Crawdad Efpl/Mobility Dataset

What is the dataset?

The efpl/mobility dataset is a dataset of mobility traces of taxicabs in San Francisco, CA. The dataset holds GPS information from tracking 500 cabs through the San Fransisco Bay Area over the course of 30 days in May 2008. The data records the latitude and longitude of the taxis at a sampling rate of about 1 data line per minute. The data lines also hold a boolean value indicating whether the taxi was occupied or unoccupied at the time of sampling and the timestamp in UNIX epoch format.

There are roughly 11,000,000 total data lines in the dataset, equating to about 20,000 per cab. About 5,000,000 of the total data lines correspond to "occupied" instances, where a rider was in the cab. There are 450,000 unique occupied trips encompassed in the dataset.

The total "occupied" trip length in the dataset is 1,500,000 miles. The average length for an occupied trip is 3.4 miles. The median trip length is 1.8 miles. The 25th and 75th percentiles for trip length are 1.0 and 3.1 miles, respectively. There are also 100,000 data point pairs for which the distance between adjacent pairs is over 1 mile (this could be relevant for snapping the data to a graph).

Who authored the dataset?

This dataset is distributed by Crawdad: a research community resource curated by Dartmouth University Computer Science that stores many wireless trace datasets (http://crawdad.org/index.html). This dataset, in particular, was provided by the Exploratorium, a science museum in San Francisco. The dataset can be found at (http://crawdad.org/epfl/mobility/20090224/), but one must request a username from the Crawdad administrators to gain access to it. The noted contributors to the dataset are Michal Pirkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser.

What others have done with the dataset

The contributors to the dataset use it in a paper concerned with modeling the mobility and formation of clusters in mobile wireless networks that possess regions of sparse connectivity as well as regions of dense connectivity. They use the dataset to show that certain macroscopic characteristics specific to clustered mobile wireless networks are prevalent in real mobility traces.

Crawdad notes that the dataset has been cited in 197 papers. One such paper uses the dataset to demonstrate how the k-nearest-neighbors algorithm can be used to extract valuable information from a vehicular GPS dataset. In that article, the authors use the dataset to relate speed of travel to time of day. Another paper uses this dataset to match raw GPS points to actual road networks, enriching the raw data with contextual information such as speed limits, changes in elevation, and attraction points. Another article uses the dataset (in conjunction with 2 other GPS datasets) to attempt to learn techniques for identifying individuals based off their GPS movements.

Authors of another article use this dataset in conjunction with 5 taxi datasets from Chinese cities to learn general trends related to taxi mobility. They focus on three statistics: the displacement of each trip, the duration of each occupied trip, and the time interval between successive occupied trips for the same taxi.