Measuring Place Function Similarity with Trajectory Embedding
Cheng Fu, Robert Weibel
Department of Geography, University of Zurich

1. Abstract
Modeling place functions from a computational perspective is a prevalent research topic. The technology of embedding enables a new approach that allows modeling the function of a place by its chronological context as part of a trajectory. The embedding similarity was previously proposed as a new metric for measuring the similarity of place functions, with some preliminary results. This study explores if this approach is meaningful for geographical units at a much smaller geographical granularity compared to previous studies. In addition, this study investigates if the geographical distance can influence the embedding similarity. The empirical evaluations based on a big vehicle trajectory data set confirm that the embedding similarity can be a metric proxy for place functions. However, the results also show that the embedding similarity is still bounded by the distance at the local scale.
2. Introduction

2.1. Background
In location-based applications, the function of places, such as the socio-economic function or land use, are commonly surveyed and labeled by predefined hierarchical taxonomies, for instance, in the form of commercial or user-contributed points-of-interest (POIs), or modeled by the spatio-temporal activity patterns per se. Recent studies started to involve contextual information of the target place for place function modeling as well. For example, the concept of interactional regions (Kempinska et al. 2018) emphasizes the human movements between places as interaction for defining functions. Embedding models, and Word2Vec (Mikolov et al. 2013) in particular, have also been adapted as a useful prototype for modeling complex geographical or behavioral contextual information. The Word2Vec model aims to model a word’s position in an embedding space by its preceding and succeeding words in sentences of a text corpus. In the embedding space, words with similar semantics are expected to have closer distances. One main application of the embedding modeling in GIScience is to model the similarity of POI categories, such as, whether cafes are more similar to restaurants than to hotels in general. The Place2Vec model (Yan et al. 2017) uses the neighboring POIs of the target POI as the spatial context for measuring the similarity of POI categories. A more recent perspective is to model the similarity of individual POIs. The Move2Vec model (Crivellari and Beinat 2019) takes the consecutive places in a movement as the context and provides the evidence that the function of places can be inferred by their preceding and succeeding neighbors in the trajectories.

This study conceptualizes the chronological neighbors of a place in trajectories as its context and explores if the embedding contextual space can measure place similarity. We examine if the embedding model also works at a finer spatial granularity compared to previous studies. We further use POIs as a reference for validating the results. We also explore if spatial proximity still has an impact on the output of this place modeling method, even though no explicit spatial information is involved in the processing.

2.2. Related work and research gaps
As briefly introduced in the previous section, one main perspective of applying the embedding models to place modeling is to understand the similarity of POI categories, which can be denoted as spatial-context embedding. Typically, POIs with the same category in an existing POI data set are treated as an identical word, and neighboring POIs within a buffering distance or the same polygonal segment are used as the context to feed into the Word2Vec model or its variants. From the embedding space, it is learned whether one POI category is more similar to another POI category (Yan et al. 2017, Yao et al. 2017, Zhai et al. 2019, Liu et al. 2020). The first perspective requires places being already labeled, which may be limited by the correctness of the employed POI database concerning the completeness and the frequency of updating. In addition, the first perspective assumes that two POIs with the same category should have the same social function. However, even the same categories of POIs may have very different activities.

The second perspective of applicants is to treat each place as a unique entity, and trajectories formed by the places are fed into Word2Vec or its variants, which can be termed trajectory embedding. The preceding and succeeding places in the same trajectory thus are used as the context of a place for embedding. Solomon et al. (2018) use the similarity of stay points learned by embedding from users’ GPS trajectories as features for classifying demographic profiles of users. However, this study uses the embedding space of the places as a black box for feature engineering but does not explore the meanings of the embedding results in geography. Crivellari and Beinat (2019) demonstrate on some examples that the similarity of places in the embedding space (denoted as embedding similarity) is associated with the similarity of social functions using call detail records (CDRs) as the input data. However, as Crivellari and Beinat (2019) use the service area of cellular towers as the geometry of places, the places correspond to very different
geographical granularities, leading to a mixture of social functions represented in a particular cell, due to the nature of the spatial layout of the cellular towers.

