Evaluating Spatial Dependency in Regional Similarities of the Population Dynamics Foundation Model

Hongyu Zhang[1] and Grant McKenzie[2]

[1] Department of Earth, Geographic, and Climate Sciences, University of Massachusetts Amherst, Amherst, United States
[2] Department of Geography, McGill University, Montreal, Canada

Correspondence: Hongyu Zhang (honzhang@umass.edu)

Submission Type. Analysis; Case Study.

BoK Concepts. Spatial dependency; Geography as a foundation for GIS; Artificial intelligence in EO and GI.

Keywords. spatial dependency, regional similarity, foundation model

Abstract

Artificial intelligence (AI) is a trending topic in GIScience, and spatially explicit AI techniques are increasingly sought after (Janowicz et al., 2020). In November 2024, Google released its Population Dynamics Foundation Model (PDFM) embeddings to support various spatial prediction tasks (Agarwal et al., 2024). The company claims that these embeddings can be used similarly to census data and socioeconomic statistics by aggregating search trends, geospatial data, busyness, weather, and air quality data at county and ZIP code levels. However, the embeddings lack explainability, as users do not know what each feature represents. How useful are these embeddings? Do they merely capture the effects of spatial dependency, or do their features reveal patterns beyond spatial relationships?

Previous research has examined regional similarities from various perspectives. Adams (2015) proposed an observation-to-generalization place model for identifying similar places. Wang et al. (2022) studied cultural semantic similarity using place names in mainland China. McKenzie and Romm (2021) compared regional similarities in Berlin and Stockholm using mobility signatures. In this exploratory study, we applied research methods in regional similarity to evaluate the usefulness of PDFM embeddings.

We calculated the pairwise similarity between all counties in the contiguous United States. The Jensen-Shannon Distance (JSD) was used to represent regional similarities. In geography, JSD has been used for land use modeling (Niesterowicz and Stepinski, 2016) and socioeconomic comparisons (McKenzie et al., 2024). The JSD equation is presented in Equation (1) where C_A and C_B are vectors of PDFM embeddings for two different counties. JSD is based on the Kullback–Leibler divergence, as shown in Equation (2). Here, $M = \frac{1}{2}(C_A + C_B)$, and x represent an individual attribute value.

$$JSD(C_A \parallel C_B) = \sqrt{\frac{D(C_A \| M) + D(C_B \| M)}{2}} \quad (1)$$

$$D(C_A \parallel M) = \sum_{x \in \chi} C_A(x) \log\left(\frac{C_A(x)}{M(x)}\right) \quad (2)$$

Before calculating JSD, we added a small ε (0.17) to all features in the PDFM embeddings to ensure they were positive. We then constructed a JSD matrix for pairwise comparisons of U.S. counties. To facilitate analysis, we designed an interactive, exploratory spatial analysis tool[1] using HTML, CSS, and JavaScript (Figure 1). Researchers can click on any county to view the corresponding JSD values of all other counties on a choropleth map. Hovering over a county displays metadata along with the JSD value. In Figure 1, District of Columbia (D.C.) is selected. The most similar counties are located around D.C., indicating strong spatial dependency. Experimenting with other counties yielded

[1] The interactive tool and its source code are available at https://github.com/hzhangic/regional-similarity.

