Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

A Robust Missing Data-Recovering Technique for Mobility Data Mining

Annam Zafar, Muhammad Kamran, Shafqat Ali Shad & Wasif Nisar

To cite this article: Annam Zafar, Muhammad Kamran, Shafqat Ali Shad & Wasif Nisar (2017) A Robust Missing Data-Recovering Technique for Mobility Data Mining, Applied Artificial Intelligence, 31:5-6, 425-438, DOI: 10.1080/08839514.2017.1378120

To link to this article: https://doi.org/10.1080/08839514.2017.1378120

APPLIED ARTIFICIAL INTELLIGENCE
2017, VOL. 31, NOS. 5–6, 425–438
https://doi.org/10.1080/08839514.2017.1378120

A Robust Missing Data-Recovering Technique for Mobility Data Mining

Annam Zafar, Muhammad Kamran, Shafqat Ali Shad, and Wasif Nisar

Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan

ABSTRACT

Based on location information, users' mobility profile building is the main task for making different useful systems such as early warning system, next destination and route prediction, tourist guide, mobile users' behavior-aware applications, and potential friend recommendation. For mobility profile building, frequent trajectory patterns are required. The trajectory building is based on significant location extraction and the user's actual movement prediction. Previous works have focused on significant places extraction without considering the change in GSM (global system for mobile communication) network and is based on complete data analysis. Since network operators change the GSM network periodically, there are possibilities of missing values and outliers. These missing values and outliers must be addressed to ensure actual mobility and for the efficient extraction of significant places, which are the basis for users' trajectory building. In this paper, we propose a methodology to convert geo-coordinates into semantic tags and we also purposed a clustering methodology for recovering missing values and outlier detection. Experimental results prove the efficiency and effectiveness of the proposed scheme.

Introduction

Human mobility data is required to understand human mobility. To collect such data, devices like mobile devices using global system for mobile communication (GSM)-based methods, especially cell global identity (CGI) information, different sensors networks, or number of global positioning system (GPS) appliances, are used. Data collected through these devices is known as the spatial temporal dataset. GPS-based data collection methods have some limitations such as the following: (i) GPS-installed devices have battery consumption issues because extra installation is required; (ii) signaling strength is weak inside a building so data collection issue may arise; and (iii) GPS appliances are expensive to use and also have some installation problems. On the other end, GSM-based methods are inexpensive, have no

CONTACT Muhammad Kamran ✉ muhammad.kamran@ciitwah.edu.pk; m.kamran.nuces@gmail.com Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/UAAI.

installation problem, do not have signal strength problem, and because no extra installation is required, they do not have much battery consumption issues. Owing to these reasons, GSM is usually preferred over GPS. This makes mobile phone data an ample source for spatial temporal data.

In mobility data mining, GSM-based spatial temporal data is frequently used for human mobility profile building, which is the main task while developing different useful mobile applications such as early warning systems (Capan et al. 2015; Storey, van der Gaag, and Burns 2011; Zambrano et al. 2014), next destination and route prediction (Horvitz and Krumm 2012; Krumm, Gruen, and Delling 2013; Merah et al. 2013), traffic management (Corman et al. 2012; Wang et al. 2012; Yan et al. 2013), advertisement, community finding (Scripps and Trefftz 2013), determining point of interest (POI) (Chon et al. 2012; Chon and Cha 2011; Yuan, Zheng, and Xie 2012), potential friend recommendation, and social networking(Dong et al. 2012; Samanthula and Jiang 2013; Shad and Chen 2012a). In the GSM network, CGI is the main factor for providing human mobility information. It includes mobile country code (MCC) (varies from country to operator), mobile network code (MNC) (binds with operator), location area code (LAC) for cell arrangement (arranged and assigned by the network operator), and Cell Id (CI) (given to every user for network connection). This CGI information is used for uniquely identifying a user over the network (Shad and Chen 2013). For mobility profile building, it is important to have this CGI information in the form of location tags.

LAC and CI are important factors of CGI information for the extraction of information about the geo-coordinates of a particular mobile user using Google API (Shad and Chen 2012b). We use these geo-coordinates and then by applying Reverse geo-coding on these geo-coordinates convert them in the form of complete location address. After this conversion, we apply geo-coding on the extracted location and convert them in the form of location tags. Sometimes, due to aging factor in the GSM data or due to abrupt change in the network structure, it is not possible to completely convert LAC and CI in the form of geo-coordinates, which results in numerous such LAC and CI that do not have associated geo-coordinates and so it is not possible to convert them in the form of location tags. This led us to much information loss. We can call this problem the missing information problem and we must have some defined methodologies to solve this problem. There are some solutions present in the literature to address this problem, for example semantic-based clustering, single imputation, multiple imputation (MI), or mean imputation (Little and Rubin 2014; Van Buuren 2012). One of the simplest solutions is ignoring those LAC and CI that do not have associated geo-coordinates. However, it is not a suitable solution.

