 if the spatial lag is type 3 should be compromised by the fact that there is only 1 observation transitioning from type 0 which has the spatial lag of type 3 at t and this very observation happens to stay at type 0. Since we are short of information, we could not conclude with the "perfectly stuck" theory. The spatial Markov tests (available upon request) including the likelihood ratio test and the χ^2 test [BB03], [RKW16] are both rejected indicating that neighboring context plays an important role in shaping the neighborhood dynamics.

Sequence Analysis to Neighborhood Change

Armed with the sequences of sociodemographic classifications for every harmonized tract in LA, the distance between these sequences can be computed. Since these sequences are intrinsically aligned in time, the Hamming distance between classifications yields an effective metric for how different places' demographic changes have been. The pairwise Hamming distance matrix for demographic transitions in LA is sufficient to recover a set of boundaries. However, alone, this metric only considers that two areas are in different sociodemographic classifications at a specific point in time. It does not consider the difference in the attribute's strength of assignment in these classifications, nor does it consider how well an area fits into its demographic classification.

Conceptually, this is important; even though the gist of the demographic classifications stay consistent over time, the members of these classes may shift around significantly over time. As a tract drifts from one classification to another classification over time, it may move within the class before it hops classifications if the movement is slow. This means that, at each point in time, tracts are more or less representative of their clusters; a transition of one area from "white working class" to "white upper class" may not necessarily reflect the same amount of social/spatial volatility as a move from "minority working class" to "white upper class," as might happen during rapid gentrification.

As such, we can also weight the edit distance based on how "expensive" the edit is in terms of the clustering distance. Using this weighting method, not all transitions from white working class to white upper class will be treated the same: observations that are "almost" white upper class but not quite will be considered more similar to white upper class tracts. But, since a reassignment is still involved, there will still be a cost associated with that edit. Clusterings for both the raw Hamming edit distance and the weighted Hamming edit distances over sociodemographic sequences are shown in Figure 7 using [Wol18]. Broadly speaking, the assignments between the two clustering methods are strongly related (with an adjusted Rand index of .68), but macro-level distinctions between assignment structures are visible, particularly in the areas of central northern LA near the Hollywood Hills, as well as the areas of east LA, near Fullerton. This means that, when the subclassification information is taken into account, clusterings can change. However, when examining spatially-contiguous clusters, the total amount of possible change is often quite constrained as well. Thus, the move from unweighted to weighted edit distances may make even more of a difference in some cases.

Future Directions

At present, we are in the early phases of the project and moving forward we will be focusing on the following directions.

Parameter sweeps: In the definition of neighborhoods, a researcher faces a daunting number of decisions surrounding treatment of harmonization, selection of variables, and choice of clustering algorithm, among others. In the neighborhood literature, the implications of these decisions remain unexplored and this is due to the computational burdens that have precluded formal examination. We plan on a modular design for OSLNAP that would support extensive parameter sweeps to provide an empirical basis for exploring these issues and to offer applied researchers computationally informed guidance on these decisions.

Data services: OSLNAP is being designed to work with existing harmonized data sets available from various firms and research labs. Because these fall under restrictive licenses, users must first acquire these sources - they cannot be distributed with OLSNAP. To address the limitations associated with this strategy, we are exploring interfaces to public data services such as CenPy [cen18] and tigris [tig18].

Interactive visualization: Apart from scripted environments demonstrated in this paper, OSLNAP is being designed with a web-based, interactive front-end that allows users to explore the results of different neighborhood analyses with the assistance of linked maps, charts, and tables. Together, these linked "views" allow a researcher to interrogate their results in a manner far richer than creating a series of static maps.

Reproducible Urban Data Science: A final direction for future research is the development of reproducible workflows as part of OSLNAP. Here we envisage leveraging our earlier work on provenance for spatial analytical workflows [ARL14] and extending it to the full longitudinal neighborhood analysis pipeline.

Conclusion

In this paper we have presented the motivation for, initial design, and implementation of OSLNAP. We feel that, even at this early stage in the project, OSLNAP has benefitted from the scope and deep nature of the PyData stack as we have been able to move from conceptualization to prototyping in fairly short order. At the same time, we see OSLNAP playing an important role in widening the use of Python in urban and spatial data science. We are looking forward to the future development of OSLNAP and interaction with both the PyDATA community and the broader community of computational social sciences.

Acknowledgment

This research was supported by NSF grant SES-1733705.

References

[ARL14] Luc Anselin, Sergio J. Rey, and Wenwen Li. Metadata and provenance for spatial analysis: the case of spatial weights. *International Journal of Geographical Information Science*, 28(11):2261–2280, May 2014. doi:10.1080/13658816.2014.917313.

[AT00] Andrew Abbott and Angela Tsay. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1):3–33, 2000. doi:10.1177/0049124100029001001.

[BB03] F. Bickenbach and E. Bode. Evaluating the Markov property in studies of economic convergence. *International Regional Science Review*, 26(3):363–392, 2003. doi:10.1177/0160017603253789.

[cen18] cenpy Developers. cenpy. https://github.com/ljwolf/cenpy, 2018.

[Del16] Elizabeth C Delmelle. Mapping the DNA of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change. *Annals of the American Association of Geographers*, 106(1):36–56, 2016. doi:10.1080/00045608.2015.1096188.

