Investigating Causality in Human Behavior from Smartphone Sensor Data: A Quasi-Experimental Approach Supplementary Material

Fani Tsapeli[1] and Mirco Musolesi[2],[1]
[1]School of Computer Science, University of Birmingham, United Kingdom
Department of Geography, University College London, United Kingdom

Location Clustering and Labeling

We create location clusters using raw GPS traces. In order to increase the quality of the location estimation, we consider only GPS samples with less than 50 meters accuracy. Moreover, we ignore any sample that was collected while the user was moving. For each new GPS point, we create a cluster only if the distance of this point from the centroid of any of the otherexisting clusters is more than 50 meters. Otherwise, we update the corresponding cluster with the new GPS sample. Every time a new GPS sample is added to a cluster, the centroid of the cluster is also updated. The pseudo code of the location clustering algorithm is presented in Algorithm 1.

Each location cluster is labeled as *home*, *work/university*, *gym/sports-center*, *socialization venue* or *other*. The label *socialization venue* is used to describe places like pubs, bars, restaurants and cafeterias. The label *other* is used to describe any place that does not belong to the above mentioned categories. We label as *home* the place where people spend most of the night and early morning hours. In order to find clusters that correspond to gyms/sports-centers or socialization venues we use the Google Maps JavaScript API [1]. The Google Maps JavaScript API enables developers to search for specific types of places that are close to a GPS point. The type of place is specified using keywords from a list of keywords provided by this API. We use the centroid of each unlabeled cluster to search for nearby places of interest. Places that correspond to *gym/sports centers* are specified by the keyword *gym* and places that correspond to socialization venues are specified by the keywords *bar*, *cafe*, *movie_theater*, *night_club* and *restaurant*. For each unlabeled cluster we conduct a search for nearby points of interest. If a point of interest with distance less than 50 meters from the cluster centroid is found, we label the cluster as *gym/sport-center* or *socialization venue* depending on the point of interest type. Otherwise the cluster is labeled as *other*. Any place within the university campus that is not labeled as *gym/sport-center* or *socialization venue* is labeled as *work/university*. In Fig. 1 we present the percentage of time that students spend on average in each of the five location categories that were mentioned above. According to Fig. 1 students spend the majority of the their time at home and at the university. During night and early morning hours, the location of around 60% of the samples has been labeled as *home* while the majority of samples from 9:00 to 20:00 are labeled as *university*.

Data: Set of location points $L = \{l_1, l_2, ..., l_n\}$
Result: Set of Clusters $C = \{c_1, c_2, ..., c_m\}$
$C := \{\}$;
for *each* $l \in L$ **do**
  if *accuracy(l)>50* **then**
    continue;
  end
  $locationClusteredFlag := 0$;
  for *each* $c \in C$ **do**
    $H := \{Z^{j,k} : Z^{j,k} \in P\}$;
    if *distance(l, centroid(c))<50* **then**
      $c := c \cup \{l\}$;
      $locationClusteredFlag := 1$;
      break;
    end
  end
  if $locationClusteredFlag = 0$ **then**
    $newCluster := \{l\}$;
    $C := C \cup \{newCluster\}$;
  end
end

Algorithm 1: Location clustering

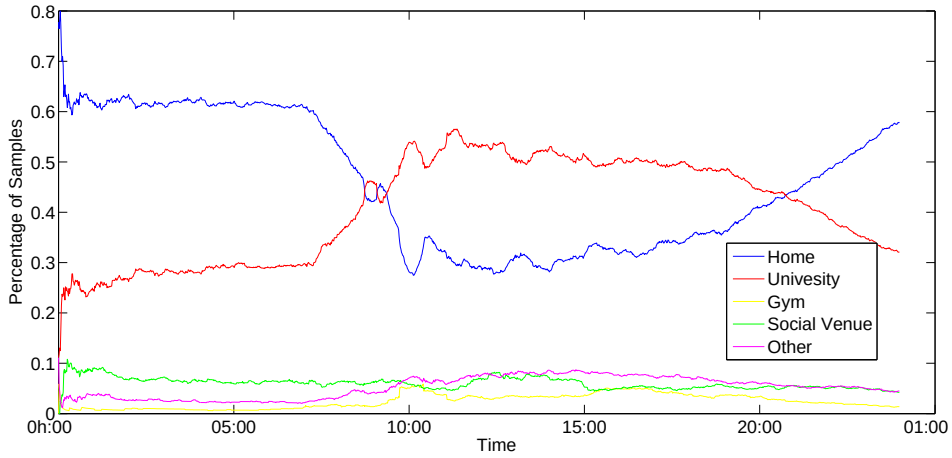

Figure 1: Percentage of time that students spend on average in each of the five location categories

Matching Method

For matching the treatment and control units we use the MatchIt R package [2] that includes an implementation of the genetic matching algorithm described above. Several optimization criteria can be used with Genetic Matching [3]. Here, the balance metric that the Genetic Matching algorithm optimizes is the mean standardized difference of all the confounding

variables. We use matching with replacement, i.e., each control unit can be matched to more than one treatment units. Matching with replacement can reduce the bias, since control units which are very similar to treatment units can be exploited more. We use a matching ratio equal to 2. This means that each treatment unit is matched with up to 2 control units.

References

[1] Google Maps Places API. https://developers.google.com/maps/documentation/javascript/places

[2] Ho, D.E., Imai, K., King, G., Stuart, E.: Matchit: nonparametric preprocessing for parametric causal inference. PhD thesis (2006)

[3] Diamond, A., Sekhon, J.S.: Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Review of Economics and Statistics **95**(3), 932–945 (2013)