By taking this problem into consideration, in this paper, we propose a robust missing values and outlier removal algorithm. We first generated a new attribute that contains information about semantic tags and then clustered the whole set based on user-provided sematic tags and tags generated through Google API. As a result, we obtained n number of clusters. One of these clusters is a type of cluster that does not contain any associated semantic tag. Then we used the distance threshold-based mechanism and placed all the values present in the missing semantic tag cluster in the appropriate cluster. Then, by using the MI technique, we estimated and recovered values for the missing locations.

Related work

It is very important for us to know that before going to mine through the data for mobility profile building, we must have data with good quality. The data obtained from GSM may be incomplete, noisy, or inconsistent (Tyagi, Solanki, and Tyagi 2010). With the immense growth in the development of applications that utilize location-based information, there are several methods available in the literature that are used for outlier removal and recovering missing location information because in research study missing data is ubiquitous. This is one of the important and pervasive problems. Despite its importance and pervasiveness, many researchers simply use standby techniques such as deleting (list-wise or pair-wise), or ignoring the missing data records that are not suitable for practical applications. We can broadly divide other more suitable missing data techniques as Maximum Likelihood (ML) estimation, multiple and mean imputation techniques that do not have very rigorous assumptions, and alleviate the pitfalls of excluding data technique (Baraldi and Enders 2010). For handling missing values in multidimensional data, principal component analysis (PCA)-based approaches are more suitable because PCA reduces the dimensionality of data and also helps isolate noise that results in increasing imputation stability and decreasing variability of the estimator. A formulation of probabilistic PCA (PPCA) is helpful for missing values imputation. Thus, an algorithm, named as Variational Bayesian PCA (VBPCA), had been proposed in this regard (Ilin and Raiko 2010). A technique developed for missing traffic flow data, proposed by Chiou et al., is based on the functional data approach to handle missing values. They used the functional principal component analysis (FPCA) technique for missing values imputation. Their proposed technique not only imputes missing values but also handles outliers (Chiou et al. 2014). Similarly, another approach for continuous dataset based on Bayesian principal component analysis (BPCA) for missing values imputation was proposed by Audigier, Husson, and Josse (2014). However, an improved version of K nearest neighbor (KNN), called the local least square (LLS) method, was proposed by Chang, Zhang, and Yao (2012), which performs better than BPCA. BPCA uses correlation between global data information, whereas LLS exploits local correlation. Because BPCA does not perform better on data with strong local similarity, another method called the Bi cluster-based BPCA was proposed by Meng, Cai, and Yan (2014), which fully exploits the local structure of the data matrix although it takes more computational time than either BPCA or LLS. There are some estimation-based methods in the literature to recover missing values, such as the one proposed by Le Gruenwald et.al. (Gruenwald et al. 2010); their work is related to the recovery of mobile sensor network (MSN) missing data. Their methodology works by converting mobile sensor readings into static virtual readings and by mining the spatial and temporal relationship between readings. Some methods such as PPCA and Kernel PPCA (KPPCA) perform better if we consider spatio-temporal dependencies (Li, Li, and Li 2013). To deal with missing data, the Tucker decomposition-based algorithm (TDA) (Tan et al. 2013) works by creating a multi-way dataset of the provided traffic dataset. MI-based methods are also widely being used for missing data handling (Díaz-Ordaz, Kenward, and Grieve 2014; Lee and Carlin 2010).

Proposed work

Our proposed algorithm is also a kind of mixture of the estimation-based method and the MI techniques. From the available MIT reality mining dataset (http://realitycommons.media.mit.edu/realitymining4.html), **all-locs** and **cell_names** tables contain our required information, so we use these tables of user x. The all-locs table contain information about unique towers seen by user x during all his/her mobility information-gathering procedure. The table cell names contain information about the semantic tags provided by that particular user against some of his/her visited location. User-provided semantic tags are very less in number; for example, Table 1 describes the unique location information of that x user and the number of associated semantic tags.

Another problem that we find is that most of the users-provided semantic tags do not contain associated geo-coordinates. As this information about associated semantic tags is not sufficient enough required for mobility profile building, we need to use some method that can help increase this information. Another benefit of having semantic tags associated with the maximum number of visited locations is that these will help us in clustering during missing value retrieval. Thus, by keeping all these benefits in our mind, we decided to first retrieve more semantic tags before actually extracting missing values. In our dataset, we also have geo-coordinate information provided by Shad and Chen (2012b). We use these geo-coordinates and apply reverse geocoding on these coordinates to extract the complete location addresses. After

Table 1. Mobility information of single user (x) before applying geo-coding.

Table names	Information
All-locs	**1744** uniquely visited locations
Cell names	93 associated user-provided semantic tags

Table 2. After applying geo-coding.

# of unique cells visited by x user	1744	100%
User-provided semantic tags	93	5.3%
Semantic tags extracted through Google API	789	45%
Total semantic tags	833	48%
Missing values	911	52%

that, by applying geo-coding on these extracted locations, we convert them in the form of semantic tags. As a result, we now have associated semantic tags against almost 48% of all uniquely visited locations.