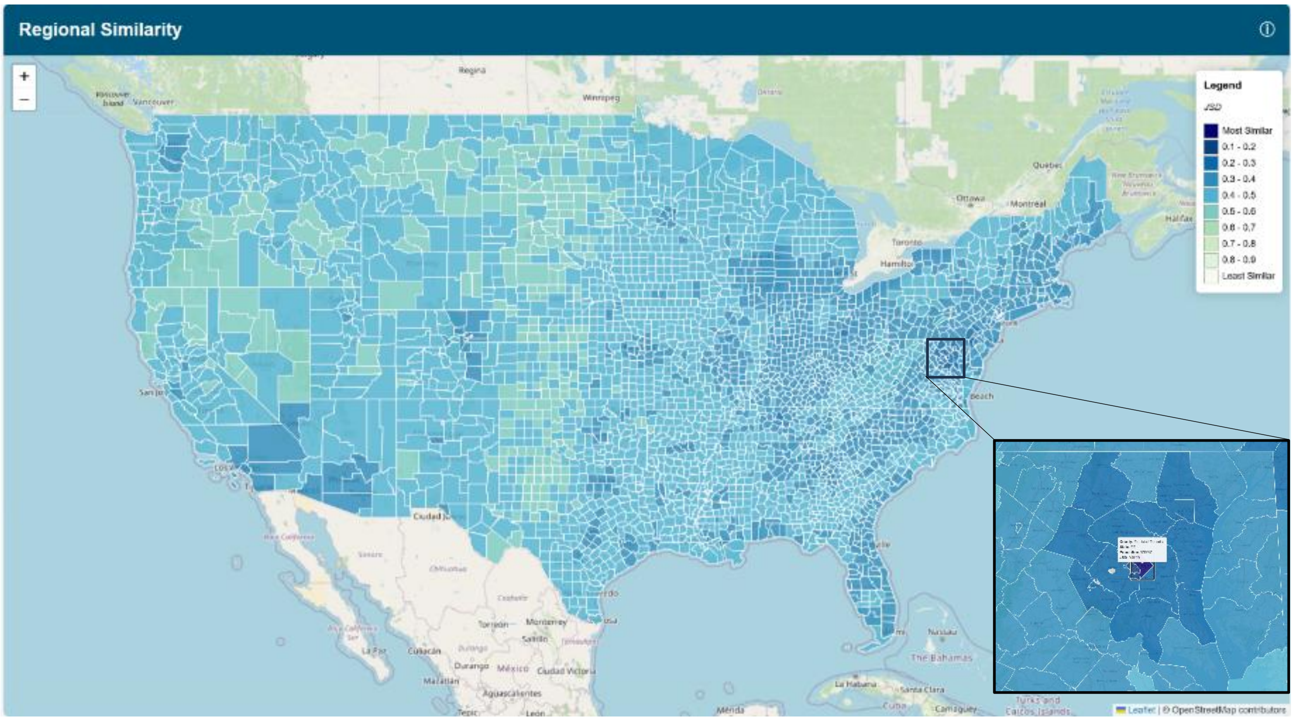

Figure 1. A screenshot of the regional similarity web tool developed for exploratory spatial analysis.

similar results. Thus, we hypothesize that spatial proximity plays a significant role in the underlying structure of PDFM embeddings.

To validate our hypothesis, we conducted the following analysis. For each county, we sorted the pairwise JSD values and selected 10 counties with the smallest JSD values greater than 0. These represent the top 10 most similar counties to the home county. We then calculated the percentage of the top 10 most similar counties that are located within the same state as the home county. For states with fewer than 10 counties, the maximum number of selected counties was adjusted accordingly (e.g., 8 for Connecticut, 5 for Rhode Island, 3 for Delaware, with D.C. excluded). The national percentage was 82%, indicating that spatial dependency plays a significant role in determining regional similarities using PDFM embeddings. Detailed results are visualized in Figure 2. The results show that, for most states, the most similar counties are within their own state, including counties in the Great Lakes region, the South, and the Southeast. Exceptions were observed in the western United States (excluding California and Montana) and New England. For counties in these regions, the PDFM embeddings appear to capture unique features that identify similar counties beyond state boundaries.

One question that arose was whether similar counties that are less spatially dependent are located in states with fewer counties. To address this, we used ordinary least squares (OLS) regression to analyze the relationship between the number of counties per state and the average percentage of the top 10 most similar counties within each state. The overall R-squared value was 0.42, indicating

