Why We Go Where We Go: Profiling User Decisions on Choosing POIs Supplementary Material

Renjun Hu,[1*] **Xinjiang Lu,**[2*] **Chuanren Liu,**[3] **Yanyan Li,**[2] **Hao Liu,**[2] **Shuai Ma,**[1] **Hui Xiong**[2]
[1]SKLSDE Lab & BDBC, Beihang University, Beijing, China, {hurenjun, mashuai}@buaa.edu.cn
[2]The Business Intelligence Lab, Baidu Research, National Engineering Laboratory of Deep Learning Technology and Application, Beijing, China, {luxinjiang, liyanyanliyanyan, liuhao30}@baidu.com xionghui@gmail.com
[3]Department of Business Analytics and Statistics, University of Tennessee, Knoxville, TN, cliu89@utk.edu

Factor index	Factor name
0	user identifier
1–5	frequently-visited areas: business areas
6–10	frequently-visited areas: scenic spots
11–15	frequently-visited areas: malls
16–20	frequently-visited areas: sports places
21–25	frequently-visited areas: hospitals
26–30	frequently-visited areas: universities
31	POI identifier
32	POI category
33	POI brand
34	POI popularity
35	decision time (*i.e.,* hour)
36	distance to home
37	distance to work
38	distance to POI

Table 1: Summary of decision factors on BEIJING

Appendix A: Detailed Factors

The detailed factors for profiling user decisions on BEIJING and NYC are summarized in Tables 1 and 2, respectively. Note that we keep the top-5 areas/POIs for each frequently-visited type factor and add "null" placeholders if the numbers of valid areas/POIs are less than 5. These frequently-visited type factors are helpful to capture the complementary patterns between POIs and areas. As stated in the main body of our paper, we discretize the continuous POI popularity into six levels and discretize distance into five levels. Thus, all our factors are concrete and explainable items for which we learn hidden representations. Finally, we profile each user decision with a subset of explainable factors.

Appendix B: Popularity and Distance Discretization

We now present justification for our discretization on the continuous POI popularity and distance.

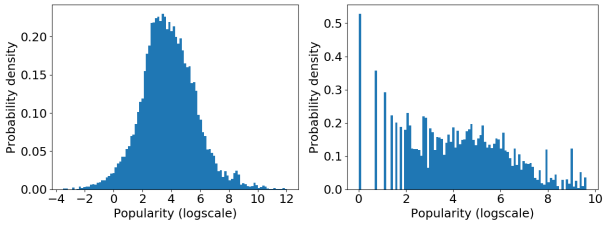
The statistics of POI popularity on both data sets are illustrated in Fig. 1. As can be seen, the log-scaled POI popularity is normal-distributed in general. On NYC, the probability concentrates on specific values for low POI popularity, *i.e.,* left part of Fig. 1(b). This is due to precision limits. However, the overall shape is still gaussian. Inspired, we discretize POI popularity by standard scores (z-scores). Assuming that the mean and standard derivation of the normal distribution are μ and σ, respectively, the z-score of a log-scaled POI popularity p is given by:

$$z(p) = \frac{p - \mu}{\sigma}. \quad (1)$$

The standard derivation σ is about 5.0 and 3.6 on BEIJING and NYC, respectively. That is, the corresponding z-scores

Factor index	Factor name
0	user identifier
1–5	frequently-visited POIs: education
6–10	frequently-visited POIs: sports
11–15	frequently-visited POIs: health care
16–20	frequently-visited POIs: shopping
21–25	frequently-visited POIs: food
26–30	frequently-visited POIs: leisure
31–35	frequently-visited POIs: service facility
36–40	frequently-visited POIs: transport
41	POI identifier
42	POI category
43	POI popularity
44	decision time (*i.e.,* hour)

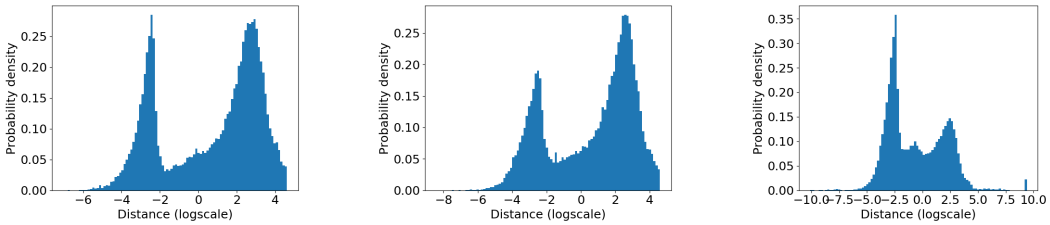
Table 2: Summary of decision factors on NYC



(a) BEIJING (b) NYC

Figure 1: Statistics of POI popularity (X-axis is log-scaled and its ticks do not have meanings.)

[*]Both authors contribute equally to this work.



(a) Distance to home (b) Distance to work (c) Distance to POI

Figure 2: Statistics of distance on BEIJING (X-axis is log-scaled and its ticks do not have meanings.)

are mainly within $[-1, 1]$ for both BEIJING and NYC. Combining the above, we discretize POI popularity into six levels based on z-scores z of log-scaled popularity: (i) *strongly unpopular* if $z \leq -1$, (ii) *unpopular* if $-1 < z \leq -0.5$, (iii) *weakly unpopular* if $-0.5 < z \leq 0$, (iv) *weakly popular* if $0 < z \leq 0.5$, (v) *popular* if $0.5 < z \leq 1$, and (vi) *strongly popular* if $z > 1$.

The statistics of log-scaled distance (to home, work, and POI) on BEIJING is illustrated in Fig. 2. Recall that distance factors are omitted on NYC as the decision locations are unknown. Different from POI popularity, the distance is captured by multimodal Gaussian distributions. Each distribution has two main peaks. Notably, the probability peaks occur at similar positions for the three distributions. In other words, the two-peak distance patterns are universal. For discretizing distance, the thresholds we choose are more of domain experience. More specifically, we discretize distance into five levels: the first *1km and less* is treated as walking distance, the second *between 1 and 3 km* is treated as bicycling distance, and the rest *between 3 and 7 km*, *between 7 and 15 km*, and *15 km and more* are treated as automobile distance. Finally, we note that the increment of thresholds is inspired by the log-scale characteristic.