The above-mentioned description in Table 2 shows that we have 52% missing location information. We propose a robust semantic tags-based clustering technique that helps us recover these missing values.

Methodology

Before applying our proposed algorithm, we are assuming that there are attributes such as LAC, CI, Lat (Latitude), Lon (Longitude), user-provided semantic information (USI), and geo-coding-based semantic information (GSI). Now we are considering having this data in the form of matrix M_{m*6}. So we can define this matrix as

$$\mathrm{M} = [\mathrm{A\ B\ C\ D\ E\ F}]$$

Each of its element in itself is a column matrix and each one A, B, C, D, E, and F is equivalent to LAC, CI, Lat, Lon, USI, and GSI, respectively. Now first of all, we had generated a new attribute by the concatenation of E and F such that

$$E = \begin{bmatrix} fII \\ f21 \\ . \\ . \\ . \\ fn1 \end{bmatrix}$$

$$F = \begin{bmatrix} g11 \\ g21 \\ . \\ . \\ . \\ gn1 \end{bmatrix}$$

$$E||F = \begin{cases} fij\,||\,gij = fij \ \ if \ \ gij = 0 \\ fij\,||\,gij = gij \ \ if \ \ fij = 0 \\ fij\,||\,gij = fij \ \ if \ \ gij, fij\ ! = 0 \end{cases} \tag{1}$$

And we name this new attribute E||F, generated through concatenation simply as semantic information (SI). This SI is also a column matrix. Thus, this can be expressed as

$$\mathrm{SI} = \begin{bmatrix} f11 \,||\, g11 \\ f21 \,||\, g21 \\ . \\ . \\ . \\ fm1 \,||\, gm1 \end{bmatrix}$$

By performing all the above-mentioned procedures, now we are considering having a new matrix:

$$\mathrm{N} = [\mathrm{A\ B\ C\ D\ S\ I}]$$

We apply our proposed clustering algorithm on this newly generated matrix N_{m*5}. We define clustering as follows:

There are three possibilities for each S_i ϵ SI and for each of these possibilities, we define clustering by a mapping.

Case I: $s_i = s_j$ for some s_i, s_j ϵ SI and s_i, s_j ! = 0
Let us define mapping Ψ: SI → J defined by s_i ! = 0
Ψ (s_i) = {X_i: s_i =s_j and s_i, s_j ϵ SI

where X_i is the matrix of the order p × k, where k = 5 and 1 ≤ p ≤ m. For each i convert X_i into a row matrix of order 1 × 5 by replacing each column with their respective mean values. Hence, as a result, for each i we will obtain a new row matrix P_i.

Case II: s_i = 0 = s_j for some s_i ϵ SI
Let us define another mapping ϕ: SI → J defined by s_i = 0
φ (s_i) = {Z: s_i =s_j and s_i, s_j ϵ SI

where we can also call this Z as the missing values cluster.

Case III: unique s_i ϵ SI

Now we define another mapping δ: SI → J, where J is a set of row matrices of order $Y_{i's}$ formed by considering a complete row corresponding to each unique value of S_i including s_i.

δ(s_i) = {Y_i: s_i ϵ SI, where s_i ! =s_j, s_i! = 0 ! = s_j

Now we define a new collection ω of matrices W_i. To each first element of all rows of Z, there is an associated matrix in this collection ω. The rows of these matrices are formed based on a comparison between each first entry of Z and each first entry of P_i and Y_i.

Let θ: Z $\rightarrow$ ω
θ $(z_i) = \{W_i: \; if \; z_i = p_i$ or $z_i = y_i$

where $\forall \; z_i \in$ Z and p_i and y_i represent the first entry of the each row of P_i and $Y_i \; \forall \; I$, respectively. Here, the first row of $W_i \; \forall \; i$ is the row corresponding to each z_i. Now consider each matrix W_i as a single cluster. Now for each cluster W_i belonging to the collection ω, estimate and impute the value using the following procedure.

For the first value of CI belonging to any selected cluster W_i, calculate the Euclidean distance (D) between all the rest of the values present in the same column (Cl) and only select the value having the minimum distance (MD). Now compare this MD with a set threshold value (TD). If MD < TD, then impute the MD CI's associated values that include Lat and Lon in the first row of the selected cluster at equivalent places. If MD > TD, mark the first row of the selected cluster as an outlier. Now apply reverse geo-coding on the imputed geo-coordinates to extract the location address and then by applying geo-coding on the extracted location, extract the semantic tags against it.

Robust and precise missing value extraction and outlier removal algorithm

INPUT: A spatial dataset M contains unique spatial temporal information, including missing values consisting of LAC, CI, Lat, Lon, USI, GSI, d_t//distance threshold
OUTPUT: A dataset N (that includes attribute SI) with the recovered missing values.

