Point2Hex: Higher-order Mobility Flow Data and Resources

(Data and Resources Paper)

Ali Faraji[§], Jing Li[§], Gian Alix, Mahmoud Alsaeed, Nina Yanin, Amirhossein Nadiri, Manos Papagelis
{ faraji, jliellen, gcalix, mahmoud2, anadiri, papaggel }@yorku.ca, nina27@my.yorku.ca
York University – Toronto, Ontario, Canada

ABSTRACT

Research on trajectory data mining relies on appropriate datasets, including Gps-based geolocations, check-in data to points of interest (Pois), and synthetic datasets. Even though some data are accessible, the majority of mobility datasets are typically discovered through ad-hoc searches and lack comprehensive documentation of their generation process or source to reproduce curated or customized versions of them. At the same time, there has been a growing interest in a new type of mobility data, describing trajectories as sequences of higher-order geometric elements like hexagons that offer several benefits: (**i**) reduced sparsity and analysis at different granularity levels, (**ii**) compatibility with popular machine learning architectures, (**iii**) improved generalization and reduced overfitting, and (**iv**) efficient visualization. To this end, we present Point2Hex, a method and tool for generating higher-order mobility flow datasets from raw trajectory data. We used Point2Hex to create higher-order versions of seven popular mobility datasets typically employed in trajectory-related technical problems and downstream tasks, such as trajectory prediction, classification, clustering, imputation, and anomaly detection, to name a few. To promote reuse and encourage reproducibility, we provide the source code and documentation of Point2Hex, as well as the generated higher-order mobility flow datasets in publicly accessible repositories.

CCS CONCEPTS

• **Information systems** → **Geographic information systems**; **Location based services**; **Global positioning systems**.

KEYWORDS

trajectory datasets, higher-order mobility flow datasets, generator

ACM Reference Format:
Ali Faraji[§], Jing Li[§], Gian Alix, Mahmoud Alsaeed, Nina Yanin, Amirhossein Nadiri, Manos Papagelis. 2023. Point2Hex: Higher-order Mobility Flow Data and Resources: (Data and Resources Paper). In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589132.3625619

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0168-9/23/11...$15.00
https://doi.org/10.1145/3589132.3625619

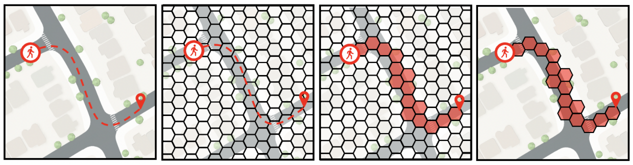

Figure 1: An example of a trajectory represented as a sequence of hexagon cells (*higher-order mobility flow data*).

1 INTRODUCTION

Motivation & Limitations. The proliferation of location-based technologies, geo-enabled smart devices, and advanced global positioning systems has resulted in the accumulation of vast amounts of *trajectory or mobility data* that includes information of an object's movement over time. Mining interesting patterns and extracting useful information from mobility data can find application in diverse domains, including intelligent transportation systems [8], urban planning and environmental monitoring [9], and public health [15, 16]. Due to the broader impact, several trajectory-related research problems have been of interest, including trajectory prediction [19], simplification [1, 18], clustering [7], classification [3, 11], and imputation [13]. While some mobility datasets are available for research purposes, the majority of them are typically discovered through ad-hoc searches and lack comprehensive documentation. Moreover, researchers frequently resort to generating their own synthetic datasets, which are often unpublished and lack details of their generation process and potential reproducibility. Moreover, working with mobility data presents its own challenges, due to its large size, high sparsity, and complexity.

Our Approach & Contributions. To address these problems, there has been a growing interest in a new type of mobility data, describing trajectories using *higher-order geometric elements* [6]. First, a map is uniformly tessellated using a geometric element, say hexagons, and then trajectories are represented as sequences of hexagons (see Fig. 1). The representation provides some advantages: it decreases sparsity and allows for analysis at varying granularity based on map tessellation; it is compatible with well-known machine learning models; and it promotes better generalization, minimizes overfitting, and supports effective visual analytics, such as heatmaps. Based on these observations, we contribute:

- We present Point2Hex, a method and tool for generating higher-order mobility flow datasets from raw trajectory data.
- We use Point2Hex to create higher-order versions of seven popular mobility datasets typically employed in trajectory-related research problems and downstream tasks.
- We provide the source code and documentation of Point2Hex, as well as the generated higher-order mobility flow datasets in publicly accessible repositories, Github and Zenodo, respectively.

