Spatially weighted structural similarity index: A multiscale comparison tool for diverse sources of mobility data

Jessica Embury
Department of Geography
San Diego State University
San Diego, CA, USA
jembury8568@sdsu.edu

Atsushi Nara, PhD
Department of Geography
San Diego State University
San Diego, CA, USA
anara@sdsu.edu

Chanwoo Jin
Department of Humanities and Social Sciences
Northwest Missouri State Univ.
Maryville, MO, USA
jchanwoo@nwmissouri.edu

ABSTRACT

Data collected about routine human activity and mobility is used in diverse applications to improve our society. Robust models are needed to address the challenges of our increasingly interconnected world. Methods capable of portraying the dynamic properties of complex human systems, such as simulation modeling, must comply to rigorous data requirements. Modern data sources, like SafeGraph, provide aggregate data collected from location aware technologies. Opportunities and challenges arise to incorporate the new data into existing analysis and modeling methods.

Our research employs a multiscale spatial similarity index to compare diverse origin-destination mobility datasets. Established distance ranges accommodate spatial variability in the model's datasets. This paper explores how similarity scores change with different aggregations to address discrepancies in the source data's temporal granularity. We suggest possible explanations for variations in the similarity scores and extract characteristics of human mobility for the study area.

The multiscale spatial similarity index may be integrated into a vast array of analysis and modeling workflows, either during preliminary analysis or later evaluation phases as a method of data validation (e.g., agent-based models). We propose that the demonstrated tool has potential to enhance mobility modeling methods in the context of complex human systems.

CCS CONCEPTS

• **Applied computing** → Law, social and behavioral sciences

HANIMOB '22, November 1, 2022, Seattle, WA, USA

ACM 978-1-4503-9534-2/22/11…$15.00
https://doi.org/10.1145/3557921.3565542

KEYWORDS

human mobility, complex adaptive systems, spatial analysis, exploratory data analysis, validation

ACM Reference Format:

Jessica Embury, Atsushi Nara, and Chanwoo Jin. 2022. Spatially weighted structural similarity index: A multiscale comparison tool for diverse sources of mobility data. In *HANIMOB'22: The 2nd ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility Proceedings. November 1, 2022, Seattle, WA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557921.3565542

1 INTRODUCTION

Our society is defined by the basic choices and daily activities of individuals. Although simple tasks like commuting to work may seem mundane, they ultimately give rise to complex properties when considered in aggregate. Accurate insights into the dynamics of human activity and mobility are required for transportation planning, emergency management, and pandemic response as well as diverse applications in the fields of geography, ecology, economics, and computer science [1, 2, 3]. For example, travel forecast models are used to plan road infrastructure projects and public transit schedules. However, traffic congestion, a complex phenomenon, causes nonlinear travel time delays that are not well predicted by conventional equation-based models [4, 5].

Methods capable of analyzing complex human dynamics tend to have rigorous data requirements. For instance, agent-based models (ABMs) need fine-resolution data to simulate the activities and mobility of individuals. Limited data availability has hindered model development and evaluation since the inception of agent-based modeling in the late 1980s [6]. Fortunately, modern sources of big data collected from location aware technologies such as mobile phones mitigate data availability problems. SafeGraph (https://www.safegraph.com) is a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places, via the SafeGraph Community. SafeGraph provides mobility datasets at fine spatiotemporal resolutions from a sample population of approximately 10% of mobile devices in the United States of America (US) [7].

Detailed datasets from sources like SafeGraph create an opportunity to enhance existing analysis and modeling methods. Modelers and analysts are challenged to establish practices that fully integrate modern data into traditional methodologies. In addition, the inherently geographic nature of modern data calls for the innovation of spatially explicit methods with improved capacity to detect and represent multiscale complex behaviors.

Our research demonstrates a multiscale data comparison tool that reveals the similarities and differences between mobility datasets from diverse sources. We focus on the application of a spatial similarity index to compare two origin-destination datasets and extract characteristics of human mobility across space and time that guide future research activities. A better understanding of the relationship between modern and traditional datasets can highlight avenues through which new data might complement existing analysis and modeling methods.