- Read or Export M
- **Do** USI || GSI based on conditions mentioned in equation (1) and obtain a new attribute **SI**.
- Now have a new dataset **N**.
- **For each** SI = 0 select complete data row (say d_r) and MVC $\leftarrow d_r$, where MVC is missing values cluster.
- It will return dataset X = N – MVC.
- From X **Select** complete instance of each unique $s_i \in$ **SI** and consider each of them as a single cluster to get Y_i (single-instance clusters).
- It results in dataset Z = X − Y_i
- Make **n** clusters of Z based on the following conditions:
  (1) where SI ! = 0
  (2) If $s_i = s_j$ for some $s_i, \; s_j \in$ **SI** consider them in a single cluster (*say* SC_i).
  (3) **For each** SC_i calculate the statistical mean of LAC, CI, Lon, Lat and get the row matrix R_i
  (4) $MC_i \leftarrow R_i$ (where *MC* is the mean value cluster).

- **For all** clusters (MVC, Y_i, MC_i) do the following:
  1. Select one d_r from MVC
  2. Now compare the LAC of the selected d_r with the LAC of Y_i and MC_i
  3. Select only the matching LAC of Y_i and MC_i
  4. Calculate Euclidean distance (say d) between the selected MVC's CI and CI of matching Y_i and MC_i
  5. Select Y_i and MC_i with minimum d.
  6. Compare d with d_t.
  7. If d $<d_t$, then place the selected d_r in the selected cluster $(Y_i$ or $MC_i)$; otherwise, mark the missing value as the outlier.
  8. Recover the missing coordinate by imputing the mean lat/long in d_r of the selected cluster $(Y_i$ or $MC_i)$t.
  9. Extract the location of the imputed coordinate by applying Reverse Geocoding.
  10. Repeat steps 1–9 for all instances of MVC.

Experiments and results

In this section, we report the result of our experimental study.

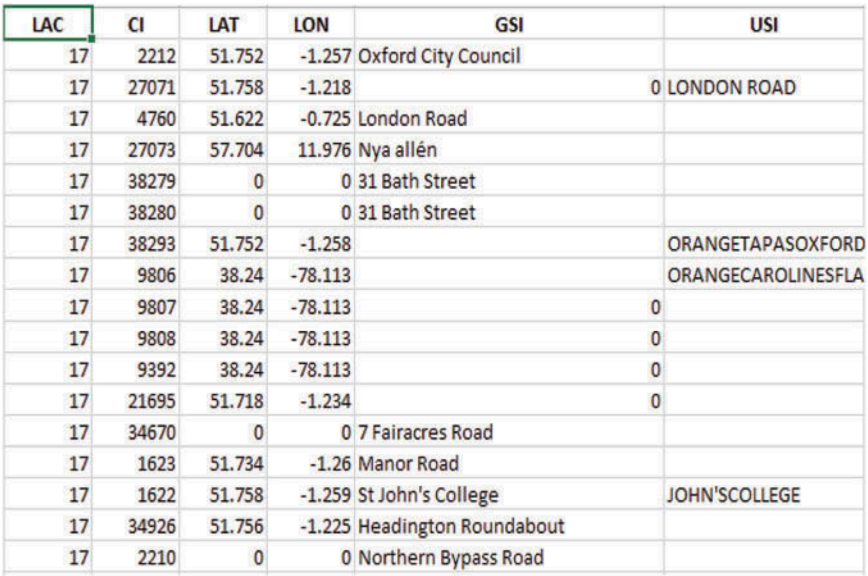
Dataset

For experimental purposes, we will use the dataset provided by MIT reality mining (http://realitycommons.media.mit.edu/realitymining4.html). It is a publically available dataset. It consists of the mobility information of 106 users and of 9-month duration. This data is collected using Nokia 6600. As this dataset contains partial mobility information, we use the lat/long information provided by Shad and Chen (2012b).

Results and discussion

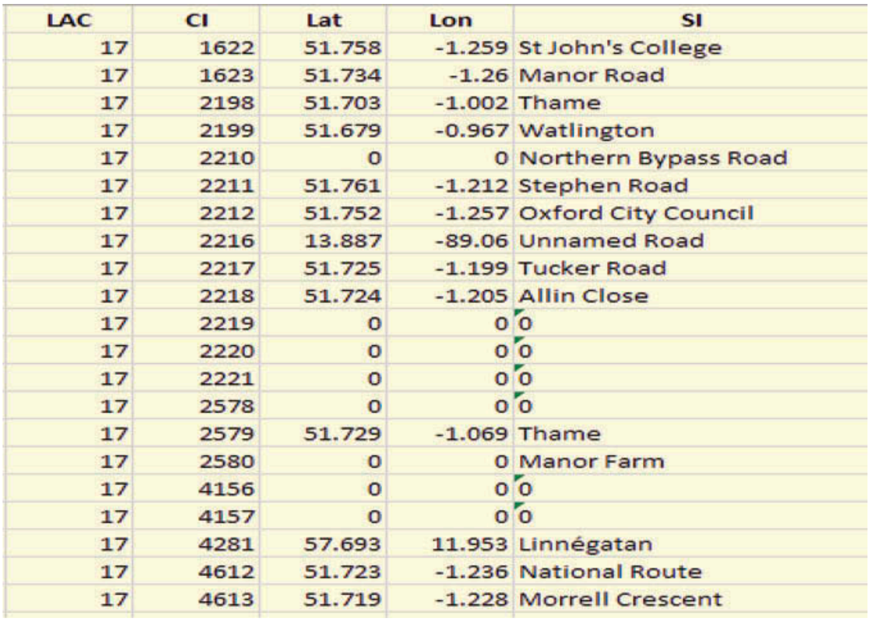
Our dataset consisted of LAC, CI, user-provided semantic tags, latitude and longitude attributes. Using Google API through Reverse Geocoding, we extracted the location attribute. After generating this attribute, we applied geo-coding on it and converted this location information in the form of location tags. Figures 1–3 show the description dataset before and after applying our proposed algorithm dataset, and Figures 4 and 5 geographically show these location on a google map.

