Predicting visit frequencies to new places

Nina Wiedemann[1]
Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Ye Hong
Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Martin Raubal
Institute of Cartography and Geoinformation, ETH Zürich, Switzerland

Abstract

Human mobility exhibits power-law distributed visitation patterns; i.e., a few locations are visited frequently and many locations only once. Current research focuses on the important locations of users or on recommending new places based on collective behaviour, neglecting the existence of scarcely visited locations. However, assessing whether a user will return to a location in the future is highly relevant for personalized location-based services. Therefore, we propose a new problem formulation aimed at predicting the future visit frequency to a new location, focusing on the previous mobility behaviour of a single user. Our preliminary results demonstrate that visit frequency prediction is a difficult task, but sophisticated learning models can detect insightful patterns in the historic mobility indicative of future visit frequency. We believe these models can uncover valuable insights into the spatial factors that drive individual mobility behaviour.

2012 ACM Subject Classification Information systems → Geographic information systems; Computing methodologies → Neural networks; Applied computing → Transportation; Information systems → Location based services

Keywords and phrases Human mobility, Visitation patterns, Place recommendation, Next location prediction

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.84

Category Short Paper

Supplementary Material *Software (Source code)*: `https://github.com/mie-lab/predict-visits` archived at `swh:1:dir:c0c080878ee26ac806daac00fd25458dbfeb5406`
Text (Implementation details): `https://github.com/mie-lab/predict-visits/blob/main/supplementary_information.pdf`, archived at `swh:1:cnt:5eb98f0df940d22a9e3b0a1dbf883bef1b029688`

1 Introduction

Large-scale tracking data collected from mobile phone users are crucial for location-based services such as place recommendations [8]. One field of research is the so-called next location[2] prediction, which is concerned with finding the immediate next location an individual will visit [7]. Such predictions could be used for recommendations, navigation advice or on-demand transport services. The developments in this field, however, suffer from the heavy-tailed distribution of visit frequencies [3]; i.e., many locations are visited only once and are thus difficult if not impossible to predict [11]. Specifically, Cuttone et al. [1] find that 70% of locations are visited only once, and 20-25% of the visits are to new locations. The interest of users in these locations is primarily assessed upfront via recommendation systems that

[1] Corresponding author
[2] Since the term ”place” describes the subjectively experienced form of a geographic location [19], we will use the term ”location” throughout the paper to objectively denote a user's activity cluster.

leverage insights from aggregated user behaviour, e.g., the general popularity of a place. Many systems were developed for this purpose, mainly based on data from location-based social networks (LBSN), and employ (context-aware) collaborative filtering [9, 2, 18]. For a recommender system to successfully suggest entirely new locations to the user, data from many users in the same region must be available, which is often hampered due to the sensitive nature of tracking data [4].

At the same time, the mobility of a *single* user already allows one to draw insights about a user's interest in new locations. For example, the spatial layout [10] and topology of the mobility behaviour [20, 16], or the category frequencies of the user's previously visited locations [17] can help to estimate the spatial distribution of future visitation patterns [11]. In light of this possibility, we argue for a new problem formulation: *Predicting the frequency of future visits to a newly visited location, given the historic mobility of a single user.* In other words, assuming that we observe a user visiting a location for the first time, can we predict whether they will return to this location, in a scenario where knowledge about collective mobility patterns is scarce? We argue that this problem is mistaken as a subtask of recommender systems or next location prediction. In contrast, it requires special modelling approaches to learn efficiently from individual historic mobility patterns. Successful approaches could decide whether a location will become part of a user's activity set [6] and possibly unveil hidden patterns in the user's location preferences. Moreover, the gained knowledge will support the online next location prediction that needs to consider new locations at runtime, or improve individualized transport recommendation and planning.

In this paper, we formalize our new problem termed "visit frequency prediction", and present an approach to frame it as a supervised learning task. We experiment with self-attention-based and graph-based neural network models to efficiently process the historical tracking data. As expected, predicting the visit frequency to new locations is challenging due to the lack of information about the user's motives for visiting the location. Nevertheless, we find that neural network models can find patterns in the historic mobility that are predictive of future visits, improving over the baseline methods.

