Data Description

The document introduces the data of feature matrices and noise tensor used in paper [1] and [2]. Please cite the following two papers when using the datasets.

[1] Yu Zheng, Tong Liu, Yilun Wang, Yanchi Liu, Yanmin Zhu, Eric Chang. Diagnosing New York City's Noises with Ubiquitous Data. In Proceedings of UbiComp 2014.
[2] Wang, Y., Zheng, Y., Liu, T. A noise map of New York City. In Proc. of UbiComp 2014.

There are four data sources we use for diagnosing NYC's noises: 311 data about noise, user check-in data, road networks and POIs. First, we have segmented NYC map into 1199 disjointed regions according to its major roads. Each region is a basic geographic unit to study noises and features.

We modeled the noises using a tensor $\mathcal{A}$ with three dimensions denoting regions, noise categories and time slots, respectively. According to the 311 data, we extracted 14 noise categories. We divided a day into 24 equal time slots, one hour for each time slot. Therefore, there is $\mathcal{A} \in \mathbb{R}^{1199\times 14\times 24}$. An entry $\mathcal{A}(\mathrm{i,j,k})$ stores the total number of 311 complaints of category c_j in region r_i and time slot t_k over the given period of time (e.g., 168 weekdays or 68 weekends). A tensor is stored in ***A.mat***.

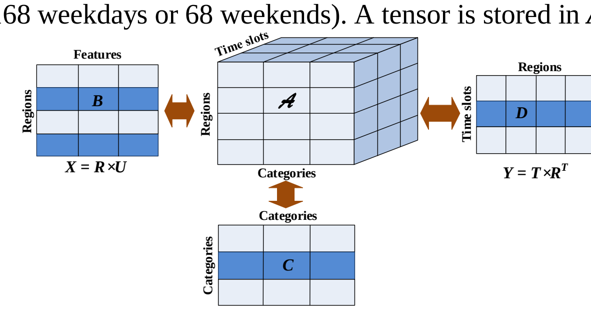

Figure 1. Tensor and feature matrices

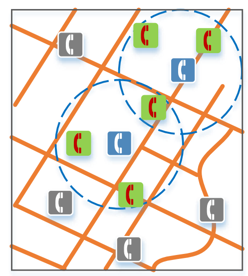

Figure 2. 311 complaints of a location

To deal with the data sparsity problem, we extract three categories of features: geographical features, the noise category correlation features, and human mobility features (denoted by matrices *B*, *C*, and *D*, as shown in Figure 1).

The geographical feature B $\in \mathbb{R}^{1199\times 22}$, is comprised of two parts: POI features and road network features. POI features, in the first 15 columns of *B*, are extracted from POIs falling in a region, consisting of the total number of POIs over 15 categories: *Entertainment & Arts, Vehicles, Business to Business, Computers, Education, Food & Dining, Government, Health & Beauty, Home & Family, Legal & Finance, Professional & Services, Estate & Construction, Shopping, Sports & Recreation, and Travel*. Road network features consist of the number of intersections and the total length of road segments in different levels ($\in [1,6]$), which are in the last 7 columns of *B*. Matrix *B* is stored in ***B.mat***.

The correlation between different noise categories $C \in \mathbb{R}^{14\times 14}$ can be learned from the 311 data itself. The complaints within a circle distance of 100 meters to a location is counted for the location, as shown in Figure 2. For each complaint of a noise category, we counted the number of complaints of each other category around it and summed together. An entry c_{ij} represents the total number of noise complaints in category j around noise complaints in category i. Matrix *C* is stored in ***C.mat***.

Human mobility features $D \in \mathbb{R}^{1199\times 24}$ are derived from check-ins created by users in different regions and time slots. An entry d_{ki} of matrix D denotes the number of check-ins generated in region r_i and time slot t_k. Matrix *D* is stored in ***D.mat***.

As weekdays and weekends have different noise patterns and features, we built a tensor and feature matrices for them separately. Each of file folders ***weekday*** and ***weekend*** encloses a tensor ***A.mat*** and three features ***B.mat***, ***C.mat*** and ***D.mat***.

Contact:

Dr. Yu Zheng, yuzheng@microsoft.com

Lead Researcher at Microsoft Research

http://research.microsoft.com/en-us/people/yuzheng/