As we mentioned before in Table 2, we have 52% missing locations in the overall mobility of user x. By applying our proposed missing values clustering algorithm, we recover more 28% missing values so now we have almost 76% values with complete location information as mentioned in Table 3, while only 26% values do not contain complete location information. In other



LAC	CI	LAT	LON	GSI	USI
17	2212	51.752	-1.257	Oxford City Council	
17	27071	51.758	-1.218	0	LONDON ROAD
17	4760	51.622	-0.725	London Road	
17	27073	57.704	11.976	Nya allén	
17	38279	0	0	31 Bath Street	
17	38280	0	0	31 Bath Street	
17	38293	51.752	-1.258		ORANGETAPASOXFORD
17	9806	38.24	-78.113		ORANGECAROLINESFLA
17	9807	38.24	-78.113	0	
17	9808	38.24	-78.113	0	
17	9392	38.24	-78.113	0	
17	21695	51.718	-1.234	0	
17	34670	0	0	7 Fairacres Road	
17	1623	51.734	-1.26	Manor Road	
17	1622	51.758	-1.259	St John's College	JOHN'SCOLLEGE
17	34926	51.756	-1.225	Headington Roundabout	
17	2210	0	0	Northern Bypass Road	

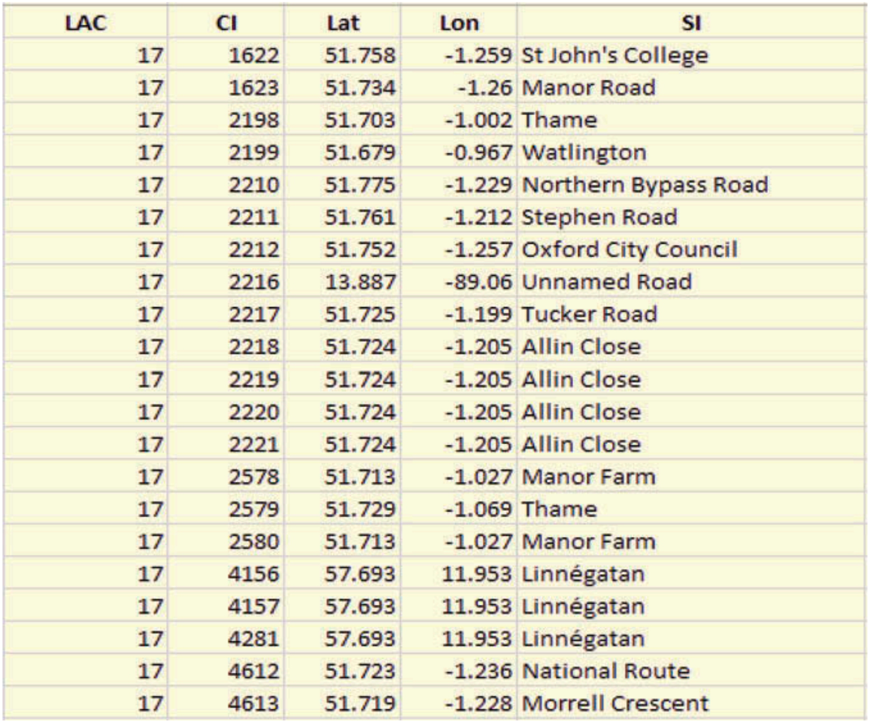
Figure 1. Missing values.