2 Problem formulation

Let $L^u = \{l_1^u, \ldots, l_m^u\}$ denote the set of all locations visited by user u. A location is defined by point coordinates or an area, where the user performed some stationary *activity* (e.g., working or catering). Locations can be derived, e.g., by clustering GNSS data or from check-ins to known POIs in LBSN. In practice, a user visits these locations sequentially, represented as a list S^u of n visit events, for example, $S^u = [l_2^u, l_1^u, l_2^u, l_4^u]$ with $n = 4$. The visit frequency $\nu(l)$ is thereby the number of visits to location l, e.g., $\nu(l_2^u) = 2$ in the example. Let $S_{i:j}$ be the excerpt of the chain from the i-th until the j-th element in S (excluding the j-th). We assume that at a specific point t, we observe that a new location $l_\theta^u \notin S_{1:t}^u$. The task is to predict $\nu(l_\theta^u)$ in $S_{t:n}^u$ given the historic mobility $S_{1:t}^u$.

One potential approach is to train a model to learn a mapping g such that, optimally, $\nu(l_\theta^u) = g(S_{1:t}^u, l_\theta^u, u)$, where the model could leverage 1) feature representation of the previously visited locations $f(l)$, $l \in S_{1:t}^u$ and the visit frequency $\nu(l)$ of these locations, 2) user characteristics u, and 3) features of the new location $f(l_\theta^u)$. Note that the model can be fitted to the data of many users, but, at inference time, it should be possible to apply the model to the data from a single user, potentially in a different geographical region.

3 Methods

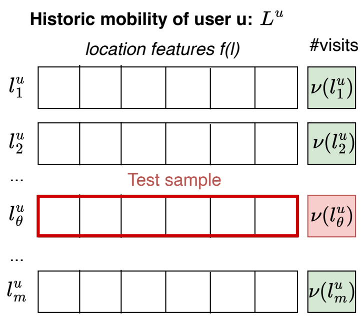


Figure 1 Approaching the visit frequency prediction problem as a supervised task.

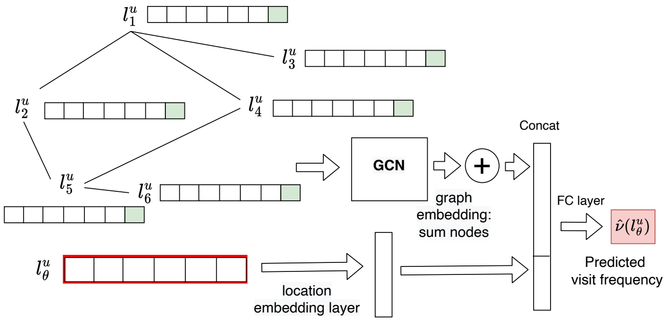


Figure 2 Graph-based model for learning visit frequency of new locations from the historic location graph. The graph is embedded and concatenated with the new location's features.

We propose a supervised approach to tackle the visit frequency prediction problem. In each training step, one location l_θ^u is removed from the user's overall tracking data (see Figure 1). The pruned mobility data $L^u \setminus l_\theta^u$ and $S^u \setminus l_\theta^u$ (i.e., the historic mobility, pretending that l_θ^u was never visited), as well as features of the removed location $f(l_\theta^u)$ are provided as input, and the visit frequency $\nu(l_\theta^u)$ is the desired output. We utilize the following features as f: The projected coordinates of l relative to the home location, the location purpose encoded as a one-hot vector, the average start hour of visits to l, and POI features. This leads to a vector of 24 entries. We implement a simple median and a k-nearest neighbor (kNN) approach as baselines and then test a fully connected neural network (MLP), a multi-head self-attention (MHSA) model, and a graph convolutional network (GCN) on the task. Each model is described in the following. For implementation details, see our code and supplementary material available at `https://github.com/mie-lab/predict-visits`.