[§]Authors contributed equally.

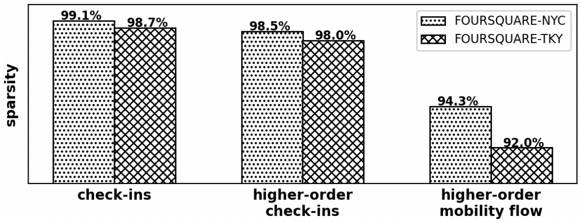

Figure 2: Impact of higher-order abstraction on sparsity.

Resource type: Datasets and Data Generation Tool
Source code: https://github.com/alifa98/point2hex
Dataset DOI: https://doi.org/10.5281/zenodo.7879595

Paper Organization. The remainder of the paper is organized as follows. Section 2 provides the rationale for operating on higher-order mobility flow data and preliminaries. Section 3 discusses the data generation pipeline followed by Point2Hex. Section 4 provides descriptions of the generated datasets, along with an insightful discussion in section 5. Finally, we conclude in section 6.

2 HIGHER-ORDER MOBILITY FLOW DATA

In this section, we present the rationale for working with higher-order mobility flow data and its advantages.

Rationale. Working with raw trajectory datasets (either in the form of GPS data points or POI check-ins) is challenging because (longitude, latitude) geo-coordinates: (*i*) are sparse and large amounts are needed to learn meaningful relationships, and (*ii*) are not very compatible (as input) to popular machine learning architectures, due to their continuous nature. We therefore propose to resort to a higher level of abstraction for representing trajectories. This transformation is done by first obtaining the routes by connecting raw trajectory data points through publicly available routing algorithms[1]. Then, the trajectory is represented as a sequence of the higher-order elements (hexagons) traversed by the route.

Advantages. The benefits of such a transformation are multi-fold. First, the sparsity will be decreased, as multiple data points (or checkins) will belong to the same higher abstraction element (e.g., hexagons). For instance, Fig. 2 illustrates the impact of higher-order representations on decreasing the sparsity on two benchmark datasets (Foursquare-Nyc and Foursquare-Tky [20]). We observe about a 1% decrease in sparsity when using higher-order check-ins, and more than a 5% decrease in the case of higher-order mobility flow. Furthermore, trajectories can now be treated as sequences of a predefined set of higher abstraction elements (e.g., hexagons), which are compatible (as input) to popular machine learning models, such as sequence models and Transformer-based models. In addition, the higher-order mobility flow data represents continuous routes/paths between location data points. While these routes are not representing the actual path an object has followed (this is unknown), they can capture implicit information and other semantics that lie in-between the different data points. The premise of operating on higher-order mobility data is that the deep learning models can avoid overfitting and generalize better as it simplifies the data in a way that omits specific details.

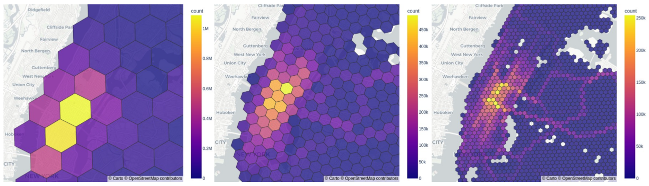

Figure 3: An illustrative example of higher-order mobility flow datasets of varying resolutions – 7, 8, 9 respectively.

2.1 Definitions and Notations

Definition 1 (Map). Let $\mathcal{M}$ be a map over a predefined, finite, and continuous geographical area.

Definition 2 (Poi). Let $\mathcal{P} = \{p_1, p_2, \ldots, p_{|P|}\}$ be a set of points of interests (Pois) on a map $\mathcal{M}$.

Definition 3 (Visits or Check-ins). In location-based services, a visit or check-in of a person to a location or place at a particular time is a record represented by a quadruplet $r = (u, l, t, \langle x, y \rangle)$, for user u, location ID l, time of visit t, and longitudinal-latitudinal tuple $\langle x, y \rangle$. We represent the set of all visits or check-ins by R. In this research, we use the term visit or check-in interchangeably.

Definition 4 (Gps Trace Points). A Gps trace point can similarly be defined using a triplet $s = (o, t, \langle x, y \rangle)$, for object o, timestamp t, and geocoordinate tuple $\langle x, y \rangle$. We denote the set of all traces by S.