[Del17] Elizabeth C Delmelle. Differentiating pathways of neighborhood change in 50 U.S. metropolitan areas. *Environment and Planning A*, 49(10):2402–2424, oct 2017. doi:10.1177/0308518X17722564.

[Ehr12] Alan Ehrenhalt. *The great inversion and the future of the American city*. Random House, 2012.

[FE05] Marc Farr and Andy Evans. Identifying 'unknown diabetics' using geodemographics and social marketing. *Journal of Direct, Data and Digital Marketing Practice*, 7(1):47–58, aug 2005. doi:10.1057/palgrave.dddmp.4340504.

[FG89] R Flowerdew and W Goldstein. Geodemographics in Practice: Developments in North America. *Environment and Planning A*, 21(5):605–616, may 1989. doi:10.1068/a210605.

[Gal01] George Galster. On the Nature of Neighbourhood. *Urban Studies*, 38(12):2111–2124, nov 2001. doi:10.1080/00420980120087072.

[Geo18] GeoPandas Developers. GeoPandas 0.3.0. http://geopandas.org/index.html, 2018.

[gid18] giddy Developers. GeospatIal Distribution DYnamcis. http://github.com/pysal/giddy.html, 2018.

[LSX16] John R. Logan, Brian J. Stults, and Zengwang Xu. Validating population estimates for harmonized census tract data, 2000-2010. *Annals of the American Association of Geographers*, 106(5):1013–1029, Jun 2016. doi:10.1080/24694452.2016.1187060.

[LXS14] John R. Logan, Zengwang Xu, and Brian J. Stults. Interpolating U.S. decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer*, 66(3):412–420, May 2014. doi:10.1080/00330124.2014.905156.

[Ope84] S Openshaw. Ecological Fallacies and the Analysis of Areal Census Data. *Environment and Planning A*, 16(1):17–31, jan 1984. doi:10.1068/a160017.

[PGL+11] Jakob Petersen, Maurizio Gibin, Paul Longley, Pablo Mateos, Philip Atkinson, and David Ashby. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems*, 13(2):173–192, 2011. doi:10.1007/s10109-010-0113-9.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RA10] Sergio J. Rey and Luc Anselin. PySAL: A Python Library of Spatial Analytical Methods. In *Handbook of Applied Spatial Analysis*, volume 37, pages 175–193. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-03647-7_11.

[RAF+11] Sergio J. Rey, Luc Anselin, David C. Folch, Daniel Arribas-Bel, Myrna L. Sastré Gutiérrez, and Lindsey Interlante. Measuring Spatial Dynamics in Metropolitan Areas. *Economic Development Quarterly*, 25(1):54–64, feb 2011. doi:10.1177/0891242410383414.

[Rey01] S. J. Rey. Spatial empirics for economic growth and convergence. *Geographical Analysis*, 33(3):195–214, 2001. doi:10.1111/j.1538-4632.2001.tb00444.x.

[Rey15] Sergio J. Rey. Python Spatial Analysis Library (PySAL): An update and illustration. In Chris Brunsdon and Alex Singleton, editors, *Geocomputation: A Practical Primer*, pages 233–253. SAGE Publications Ltd, 2015. doi:10.1007/978-3-642-03647-7_11.

[RKW16] Sergio J. Rey, Wei Kang, and Levi Wolf. The properties of tests for spatial effects in discrete markov chain models of regional income distribution dynamics. *Journal of Geographical Systems*, 18(4):377–398, 2016. doi:10.1007/s10109-016-0234-x.

[Sch17] Jonathan P. Schroeder. Hybrid areal interpolation of census counts from 2000 blocks to 2010 geographies. *Computers, Environment and Urban Systems*, 62:53–63, Mar 2017. doi:10.1016/j.compenvurbsys.2016.10.001.

[SL09] Alexander D Singleton and Paul A Longley. Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science*, 88(3):643–666, aug 2009. doi:10.1111/j.1435-5957.2008.00197.x.

[SQ13] Harini Sridharan and Fang Qiu. A Spatially Disaggregated Areal Interpolation Model Using Light Detection and Ranging-Derived Building Volumes. *Geographical Analysis*, 45(3):238–258, jul 2013. doi:10.1111/gean.12010.

[SR16] Matthias Studer and Gilbert Ritschard. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511, 2016. doi:10.1111/rssa.12125.

[SS14] Alexander D Singleton and Seth E Spielman. The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom. *The Professional Geographer*, 66(4):558–567, oct 2014. doi:10.1080/00330124.2013.848764.

[Tap10] Anna F. Tapp. Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources. *Cartography and Geographic Information Science*, 37(3):215–228, 2010. doi:10.1559/152304010792194976.

[Tat] Peter Tatian. Local scene: Neighborhood change database (ncdb). *PsycEXTRA Dataset*. doi:10.1037/e479172006-003.

[tig18] tigris Developers. tigris. https://github.com/walkerke/tigris, 2018.

[Tob70] W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240, 1970. doi:10.2307/143141.

[U.S10] U.S. Census. *Understanding the 2010 Tract Relationship Files*, 2010. URL: https://www2.census.gov/geo/pdfs/maps-data/data/rel/tractrelfile.pdf.

[Wol18] Levi John Wolf. Spatially-Encouraged Spectral Clustering : A Critical Revision of Spatially-Constrained Spectral Clustering. 2018.

[Xie95] Yichun Xie. The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems*, 19(4):287–306, 1995. doi:10.1016/0198-9715(95)00028-3.