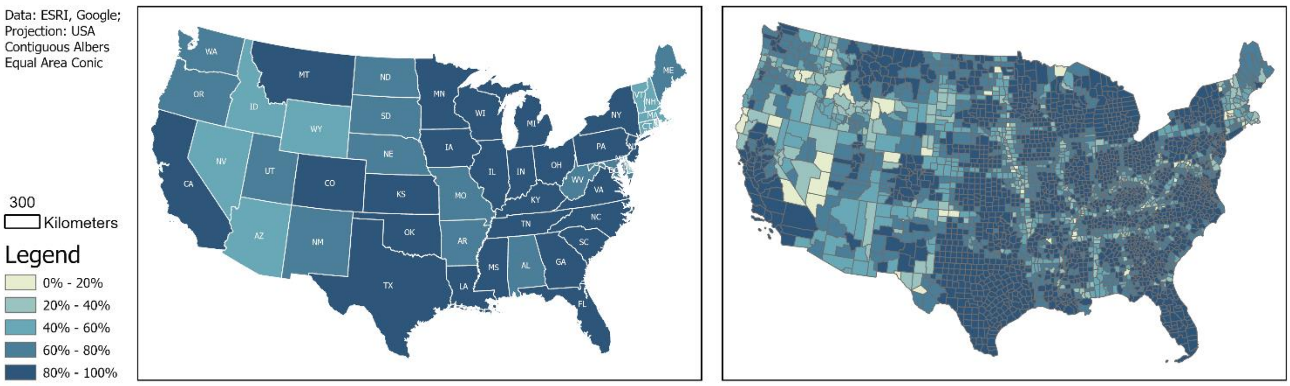

Figure 2. Percentages of the most similar counties within their own state (left: state-level; right: county-level).

that 42% of the trend can be explained by the number of counties per state. This suggests that other factors also contribute to the observed patterns.

In conclusion, this preliminary study examined the utility of PDFM embeddings in capturing regional similarities and assessed whether these embeddings reflect more than just spatial dependency. Our analysis indicates that, while counties with the smallest JSD values are generally located within the same state, which demonstrates the influence of spatial proximity, there are notable exceptions in regions such as the Western United States and New England. In these areas, the top 10 most similar counties often cross state boundaries, suggesting that the embeddings are capturing shared characteristics in search trends, location-based indicators, human activity levels, and climate conditions in addition to simple spatial relationships. These findings suggest that PDFM embeddings are not merely encoding spatial dependency but can reveal deeper patterns in specific regions. Further analysis is needed to explain the nature of these similarities. The exploratory spatial analysis web tool developed in this study facilitates this by enabling users to identify and visualize highly similar counties, which supports the investigation in future work.

References

Adams, B.: Finding similar places using the observation-to-generalization place model, Journal of Geographical Systems, 17, 137–156, https://doi.org/10.1007/s10109-015-0209-3, 2015.

Agarwal, M., Sun, M., Kamath, C., Muslim, A., Sarker, P., Paul, J., Yee, H., Sieniek, M., Jablonski, K., Mayer, Y., Fork, D., de Guia, S., McPike, J., Boulanger, A., Shekel, T., Schottlander, D., Xiao, Y., Manukonda, M. C., Liu, Y., Bulut, N., Abu-el-haija, S., Eigenwillig, A., Kothari, P., Perozzi, B., Bharel, M., Nguyen, V., Barrington, L., Efron, N., Matias, Y., Corrado, G., Eswaran, K., Prabhakara, S., Shetty, S., and Prasad, G.: General Geospatial Inference with a Population Dynamics Foundation Model, arXiv preprint, arXiv:2411.07207, 2024.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., and Bhaduri, B.: GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, International Journal of Geographical Information Science, 34, 625–636, https://doi.org/10.1080/13658816.2019.1684500, 2020.

McKenzie, G., Battersby, S., and Setlur, V.: MixMap: A user-driven approach to place-based semantic similarity, Cartography and Geographic Information Science, 51, 583–598, https://doi.org/10.1080/15230406.2023.2176930, 2024.

McKenzie, G., and Romm, D.: Measuring urban regional similarity through mobility signatures, Computers, Environment and Urban Systems, 89, 101684, https://doi.org/10.1016/j.compenvurbsys.2021.101684, 2021.

Niesterowicz, J., and Stepinski, T. F.: On using landscape metrics for landscape similarity search, Ecological Indicators, 64, 20–30, https://doi.org/10.1016/j.ecolind.2015.12.027, 2016.

Wang, H., Zhang, H., Jiang, S., Tang, G., Zhang, X., and Zhou, L.: City association pattern discovery: A flow perspective by using cultural semantic similarity of place name, Applied Geography, 139, 102629, https://doi.org/10.1016/j.apgeog.2021.102629, 2022.