2 METHODS

We retrieved datasets from a modern data source, SafeGraph [8], and a traditional data source, the US Census Bureau's Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) [9]. Both datasets provide origin-destination mobility flow data between census block groups (CBGs). Although the datasets share a spatial resolution, LODES data has an annual temporal granularity while SafeGraph provides daily values. For the analysis, we aggregated the SafeGraph data by year, month, and day of the week. LODES data is a representation of home-work commuter mobility and SafeGraph reports mobility for all activity contexts (e.g., work, school, recreation). SafeGraph excludes CBG information if fewer than five devices visited an establishment in a month from a given CBG.

Our study area is San Diego County, CA, US. San Diego County has 1,794 CBGs. We used 2019 data to avoid the influence of the COVID-19 pandemic on human mobility patterns.

To compare the LODES and SafeGraph data, we adopted a spatially weighted structural similarity index (SpSSIM) that was previously used to analyze social media data [10]. SpSSIM extends an image structural similarity index (SSIM) [11] to compare different sources of mobility flow data in origin-destination matrices. While SSIM compares the attributes of each pixel and its surrounding neighborhood, the SpSSIM tool compares mobility flows within set distance ranges (e.g., 0km-10km, 10km-20km).

After initial data processing, we standardized the origin-destination matrices by transforming the raw values to flow probabilities using Equation 1. Weights were applied to the flow probability matrices so that only flows in the same distance range are grouped for comparison. Equation 2 displays the formula for calculating SpSSIM [10].

$$x'_{ij} = \frac{x_{ij}}{\sum_{j=0}^{m} x_{ij}} \ \forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, m\} \qquad (1)$$

where x_{ij}' is the flow probability of CBG pair *ij*,
x_{ij} is the raw value of CBG pair *ij*,
i is the origin CBG ($i = 1, \dots, n$), and
j is the destination CBG ($j = 1, \dots, m$)

$$SpSSIM\ (x, y, w) = \frac{(2\mu_{wx}\mu_{wy} + C_1)(2\sigma_{wx,wy} + C_2)}{(\mu_{wx}^2 + \mu_{wy}^2 + C_1)(\sigma_{wx}^2 + \sigma_{wy}^2 + C_2)} \qquad (2)$$

where *x* and *y* are origin-destination matrices,
w is a weights matrix,
μ_{wx} is the mean value of matrix product *wx,*
μ_{wy} is the mean value of matrix product *wy*,
σ_{wx}^2 is the variance of *wx*,
σ_{wy}^2 is the variance of *wy*, and
$\sigma_{wx,wy}$ is the covariance of *wx* and *wy*

The constants, *C1* and *C2*, ensure that SpSSIM values range from 0 to 1, with higher values indicating greater similarity.

To explore the spatial variability of similarity, we selected 5 km and 10 km distance ranges. We initially grouped CBG pairs into distance bins of 10 km, per the case study in [10], and calculated distance-based SpSSIM scores for each distance range. We determined a global SpSSIM value by taking the mean of the distance-based SpSSIM scores. The 10 km distance ranges revealed a distinctive similarity trend. We also calculated SpSSIM scores using 5 km distance ranges to further investigate the pattern at a finer resolution.

We developed a Python package to automate the described process and improve the accessibility of the demonstrated technique: https://github.com/jlembury/spssim_analysis.

3 RESULTS

We anticipated major discrepancies between the LODES and SafeGraph datasets because LODES data represents home-work mobility while SafeGraph collects mobility data during all activity contexts. SpSSIM scores for the SafeGraph and LODES datasets confirmed low overall similarity between the mobility flows. However, we discovered interesting patterns of similarity that clarify San Diego County's mobility patterns and support future research activities.

Figure 1 displays the distance-based SpSSIM scores for monthly aggregations and Figure 2 illustrates the distance-based SpSSIMs for aggregations by day of the week. In general, there was an inverse relationship between distance-based SpSSIM scores and mobility flow distance. Mobility flows of less than 20 km had the highest scores overall.