LAC	CI	Lat	Lon	SI
17	1622	51.758	-1.259	St John's College
17	1623	51.734	-1.26	Manor Road
17	2198	51.703	-1.002	Thame
17	2199	51.679	-0.967	Watlington
17	2210	0	0	Northern Bypass Road
17	2211	51.761	-1.212	Stephen Road
17	2212	51.752	-1.257	Oxford City Council
17	2216	13.887	-89.06	Unnamed Road
17	2217	51.725	-1.199	Tucker Road
17	2218	51.724	-1.205	Allin Close
17	2219	0	0	0
17	2220	0	0	0
17	2221	0	0	0
17	2578	0	0	0
17	2579	51.729	-1.069	Thame
17	2580	0	0	Manor Farm
17	4156	0	0	0
17	4157	0	0	0
17	4281	57.693	11.953	Linnégatan
17	4612	51.723	-1.236	National Route
17	4613	51.719	-1.228	Morrell Crescent

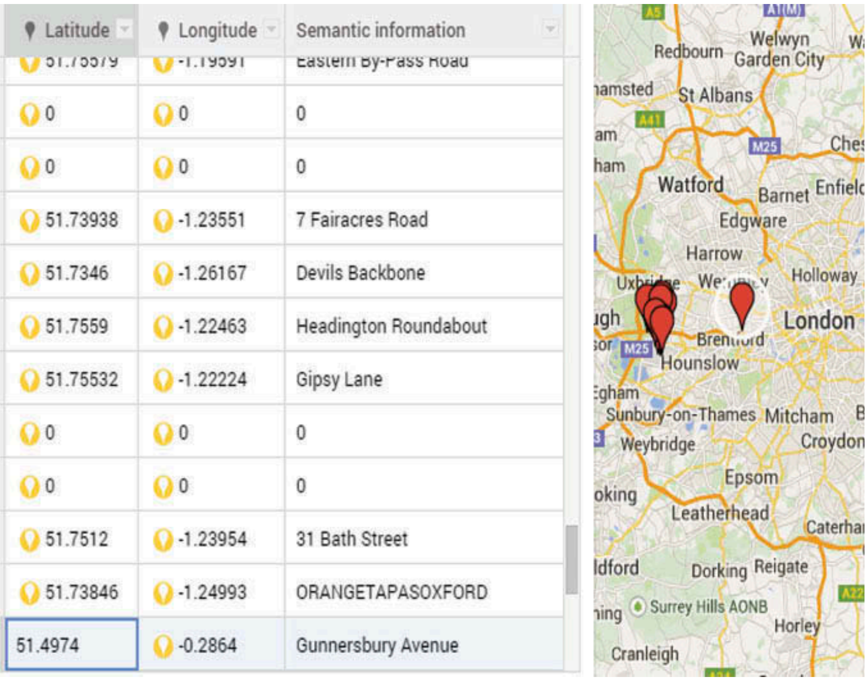
Figure 2. After concatination.

words, we can say that by applying our proposed algorithm, we extracted about 54% missing values from among 100% missing locations of user x as this user contains 911 missing values as mentioned in Table 4 and so we are left with only 420 missing locations.



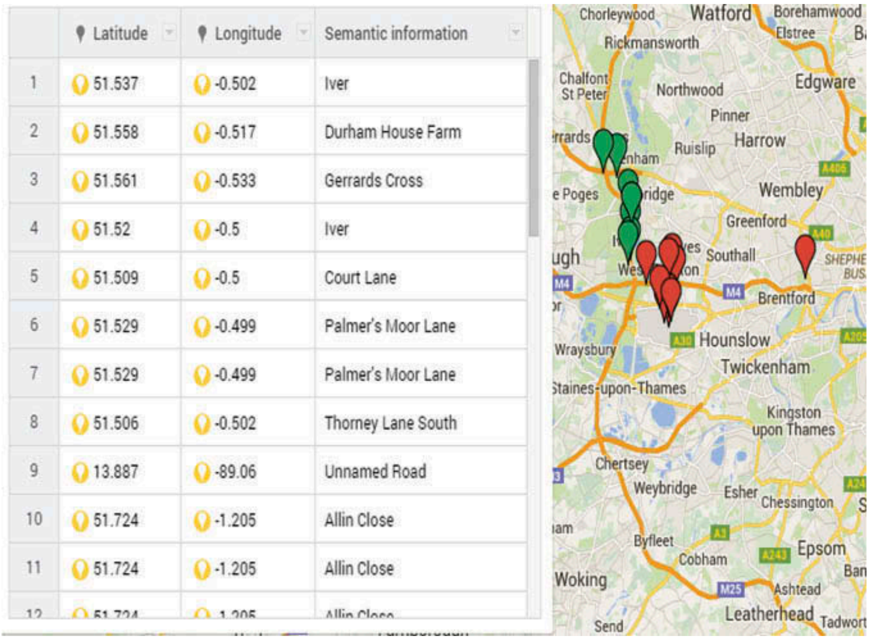
LAC	CI	Lat	Lon	SI
17	1622	51.758	-1.259	St John's College
17	1623	51.734	-1.26	Manor Road
17	2198	51.703	-1.002	Thame
17	2199	51.679	-0.967	Watlington
17	2210	51.775	-1.229	Northern Bypass Road
17	2211	51.761	-1.212	Stephen Road
17	2212	51.752	-1.257	Oxford City Council
17	2216	13.887	-89.06	Unnamed Road
17	2217	51.725	-1.199	Tucker Road
17	2218	51.724	-1.205	Allin Close
17	2219	51.724	-1.205	Allin Close
17	2220	51.724	-1.205	Allin Close
17	2221	51.724	-1.205	Allin Close
17	2578	51.713	-1.027	Manor Farm
17	2579	51.729	-1.069	Thame
17	2580	51.713	-1.027	Manor Farm
17	4156	57.693	11.953	Linnégatan
17	4157	57.693	11.953	Linnégatan
17	4281	57.693	11.953	Linnégatan
17	4612	51.723	-1.236	National Route
17	4613	51.719	-1.228	Morrell Crescent

Figure 3. After recovering missing value.