The simple median baseline is given by $\hat{\nu}(l_\theta^u) = median(\{\nu(l^u) \mid l^u \in L^u\})$. This approach yields the same output for all queried locations of a user. For a more informed baseline, we consider a kNN approach, estimating the unknown visit frequency as $\hat{\nu}(l_\theta^u) = \frac{1}{k}\sum_{l \in N(l_\theta^u)} \nu(l)$, where $N(l_\theta^u)$ is the set of k nearest neighbors of l_θ^u in $L^u \setminus l_\theta^u$. We measure the distance between locations by the Euclidean distance of their feature vectors f.

For the MLP and the MHSA model, we provide a fixed set of m locations from the historic mobility of a single user, $L^u \setminus l_\theta^u$, and the new location l_θ^u as input. We hypothesize that the locations with the highest activity are most predictive of the visit frequency to new locations, and therefore select the m locations with the highest visit frequency. They are sorted by the frequency and are featurized by f, leading to an input matrix of size $(m+1)\times 24$. The matrix is flattened to be fed into the model. The MLP is a simple fully-connected two-layer network, whereas our MHSA follows the implementation by Hong et al. [5] for location prediction. A graph approach, on the other hand, allows for a variable number of input locations per user. Our approach is shown in Figure 2. The graph is passed through a Graph-Resnet [13], and the node embeddings are combined with average pooling, yielding a single vector of fixed size. This graph embedding is then concatenated with the embedding of the new location features $f(l_\theta^u)$ passed through a single layer. The last layer yields the estimated visit frequency.

4 Results

4.1 Data

We utilize high-quality and activity-labelled GNSS data from three tracking studies: Green Class 1 (GC1), Green Class 2 (GC2) [12] and yumuv [15]. All three studies were executed in collaboration with the Swiss Federal Railways (SBB) and aimed to evaluate the impact of Mobility-as-a-Service offers. The participants were tracked via a GNSS-based app and were asked to manually label their activities. The app already preprocesses the raw GNSS track points by inferring stationary *staypoints* and continuous movement *triplegs*, which are further processed with the Python library Trackintel [14]. Trackintel derives a set of visited locations from a user's tracking data using the DBSCAN clustering algorithm. After preprocessing, we included 139 users for GC1, 48 for GC2 and 653 for yumuv, who visited 104.5k, 35.7k and 127.3k distinct locations respectively. To align the tracking period, we split the data into time bins of three months. Finally, following Martin et al. [16], we transform the tracking data into a location graph for the GCN-based approach with the same hyperparameter setting. By the visit frequency prediction definition given above, the model should be applicable to unseen users in other geographic regions. Therefore, we split the data into train and test set on a *dataset*-level for the experiment; i.e., the train set $\mathcal{D}_{train}$ comprises randomly sampled data from the GC1 and yumuv studies, and $\mathcal{D}_{test}$ is sampled from GC2. To focus on rarely-visited locations, we only use locations that were visited up to ten times as test locations (l_θ). This cutoff on average excludes three locations per person.

4.2 Model comparison

Figure 3 shows the results for all tested models. We first consider the mean absolute error (MAE), i.e. $\frac{1}{|\mathcal{D}_{test}|}\sum_{l_\theta^u \in \mathcal{D}_{test}} |\hat{\nu}(l_\theta^u) - \nu(l_\theta^u)|$. The absolute error is generally low (around 1.8) for all models, and complex models only improve marginally over the baselines. However, the MAE is misleading due to the imbalance between the visit frequencies: Many locations are visited only once, whereas very few are visited ten times. For a more insightful evaluation, we propose to consider the balanced MAE: $\frac{1}{10}\sum_{i=1}^{10}\left(\frac{1}{|\mathcal{D}_{test}|}\sum_{l_\theta^u \in \mathcal{D}_{test} \text{ with } \nu(l_\theta^u)=i} |\hat{\nu}(l_\theta^u) - \nu(l_\theta^u)|\right)$

As Figure 3 (middle) shows, the balanced MAE is 3.99 for the simple median baseline and improves to 3.78 for the best kNN model. The neural network models yield a substantial improvement if they are also trained with *balanced* data (denoted by "bal." in Figure 3), meaning that the batches at train time were sampled such that each visit frequency from 1 and 10 appear equally often. The balanced GCN and balanced MHSA model yield the best performance with a balanced MAE of 2.43, indicating that these models can indeed learn patterns in historic mobility. The results for the balanced GCN are also visualized as a