Definition 5 (Trajectory). A temporarily ordered sequence of a user's visits to places (or traces), observed during a time period, can be used to describe a trajectory $Tr = \{r_1, r_2, ..., r_m\}$ (for check-ins) or $Tr = \{s_1, s_2, ..., s_m\}$ (for trace points), where m represents the trajectory's length. We denote the set of all trajectories by $\mathcal{T}$. Specifically for check-ins, each record relates to a specific Poi $p \in \mathcal{P}$; a trajectory can therefore be represented by a sequence of Pois $Tr = \{p_1, p_2, ..., p_m\}$, where p_i is a Poi at location l of the record r_i.

Definition 6 (Grid). Let $\mathcal{G} \in \{g_1, g_2, \ldots, g_n\}$ be a set of (regular) disjoint grid cells that can fully tessellate map $\mathcal{M}$, filling the plane with no gaps and thus forming a regular tiling. For the map tessellation, we opt for *hexagons*. Hexagons are preferable over other shapes, such as squares or triangles, when tessellating a map for various mobility data analysis tasks (e.g., trajectory's movement paths, connectivity, etc.). This is because hexagons are "circularly-shaped" polygons that enable a natural representation of curvatures in trajectory data, thereby reducing sampling bias in mobility data that is due to the edge effects of a grid map's shape [4]. Note as well that the tessellation can happen at different levels of resolution, by defining different sizes of the hexagons. The smaller the hexagon size, the higher the resolution. Table 1 shows the size properties of each hexagon for each resolution considered.

We are now in a position to define 'higher-order' geometric elements that embody a more abstract representation of the trajectory.

Definition 7 (Higher-order check-ins). Since $\mathcal{M}$ is fully tessellated, each check-in/Poi $p \in \mathcal{P}$ there is a $g \in \mathcal{G}$, s.t. p is in g.

Definition 8 (Higher-order Gps traces). Similarly, with $\mathcal{M}$ fully tessellated, for every $s \in S$ there is a $g \in \mathcal{G}$, such that s is in g.

Definition 9 (Higher-order trajectories). Since every $p \in \mathcal{P}$ (for check-ins) or $s \in S$ (for Gps traces) belongs in a $g \in \mathcal{G}$, we can translate every trajectory $Tr = \{p_1, p_2, ..., p_m\}$ (for check-ins)

[1] https://developers.google.com/maps/documentation/directions

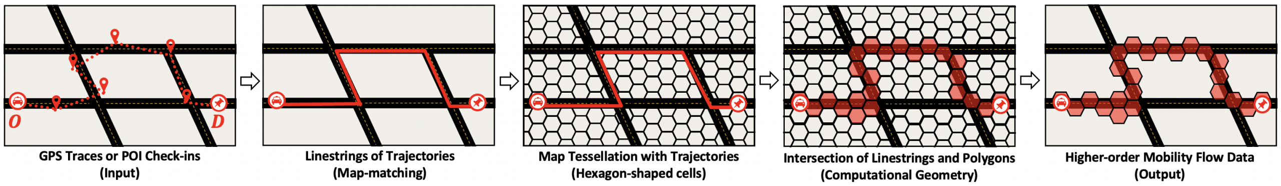
O
D
GPS Traces or POI Check-ins
(Input)
Linestrings of Trajectories
(Map-matching)
Map Tessellation with Trajectories
(Hexagon-shaped cells)
Intersection of Linestrings and Polygons
(Computational Geometry)
Higher-order Mobility Flow Data
(Output)

Figure 4: The pipeline for generating higher-order mobility flow data from either Gps trace or Poi check-in input data.

or $Tr = \{s_1, s_2, ..., s_m\}$ (for Gps traces) to a sequence of grid cells $Tr = \{g_1, g_2, ..., g_m\}$, where every $g_k \in \mathcal{G}$.

Definition 10 (Higher-order mobility flow). Given a higher-order trajectory $Tr = \{g_1, g_2, ..., g_m\}$, we can define a higher-order mobility flow as a new trajectory:

$$Tr = \{g_1, \mathcal{G}_{[1,2]}, g_2, \mathcal{G}_{[2,3]}, ..., \mathcal{G}_{[m-1,m]}, g_m\}$$

where each $\mathcal{G}_{[i,i+1]} \subset \mathcal{G}$, for $i = 1, ..., m-1$, represents the sequence of grid cells traversed between g_i and g_{i+1} of the original trajectory.