Latitude	Longitude	Semantic information
51.75579	-1.19591	Eastern By-Pass Road
0	0	0
0	0	0
51.73938	-1.23551	7 Fairacres Road
51.7346	-1.26167	Devils Backbone
51.7559	-1.22463	Headington Roundabout
51.75532	-1.22224	Gipsy Lane
0	0	0
0	0	0
51.7512	-1.23954	31 Bath Street
51.73846	-1.24993	ORANGETAPASOXFORD
51.4974	-0.2864	Gunnersbury Avenue

Figure 4. Values before clustering.



	Latitude	Longitude	Semantic information
1	51.537	-0.502	Iver
2	51.558	-0.517	Durham House Farm
3	51.561	-0.533	Gerrards Cross
4	51.52	-0.5	Iver
5	51.509	-0.5	Court Lane
6	51.529	-0.499	Palmer's Moor Lane
7	51.529	-0.499	Palmer's Moor Lane
8	51.506	-0.502	Thorney Lane South
9	13.887	-89.06	Unnamed Road
10	51.724	-1.205	Allin Close
11	51.724	-1.205	Allin Close
12	51.724	-1.205	Allin Close

Figure 5. Values after clustering.

Table 3. Location tags against overall mobility.

Total unique locations	1744	100%
Tags before clustering	**833**	**48%**
Tags after clustering	1324	76%

Table 4. Extracted missing values information.

Missing locations	911	100%
Locations retrieved through the proposed algorithm	491	54%

Accuracy/percentage error calculation

To measure the accuracy and percentage error, we adopted the following strategy.

(1) First, we applied sampling with replacement on our prepared datasets of 1342 values and chose a sample of 400 values, which is about 30% of all the datasets, and call this dataset the original sample data file (OSDF).
(2) This sample OSDF contains a number of semantic tags as described in Table 5.

Table 5. Semantic tags information.

User-provided semantic tags	API-generated semantic tags
190	210

Table 6. Cases information.

Total # of cases	# of True cases	# of False cases
200	170	30

(3) From this sample data file, we randomly removed 78 API-generated and 122 user-provided semantic (50%) values. And we call this new file the secondary sample data file (SSDF).

(4) Now as this SSDF is our generated missing values dataset, we apply our proposed methodology on it and regenerate the removed values.

(5) Then compare both OSDF and SSDF and obtain the measures as shown in Table 6.

(6) After obtaining the measures mentioned in Table 6, we calculate the accuracy and percentage error for our extracted result.

$$\begin{aligned}
Accuracy &= \frac{\#\text{ of true cases}}{\text{total }\#\text{ of cases}} * 100 \\
&= \frac{170}{200} * 100 \\
&= 85\% \\
\%\ error &= \frac{\#\text{ of errors}}{\text{actual numbers}} * 100 \\
&= \frac{30}{200} * 100 \\
&= 15\%
\end{aligned}$$

Now our data is prepared enough to be used for mobility profile building to develop some useful applications such as users' recommender systems, i.e. next destination and route prediction, or friend recommender system based on user profile similarity.

Conclusion and future work

In this work, our main focus was on data preprocessing, which will later help us develop a human mobility profile. In data preprocessing, we generated a new attribute by using Google API before recovering missing values and removing outliers. This new attribute consists of semantic tags, which basically is the conversion of geo-coordinates into a more human readable form. Based on these generated semantic tags, we consolidated our raw mobility data. Then by applying our proposed algorithm, we recovered missing values and removed outliers.

In the future, we are interested in using this preprocessed data to work on users' mobility profile building for human movement intention detection and also to find similar users for potential friend recommendations.

References

Audigier, V., F. Husson, and J. Josse. 2014. Multiple imputation for continuous variables using a Bayesian principal component analysis. *arXiv Preprint arXiv* 1401:5747.

Baraldi, A. N., and C. K. Enders. 2010. An introduction to modern missing data analyses. *Journal of School Psychology* 48 (1):5–37. doi:10.1016/j.jsp.2009.10.001.

Capan, M., J. S. Ivy, T. Rohleder, J. Hickman, and J. M. Huddleston. 2015. Individualizing and optimizing the use of early warning scores in acute medical care for deteriorating hospitalized patients. *Resuscitation* 93:107–12. doi:10.1016/j.resuscitation.2014.12.032.