2.3. Research questions
Previous studies have shown insights that the trajectory embedding has the ability to model the social function of places by movements. The embedding similarity of individual places has the potential to measure the similarity of social functions. In this study, we would like to address two additional questions based on previous studies:

RQ1: Can the similarity in the trajectory embedding space measure the similarity of social function for individual places at a fine geographical granularity?
RQ2: Does the spatial proximity, particularly the distance decay, still influence the embedding similarity?

3. Data
Two main data sets were used. The GPS trajectory data set is provided by a fleet management service company based in Greece. The data set has GPS-waypoint trajectories of 5389 vehicles over one month from June 1st, 2017 to June 30th, 2018. The majority of the vehicles are professional vehicles, such as trucks, vans, and buses. The trajectories are mainly located in Greece, but also cover continental Europe. The second data set is a POI data set consisting of OpenStreetMap point features with their amenity labels covered by three top categories in the Geofabrik taxonomy: places of worship, POI (mainly referring to commercial POIs), and transport and traffic, which contains 7,600,000 in total. This POI data set is used as a reference for validating the results from the embedding method.

4. Methodology
The study consists of two main stages: the place modeling (Figure 1) and the analytics performed on the outputs of the place modeling (Figure 2). For the first stage, the waypoint records of a vehicle over a day are identified as an independent trajectory. The stop-move model is chosen for trajectory modeling, as the stops are usually associated with activities interacting with places. Only the waypoints identified as stops are remained after applying a stop-detection algorithm that can deal with noisy waypoints (Xiang et al. 2016). The stop waypoints of a trajectory then are geocoded to a 30-m regular grid tessellation. The grids are then coded by Morton code (Morton 1966) to make sure that each grid has a unique ID. If consecutive stop waypoints are located in the same grid, only one grid is used for representing these waypoints. After the preprocessing, each GPS waypoint trajectory is then converted to a sequence of grid IDs. As an analog, each grid is understood as a unique word, and a trajectory as a sentence.

![Figure 1 Overview of the place modeling workflow](image)

Essentially, Word2Vec is a two-layer neural network that aims to represent a word by a numeric vector after a process either using a target word to predict its neighboring words in a sentence, i.e., the skip-gram approach, or using the neighboring words to predict a target word, i.e., the continues-bag-of-word (CBOW)
approach. The learned vectors then form an embedding space of the words. Therefore, the converted trajectories can also be input to the Word2Vec model. In this empirical evaluation, the skip-gram approach is used for prediction in the Word2Vec model. Cosine similarity is used for measuring the embedding similarity of the grids.

Figure 2 The framework of analytics

On the second stage of analytics, the OSM POIs are associated to the same 30-m grid tessellation as the waypoints. If there is only one POI in a grid cell, the label of the cell is assigned as the category of the POI. If there is more than one POI contained in the same grid, the cell is labeled as mixed.

To answer RQ1, we explore the embedding outputs by two tasks: Task 1 explores if an individual grid cell and its nearest neighbors share the same POI labels by geo-visual analytics. The selected grid and its most similar grids are visualized and investigated on the map. Task 2 explores if the cells with the same POI category are also close in the embedding space. The pairwise embedding similarities of cells with the same POI category are calculated. The embedding similarities of the selected cells and other types of POIs are also calculated. A t-test is applied to check if the two embedding similarity sets are significantly different. In general, we shall expect that the intra-category similarity should be higher than the inter-category similarity if the embedding similarity is indeed able to measure the similarity of social functions.

To answer RQ2, we also perform two tasks. Task 3 is a quantitative approach to explore how the pairwise embedding similarity of the grid cells, both overall and by POI categories, may decay by geographical distance using linear regression fitting. Task 4 explores how the pairwise embedding similarity of the grid cells may vary by distance using an empirical variogram. In the empirical evaluation, Euclidean distance is used to represent the geographical distance.