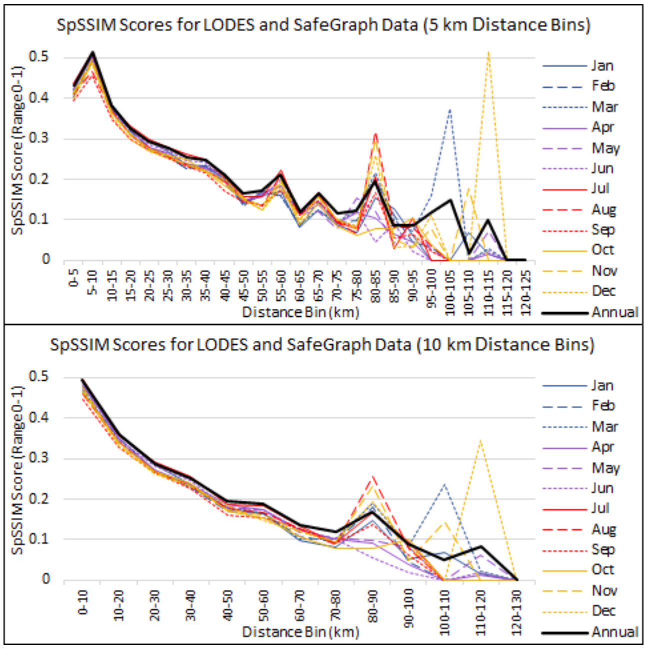

Figure 1: Spatially weighted structural similarity index scores (SpSSIM) of census block group mobility flows by distance bin size (5 km, 10 km) for US Census Bureau LODES 2019 origin-destination statistics (annual) and SafeGraph 2019 origin-destination data aggregations (monthly, annual).

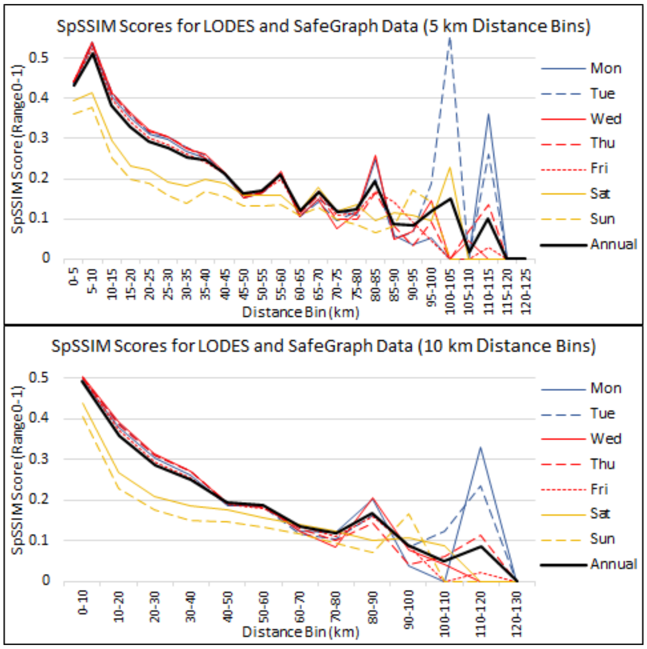

Figure 2: Spatially weighted structural similarity index scores (SpSSIM) of census block group mobility flows by distance bin size (5 km, 10 km) for US Census Bureau LODES 2019 origin-destination statistics (annual) and SafeGraph 2019 origin-destination data aggregations (day of week, annual).

Although there are outliers, distance-based SpSSIM scores were lowest for mobility flows greater than 100 km, especially for flows between 120 km and the maximum distance (123.1 km). These "high distance" bins included a relatively low number of CBG pairs in their weighted matrices, possibly influencing the results. At distances greater than 70 km, the variability of the distance-based SpSSIMs for different SafeGraph data aggregations greatly increased.