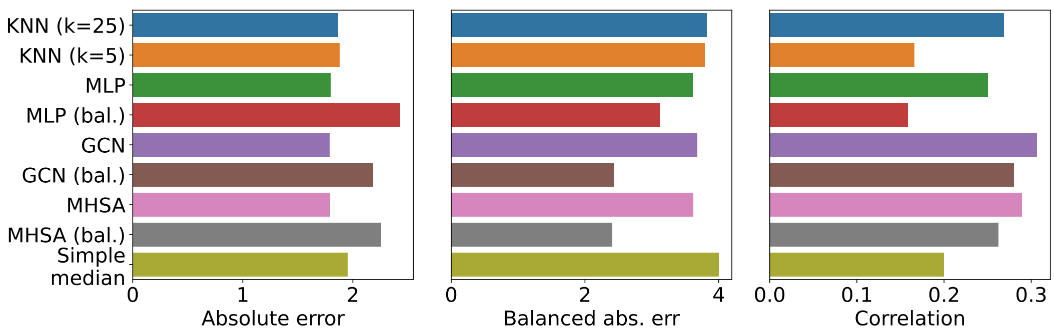
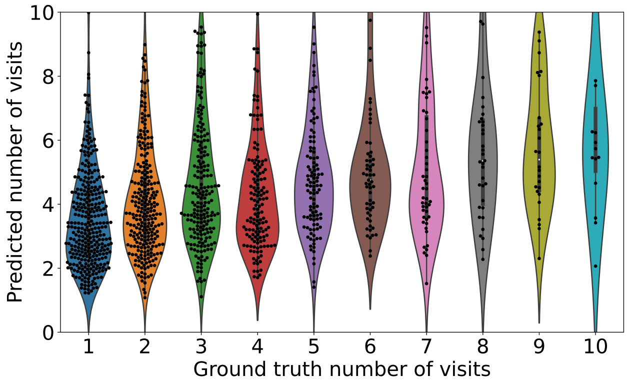


Figure 3 Model comparison on the visit frequency prediction problem for new users.



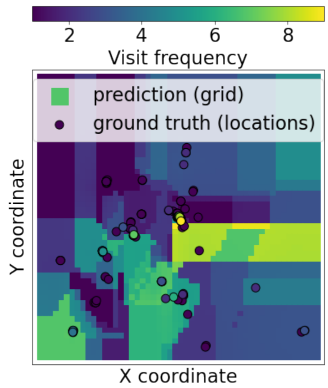


Figure 4 Violinplot of visit frequency predicted by the GCN compared to the ground truth.

Figure 5 Spatial distribution of predicted visit frequencies.

violin plot in Figure 4. While the test set is imbalanced and the predictions are very noisy, there is a clear shift in the distribution of predicted frequency with increasing ground truth visit frequency.

Finally, the Pearson correlation coefficient ρ of predicted and ground truth visit frequencies of the test data is shown in Figure 3 (right). The GCN and MHSA models again achieve the best performance with ρ up to 0.3. In general, the results indicate that predicting visit frequencies to newly visited locations is a difficult task. The value of the predicted frequencies for real applications is limited so far, even though they are more accurate than the baselines.

5 Discussion and outlook

The increasing availability of user location data gives rise to new research opportunities in the context of location recommendation and prediction. We have introduced a new problem that, for the first time, regards the importance of newly visited locations by approximating their projected visit frequency. Our preliminary results show that the task suffers from similar difficulties as next location prediction, namely noisy data, lack of information and inherent stochasticity in user decisions. The difficulty is also due to the strong imbalance of the ground-truth visit frequency. However, other models or additional context data may improve performance.

A well-trained visit frequency prediction model could also be applied to map the probability of visits to new locations. This analysis would yield insights into the spatial distribution of visit frequencies learnt by the model. An example is shown in Figure 5, where we systematically sampled locations within the convex hull of the visited locations of one user. The heatmap of predicted visits is based on hidden patterns detected in the ground-truth visit frequencies (dots, locations that are only visited once are filtered out for visibility). An analysis of the spatial visitation patterns, e.g., with respect to the spatial layout and distances of frequently visited locations, may improve the understanding of user behaviour. Thus, we believe that visit frequency prediction is an exciting endeavour, and we hope that our problem formulation and preliminary methodology inspire further research on this topic.