5. Evaluations and Results

For the stop-detection setting, a stop is defined to last more than five minutes. 960,322 30-m grid cells were then detected containing one or more stops. Trajectories were composed of 6.6 cells containing stops on average (stddev = 5.2). For the Word2Vec model, the maximum distance for defining a neighbor was set to five. The minimum visitation requirement for the grid cells was set to five visits. The embedding vector size was set to 20. Eventually, 237,822 grid cells were modeled in the final embedding space, within which 14,745 were associated with at least one POI.

For Task 1, several randomly selected grid cells were investigated, e.g., a cell with a kiosk associated in Figure 3. It can be observed that the k-nearest neighbors (with k = 5) do share similar social functions with many target cells, e.g., convenience and pharmacy are similar to kiosks that all belong to the retailing sector. Even the park cell in Figure 3.B actually has a convenience store and a pharmacy right next to it, as found by additional investigation on Google Maps. Therefore, some mismatching may be caused by position inaccuracy and the lower coverage of the OSM POI data set. It can also be observed that distance is not a dominant bound for this case as the cells sharing high cosine similarity are not nearby. However, we also observe some degree of locality because all the neighbors are located within the city.
Figure 3 A selected sample of grid cells located near Athens, Greece with POI category kiosk and its ten most similar cells in the embedding space. A: The locations of the grid cells. The target cell is colored in red. Cells with a marker are those with a POI associated. B: The similarities of the cells, their associated POI categories, and their distances to the target cells.

For Task 2, it can be observed that cells with the same POI category are significantly more similar within the category than to other types of POIs, except for the fuel stations (Table 1).

Table 1 The mean embedding similarity of POIs with the selected categories and the mean embedding similarity to other POIs. POI code and category is based on the Geofabrik taxonomy. The default sample size for randomly selected POIs with the same category is 300, unless the total number of POIs is less than 300. The number of other POIs always matches the sample size of the selected POIs.

| POI code & category | Sample size | Intra-category mean similarity | Mean similarity to other POIs | T-test |
|---------------------|-------------|-------------------------------|------------------------------|--------|
| 2101 pharmacy       | 300         | 0.39                          | 0.37                         | 16.53  |
| 2501 supermarket    | 300         | 0.39                          | 0.37                         | 24.88  |
| 2301 restaurant     | 300         | 0.37                          | 0.36                         | 5.54   |
| 2511 convenience    | 300         | 0.39                          | 0.37                         | 12.19  |
| 2562 car repair     | 100         | 0.37                          | 0.36                         | 2.39   |
| 5250 fuel           | 300         | 0.33                          | 0.34                         | -15.54 |
| 2305 bar            | 100         | 0.42                          | 0.36                         | 14.67  |
| 5260 parking        | 300         | 0.36                          | 0.33                         | 34.07  |

For Task 3, 3,000 grid cells in the embedding space were randomly selected. The sampled cells are scattered over Europe, mainly in Greece, Italy, and France (Figure 4.A). Pairwise cosine similarities and Euclidean distances were calculated, respectively. Overall, the cosine similarity decreases with increasing geographical distance (Figure 4.B), which fits the linear regression model $CS = -4.15 \times 10^{-6} \times D + 0.40$, where CS stands for the cosine similarity and D stands for the Euclidean distance. However, it can be observed that the trends of the short distance and long distance are quite different. For the cosine similarity within the local range (defined as $D < 50,000$ m), the distance decay is much stronger (Figure 4.C), which fits the model $CS = -1.19 \times 10^{-6} \times D + 0.81$. However, for the very long distance relationships, the cosine similarity actually increases with increasing distance (Figure 4.D), which roughly fits the model $CS = 9.79 \times 10^{-9} \times D - 0.65$. Therefore, it can be inferred that the similarity of the grids after the trajectory embedding is influenced by the distance and the amenity of the place at different scales. At the local level, the distance has a strong influence, as its slope is three orders of magnitude larger than the slope of the overall regression model, although the absolute value actually is still small. The positive value of the fitted model for at the long distances indicates that the embedding similarity does reflect the similarity of the social function of the places. This might be the evidence that two places far apart may still provide similar amenities, e.g., a restaurant in Paris, France and a restaurant in Athens, Greece both serve food, regardless of the geographical distance between them.
Figure 4 A: The pairwise cosine similarities against the corresponding pairwise Euclidean distances of the 3,000 sample grid cells. B: The locations of the 3,000 sample cells. C: The linear regression model fitted to the cosine similarity values with distances within 50,000 m. D: The model fitted to the cosine similarity values with distances larger than $10^8$ m.