Chang, G., Y. Zhang, and D. Yao. 2012. Missing data imputation for traffic flow based on improved local least squares. *Tsinghua Science and Technology* 17 (3):304–09. doi:10.1109/TST.2012.6216760.

Chiou, J.-M., et al. 2014. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics* 2 (2):106–29.

Chon, J., and H. Cha. 2011. Lifemap: A smartphone-based context provider for location-based services. *IEEE Pervasive Computing* 10 (2):58–67.

Chon, Y., et al. 2012. Automatically characterizing places with opportunistic crowdsensing using smartphones. Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM, September 5–8, Pittsburgh, USA.

Corman, F., A. D'Ariano, D. Pacciarelli, and M. Pranzo. 2012. Bi-objective conflict detection and resolution in railway traffic management. *Transportation Research Part C: Emerging Technologies* 20 (1):79–94. doi:10.1016/j.trc.2010.09.009.

Díaz-Ordaz, K., M. G. Kenward, and R. Grieve. 2014. Handling missing values in cost effectiveness analyses that use data from cluster randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 177 (2):457–74. doi:10.1111/rssa.12016.

Dong, Y., et al. 2012.Link prediction and recommendation across heterogeneous social networks. Data Mining (ICDM), 2012 IEEE 12th International Conference on, IEEE. December 10–13, Brussels, Belgium.

Gruenwald, L., et al. 2010. DEMS: A data mining based technique to handle missing data in mobile sensor network applications. Proceedings of the Seventh International Workshop on Data Management for Sensor Networks, ACM, September 13, Singapore.

Horvitz, E., and J. Krumm. 2012. Some help on the way: Opportunistic routing under uncertainty. Proceedings of the 2012 ACM conference on Ubiquitous Computing, ACM, September 5–8, Pittsburgh, USA.

Ilin, A., and T. Raiko. 2010. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research* 11:1957–2000.

Krumm, J., R. Gruen, and D. Delling. 2013. From destination prediction to route prediction. *Journal of Location Based Services* 7 (2):98–120. doi:10.1080/17489725.2013.788228.

Lee, K. J., and J. B. Carlin. 2010. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* 171 (5):624–32. doi:10.1093/aje/kwp425.

Li, L., Y. Li, and Z. Li. 2013. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies* 34:108–20. doi:10.1016/j.trc.2013.05.008.

Little, R. J., and D. B. Rubin. 2014. *Statistical analysis with missing data*. Hoboken, New Jersey, USA: John Wiley & Sons.

Meng, F., C. Cai, and H. Yan. 2014. A bicluster-based bayesian principal component analysis method for microarray missing value estimation. *EEE journal of biomedical and health informatics* 18 (3):863–71.

Merah, A. F., S. Samarah, A. Boukerche, and A. Mammeri. 2013. A sequential patterns data mining approach towards vehicular route prediction in VANETs. *Mobile Networks and Applications* 18 (6):788–802. doi:10.1007/s11036-013-0459-6.

Samanthula, B. K., and W. Jiang. 2015. Interest-driven private friend recommendation. *Knowledge and Information Systems* 42 (3):663–687.

Scripps, J., and C. Trefftz. 2013. Community finding within the community set space. Proceedings of the 7th Workshop on Social Network Mining and Analysis, ACM August 11–14, Chicago, IL, USA.

Shad, S. A., and E. Chen. 2012a. Precise location acquisition of mobility data using cell-id. *arXiv Preprint arXiv* 1206:6099.

Shad, S. A., and E. Chen. 2012b. Spatial outlier detection for mobility profile mining. *International Journal of Advanced Research in Computer Science* 3 (3):68–74.

Shad, S. A., and E. Chen. 2013. Unsupervised user similarity mining in gsm sensor networks. *The Scientific World Journal* 2013:1–11. doi:10.1155/2013/589610.

Storey, M. V., B. van der Gaag, and B. P. Burns. 2011. Advances in on-line drinking water quality monitoring and early warning systems. *Water Research* 45 (2):741–47. doi:10.1016/j.watres.2010.08.049.

Tan, H., G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li. 2013. A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies* 28:15–27. doi:10.1016/j.trc.2012.12.007.

Tyagi, N. K., A. Solanki, and S. Tyagi. 2010. An algorithmic approach to data preprocessing in web usage mining. *International Journal of Information Technology and Knowledge Management* 2 (2):279–83.

Van Buuren, S. 2012. *Flexible imputation of missing data*. Boca Raton, Florida, USA: CRC press.

Wang, L., et al. 2012. Rapid traffic information dissemination using named data. Proceedings of the 1st ACM workshop on emerging name-oriented mobile networking design-architecture, algorithms, and applications, ACM, June 11, Hilton Head, South Carolina, USA.

Yan, Z., et al. 2013. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4 (3):49.

Yuan, J., Y. Zheng, and X. Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, August 12–16, Beijing, China.

Zambrano, A., et al. 2014. Quake detection system using smartphone-based wireless sensor network for early warning. in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on, IEEE, March 24–28, Budapest, Hungary.