While the distance-based SpSSIM scores explored variations in similarity across spatial scales, the global SpSSIM scores are an overall indicator of how the different SafeGraph aggregations compared to the LODES data.

Table 1 contains global SpSSIM scores for all data aggregations by year, month, and day of the week. The global SpSSIM score for the LODES 2019 data and cumulative 2019 SafeGraph annual data equaled 0.191 and 0.186 for the 5 km and 10 km distance bins, respectively. Overall, similarity scores were higher when using the 5 km distance bins. Similarity scores for monthly SafeGraph data were highest during March and December, and lowest during September and October. The similarity pattern observed by days of the week was more consistent. SpSSIM scores were highest on Mondays and Tuesdays and lowest on Saturdays and Sundays.

Table 1: Global spatially weighted structural similarity index scores (SpSSIM) of census block group mobility flows for SafeGraph 2019 origin-destination data aggregations and US Census Bureau LODES 2019 origin-destination statistics, using 5 km and 10 km distance bin sizes.

Time Category	Time Period	SpSSIM (5 km bins)	SpSSIM (10 km bins)
Annual	2019	0.191	0.186
Month	January	0.165	0.167
	February	0.162	0.159
	March	0.186	0.186
	April	0.162	0.159
	May	0.164	0.165
	June	0.161	0.156
	July	0.172	0.174
	August	0.167	0.169
	September	0.154	0.154
	October	0.156	0.154
	November	0.173	0.176
	December	0.183	0.189
Week Day(s)	Sundays	0.132	0.130
	Mondays	0.193	0.202
	Tuesdays	0.218	0.205
	Wednesdays	0.187	0.184
	Thursdays	0.187	0.187
	Fridays	0.180	0.170
	Saturdays	0.160	0.154
	Weekdays (Mon-Fri)	0.198	0.195
	Weekends (Sat-Sun)	0.151	0.145

4 DISCUSSION

SpSSIM offers an innovative look at mobility datasets from diverse sources by highlighting their similarities and differences. The transformation of raw origin-destination data to probability flows facilitates direct comparison between values of (dis)similar magnitude to provide an improved understanding of the data. The use of distance ranges and multiple distance bin sizes shows how the data relationship changes across spatial scales.

The low SpSSIM scores from the analysis suggest that commuter mobility only accounts for a small portion of total mobility within San Diego County. By looking at different aggregations of SafeGraph data, we identified the overall influence of commuter mobility during different months of the year and on different days of the week. As an exploratory tool, explanation for the observed monthly variations is challenging, but seasonal fluctuations in tourism or annual academic schedules are potential underlying reasons. The similarity trend for days of the week proves easier to explain. SpSSIM scores were highest on Mondays and Tuesdays, corresponding to common working days. Conversely, Saturdays and Sundays had the lowest SpSSIM scores, reflecting regular days off for Monday-Friday workers. The highest SpSSIM scores (~0.5) were observed in the small distance ranges and suggest that about 50% of origin-destination mobility patterns in the shortest travel distance ranges are similar in the two datasets. Since LODES data represents home-work commuting mobility, this further implies that 50% of the SafeGraph data represents commuting flows while the other 50% is explained by different mobility characteristics.

An advantage of the SpSSIM tool is its flexibility to fit into a variety of analysis and modeling workflows. As demonstrated by this paper, SpSSIM can be used during preliminary analysis to better understand mobility data from different sources. Moreover, SpSSIM might be applied during later stages to evaluate generated research data against development and validation data.

We propose that this multiscale comparison technique would be beneficial when evaluating models of complex human systems. For instance, due to limited data availability, ABM validation with independent data was historically infeasible and models were typically validated using development data [6, 12, 13, 14]. Thanks to the introduction of data from modern sources such as SafeGraph, independent ABM validation is now possible. Tools like SpSSIM are well suited to provide side-by-side validation of simulated mobility outputs against development data (e.g., LODES) and independent data (e.g., SafeGraph).

