Università di Pisa	A.A. 2012-2013
Data Mining II	

Project assignments / Part 2

"Mobility Data Analysis"

General information

Objective of this project is to perform some data analysis steps over a dataset that describe the movements of several users during a long interval of time. The rules for this project are the same applied in the first part:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the analyses indicated in the text, trying to answer to each request. Any spontaneous addition to that is welcome yet optional, and cannot replace the original TODO list;
3. each group should summarize the work done in a short report (indicatively 5-15 pages), loosely following the guidelines of the CRISP model;
4. each group is totally free to choose the tools and software it prefers (although, in this case the MAtlas software appears to be the simplest choice);
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni (mirco.nanni@isti.cnr.it).

The dataset

The raw dataset is composed of GPS traces collected in (Microsoft Research Asia) Geolife project by 182 users in a period of over three years (from April 2007 to August 2012). A GPS trajectory of this

dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude and altitude. This dataset contains 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48,000+ hours. These trajectories were recorded by different GPS loggers and GPS-phones, and have a variety of sampling rates. 91 percent of the trajectories are logged in a dense representation, e.g. every 1~5 seconds or every 5~10 meters per point. This dataset recoded a broad range of users' outdoor movements, including not only life routines like go home and go to work but also some entertainments and sports activities, such as shopping, sightseeing, dining, hiking, and cycling. (Source: http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/)

Some additional information can be found on: http://research.microsoft.com/en-us/projects/geolife/

Objectives

1. Mobility profiles: select one specific user and compute his/her mobility profile (i.e., the set of trips that are frequently repeated by the user in his/her history). Experiment with different trajectory distances for the clustering step involved, including the "Start + end" distance and the "Route similarity". Describe the results.

2. O/D matrix: partition the city of Beijin into a set of regions, for instance through a regular grid (any alternative solution is welcome), in such a way that at least 10 regions are obtained. Then, extract and explore an O/D matrix of the flows among these regions. Comment on the results. Notice: here we count trajectories not considering the associated user ID, therefore, for our purposes, N trajectories performed by the same user have the same weight of N users performing 1 trajectory each.

3. Typical paths: select two regions inside Beijin from the O/D matrix study performed above. Then, study the different routes followed by the flows performed by our users to move between them.

4. Evaluate the privacy issues that might arise in this application, and discuss possible counter-measures to adopt in order to remove or limit them.

General remark: the raw dataset is extremely large (over 24M points), therefore it is highly recommended to perform a strong selection at the very beginning of the analysis, for instance limiting the time period of the analysis. That should not be needed for the mobility profile extraction, since only one user is involved.