References

1 Andrea Cuttone, Sune Lehmann, and Marta C González. Understanding predictability and exploration in human mobility. *EPJ Data Science*, 7:1–17, 2018.

2 Tuan Hung Dao, Seung Ryul Jeong, and Hyunchul Ahn. A novel recommendation model of location-based advertising: Context-Aware Collaborative Filtering using GA approach. *Expert Systems with Applications*, 39(3):3731–3739, 2012.
3 Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
4 Linrui Han. Personal Privacy Data Protection in Location Recommendation System. In *Journal of Physics: Conference Series*, volume 2138, page 012026, 2021.
5 Ye Hong, Henry Martin, and Martin Raubal. How do you go where?: improving next location prediction by learning travel mode information using transformers. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22)*, 2022.
6 Ye Hong, Henry Martin, Yanan Xin, Dominik Bucher, Daniel J. Reck, Kay W. Axhausen, and Martin Raubal. Conserved quantities in human mobility: From locations to trips. *Transportation Research Part C: Emerging Technologies*, 146:103979, 2023.
7 Ye Hong, Yatao Zhang, Konrad Schindler, and Martin Raubal. Context-aware multi-head self-attentional neural network model for next location prediction. *arXiv preprint arXiv:2212.01953*, 2022.
8 Haosheng Huang, Georg Gartner, Jukka M Krisp, Martin Raubal, and Nico Van de Weghe. Location based services: ongoing evolution and research agenda. *Journal of Location Based Services*, 12(2):63–93, 2018.
9 Defu Lian, Yong Ge, Fuzheng Zhang, Nicholas Jing Yuan, Xing Xie, Tao Zhou, and Yong Rui. Content-Aware Collaborative Filtering for Location Recommendation Based on Human Mobility Data. In *2015 IEEE International Conference on Data Mining*, pages 261–270, 2015.
10 Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)*, pages 1043–1051, 2013.
11 Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. Approaching the Limit of Predictability in Human Mobility. *Scientific reports*, 3(1):2923, 2013.
12 Henry Martin, Henrik Becker, Dominik Bucher, David Jonietz, Martin Raubal, and Kay W. Axhausen. Begleitstudie SBB Green Class - Abschlussbericht. *Working Paper No. 1439, Institute for Transport Planning and Systems, ETH Zürich*, 2019.
13 Henry Martin, Dominik Bucher, Ye Hong, René Buffat, Christian Rupprecht, and Martin Raubal. Graph-ResNets for short-term traffic forecasts in almost unknown cities. In *NeurIPS 2019 Competition and Demonstration Track*, pages 153–163, 2020.
14 Henry Martin, Ye Hong, Nina Wiedemann, Dominik Bucher, and Martin Raubal. Trackintel: An open-source Python library for human mobility analysis. *Computers, Environment and Urban Systems*, 101:101938, 2023.
15 Henry Martin, Daniel Reck, Kay Axhausen, and Martin Raubal. Empirical use and impact analysis of MaaS. Technical report, ETH Zurich, 2021.
16 Henry Martin, Nina Wiedemann, Daniel J Reck, and Martin Raubal. Graph-based mobility profiling. *Computers, Environment and Urban Systems*, 100:101910, 2023.
17 Seyyed Mohammadreza Rahimi and Xin Wang. Location Recommendation Based on Periodicity of Human Activities and Location Categories. In *Advances in Knowledge Discovery and Data Mining: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '13)*, pages 377–389, 2013.
18 Sulis Setiowati, Teguh Bharata Adji, and Igi Ardiyanto. Context-based awareness in location recommendation system to enhance recommendation quality: A review. In *2018 International Conference on Information and Communications Technology*, pages 90–95, 2018.
19 Yi-Fu Tuan. *Space and place: humanistic perspective*. Springer, 1979.
20 Nina Wiedemann, Henry Martin, and Martin Raubal. Unlocking social network analysis methods for studying human mobility. *AGILE: GIScience Series*, 3:19, 2022.