The relationship between human mobility models and SpSSIM appears to be symbiotic. While SpSSIM might serve as a valuable validation method, model simulations can recreate and explain SpSSIM findings. In this way, SpSSIM and modeling methods would mutually benefit one another to advance our current models of human mobility and provide substantive insights for real-world applications.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant No. 2031407, CS4All project titled "Collaborative Research: Encoding Geography Scaling up an RPP to achieve inclusive geocomputational education". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Ashutosh Trivedi and Shrisha Rao. 2018. Agent-Based Modeling of Emergency Evacuations Considering Human Panic Behavior. *IEEE Transactions on Computational Social Systems*, 5, 1, 277-288. https://doi.org/10.1109/TCSS.2017.2783332

[2] David O'Sullivan, Mark Gahegan, Daniel J. Exeter, and Benjamin Adams. 2020. Spatially explicit models for exploring COVID-19 lockdown strategies. *Transactions in GIS*, 24, 967–1000. https://doi.org/10.1111/tgis.12660

[3] J. Gareth Polhill, Jiaqi Ge, Matthew P. Hare, Keith B. Matthews, Alessandro Gimona, Douglas Salt, and Jagadeesh Yeluripati. 2019. Crossing the chasm: a 'tube-map' for agent-based social simulation of policy scenarios in spatially-distributed systems. *Geoinformatica*, 23, 169–199. https://doi.org/10.1007/s10707-018-00340-z

[4] Dongbin Zhao, Yujie Dai, and Zhen Zhang. 2012. Computational Intelligence in Urban Traffic SignalControl: A Survey. *IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews*, 42, 4, 485-493. https://doi.org/10.1109/TSMCC.2011.2161577

[5] Michael Batty. 2005. Agents, Cells, and Cities: New Representational Models for Simulating Multiscale Urban Dynamics. *Environment and Planning A: Economy and Space,* 37, 8, 1373-1394. https://doi.org/10.1068/a3784

[6] Andrew Crooks, Christian Castle, and Michael Batty. 2008. Key Challenges in Agent-Based Modelling for Geo-Spatial Simulation. *Computers, Environment and Urban Systems*, 32, 6, 417–430. https://doi.org/10.1016/j.compenvurbsys.2008.09.004

[7] Ryan Fox Squire. 2019. What About Bias in the SafeGraph Dataset? *SafeGraph*. Retrieved September 7, 2022, from https://www.safegraph.com/blog/what-about-bias-in-the-safegraph-dataset

[8] SafeGraph. 2019. Social Distancing Metrics [computer file]. SafeGraph [distributor], accessed at https://docs.safegraph.com

[9] U.S. Census Bureau. 2022. LEHD Origin-Destination Employment Statistics Data (2002-2019) [computer file]. Washington, DC: U.S. Census Bureau, Longitudinal-Employer Household Dynamics Program [distributor], accessed on December 1, 2021, at https://lehd.ces.census.gov/data/#lodes. LODES 7.0 [version]

[10] Chanwoo Jin, Atsushi Nara, Jiue-An Yang, and Ming-Hsiang Tsou. 2020. Similarity measurement on human mobility data with spatially weighted structural similarity index (SpSSIM). *Transactions in GIS*, 24, 104-122. https://doi.org/10.1111/tgis.12590

[11] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simmoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 4, 600–612. https://doi.org/10.1109/TIP.2003.819861

[12] Somayeh Dodge. (2021). A Data Science Framework for Movement. *Geographical Analysis*, 53, 1, 92–112. https://doi.org/10.1111/gean.12212

[13] Alison Heppenstall, Andrew Crooks, Nick Malleson, Ed Manley, Jiaqi Ge, and Michael Batty. 2021. Future Developments in Geographical Agent-Based Models: Challenges and Opportunities. *Geographical Analysis*, 53, 1, 76–91. https://doi.org/10.1111/gean.12267

[14] Paul M. Torrens. 2010. Agent-based Models and the Spatial Sciences. *Geography Compass*, 4, 5, 428–448. https://doi.org/10.1111/j.1749-8198.2009.00311.x