For exploring the experimental variograms, the same sample of 3,000 cells as used for Task 3 was employed. The bin bandwidth was set to 500 m, and the distance cutoff was set to 100 km. In addition, the variograms of selected POI categories were also calculated based on 200 random samples (Figure 5). It can be observed that the variance of cosine similarity increases with distance in general and for many categories such as supermarket and convenience stores, showing a sense of locality (Figure 5.A-D). There are exceptions, such as fuel and bar, where the variance appears to (slightly) decrease with distance. However, it should be noted that all slopes are very small, and the slope values are also on the same scale, no matter if it is an upward or downward trend.

Figure 5 Variograms of cosine similarity against pairwise distances for sample grid cells. A: all POI categories; B: supermarket; C: restaurant; D: convenience; E: fuel; F: bar.
6. Conclusion and Future Work

Given the results of the preliminary evaluation by Task 1 and Task 2, RQ1 can be answered that the trajectory-based embedding result can reflect the similarity of social function in real life. It is observed that if grids have the same POI associated, they mostly like have a higher embedding similarity as well. Based on the observation of outputs from Task 3 and Task 4, it can be concluded that the First Law of Geography still matters to the embedding similarity, even though the process of trajectory embedding does not model the distance explicitly. A possible reason might be that most vehicles still have a relatively limited area to move that the embedding model still learned some local knowledge. However, this hypothesis needs further exploration.

In summary, the embedding similarity is a result of locality and the similarity of social function. The advantage of embedding similarity as a metric is that it does not require prior hard-coded categories for POIs but may still provide semantically meaningful place recommendations given a source place as the preliminary workflow of Task 1. This actually sheds light on a task for recommending nearby and similar alternatives for a place required by a customer of a location-based service. We would like to further explore such a possibility.

Reference

Crivellari, A. and Beinat, E., 2019. From Motion Activity to Geo-Embeddings: Generating and Exploring Vector Representations of Locations, Traces and Visitors through Large-Scale Mobility Data. ISPRS International Journal of Geo-Information, 8 (3), 134.

Kempinska, K., Longley, P., and Shawe-Taylor, J., 2018. Interactional regions in cities: making sense of flows across networked systems. International Journal of Geographical Information Science, 32 (7), 1348–1367.

Liu, K., Yin, L., Lu, F., and Mou, N., 2020. Visualizing and exploring POI configurations of urban regions on POI-type semantic space. Cities, 99 (January), 102610.

Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. Nips, 1–9.

Morton, G.M., 1966. A computer oriented geodetic data base and a new technique in file sequencing. Ottawa, Canada.

Solomon, A., Bar, A., Yanai, C., Shapira, B., and Rokach, L., 2018. Predict Demographic Information Using Word2vec on Spatial Trajectories. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization - UMAP ’18. New York, New York, USA: ACM Press, 331–339.

Xiang, L., Gao, M., and Wu, T., 2016. Extracting stops from noisy trajectories: A sequence oriented clustering approach. ISPRS International Journal of Geo-Information, 5 (3).

Yan, B., Janowicz, K., Mai, G., and Gao, S., 2017. From ITDL to Place2Vec. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL’17. New York, New York, USA: ACM Press, 1–10.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., and Zhang, J., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. International Journal of Geographical Information Science, 31 (4), 825–848.

Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.-R., and Gu, C., 2019. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. Computers, Environment and Urban Systems, 74 (December 2018), 1–12.