3 THE DATA GENERATION PIPELINE

In this section, we discuss the general pipeline for generating higher-order mobility flow data from original trajectory datasets. Fig. 4 shows the various steps involved in transforming trajectory data points into sequences of hexagons. Although the procedure is straightforward, it is not trivial and can sometimes be time-consuming. This is primarily due to the involvement of specialized algorithms, such as routing, map-matching, and computationally intensive geometry tasks. We further elaborate on these issues.

The Input. Trajectory datasets are typically becoming available either in the form of GPS trace data, such as the following:

```
taxi_id, date_time, longitude, latitude
1, 2008-02-02 15:36:08, 116.51172, 39.92123
1, 2008-02-02 15:46:08, 116.51135, 39.93883
```

or, in the form of Poi check-in data, such as the following:

```
userId, venueId, venueCategoryId, venueCategory, lat, lon, time
589, 4d646c, 4bf5d8, Gym, 35.6099, 139.8256, Apr 03 20:09:41 2012
290, 4b535e, 4bf5d8, Park, 35.7495, 139.5865, Apr 03 20:14:18 2012
```

Map-Matching. Regardless of the input's form, the original dataset needs to be adequately pre-processed to connect together temporally ordered data points and form *raw trajectory sequences*. Note, however, that these sequences do not reflect the actual movement of the objects on the road network. To transform these sequences to sequences of road segments on the road network traversed by the objects, *map-matching* is used that can accurately identify the path an object has taken [14]. While popular methods, such as Ivmm [23], exist for map-matching, one can use a routing machine like Osrm [10] to first find the shortest paths between consecutive intermediate data points, and then concatenate (in the same sequence) the output shortest paths to form the map-matched trajectory of Tr.

Computational Geometry. To transform a map-matched trajectory to a sequence of hexagons, we rely on computational geometry methods. More specifically, every trajectory is modeled as a `linestring` shape type, and every hexagon as a `polygon` shape type. Then, their intersection can be computed using off-the-shelf methods of popular computational geometry libraries. Recall that a map $\mathcal{M}$ can be tessellated using hexagons of a different size, which defines the map's resolution. The datasets we provide are available in different resolutions, namely $\{6, ..., 10\}$. Table 1 shows the size of each hexagon for each resolution. Fig. 3 provides an illustrative example of higher-order mobility flow data projected on an area of a map that has been tessellated using three different resolutions. The figure shows how higher-order mobility data captures a city's road infrastructure and physical constraints (e.g., bridges). It also shows how different resolutions can be useful in diverse problems and applications, as they allow for analysis at a different level of granularity, ranging from microscopic to macroscopic [15].

Resolution	Edge Length (Km)	Area (Km^2)
Hex@6	3.725	36.129
Hex@7	1.406	5.161
Hex@8	0.531	0.737
Hex@9	0.201	0.105
Hex@10	0.076	0.015

Table 1: Size properties of a hexagon at various resolutions[2].

The Output. The final output comprises continuous hexagon sequences representing each trajectory.

4 THE DATASETS

We provide the higher-order mobility flow of the following datasets (see Table 2 for a summary of the datasets' statistics):

T-Drive [21, 22]. Taxi trajectory datasets on the Beijing road network that reaches ~9 million kilometres in total.

Porto [12]. Trajectories of Portuguese taxis that operate through a dispatch central & uses mobile data terminals installed in the taxis.

Rome [5]. Consisting of recorded traces of taxi cabs in Rome, Italy.

Geolife [24–26]. Collected from several users' Gps-based devices during their outdoor activities (a total distance of ~1.2 million km)

Nyc-Taxi [17]. Taxi cab trajectory dataset recorded by digital devices installed in the vehicles of New York City.

Foursquare [20]. Poi check-ins recorded by the phones of several users in the city of New York and Tokyo.

5 DISCUSSION

5.1 Broader Impact

In this section, we provide insights into how the new datasets (potentially with other datasets) can be used in diverse domains.

Transportation Systems and Urban Planning. Our datasets can be used to create illustrative heatmaps, such as in Fig. 3, that depict the density in a particular area or route of the map. These heatmaps can be used to visually analyze traffic patterns, congestion levels and travel times. They can also allow urban planners to design and optimize public transit networks and infrastructure.

[2] https://h3geo.org/docs/core-library/restable/

	Dataset	# Objects	# Trajectories	Time Period	Resolutions	File Size
Gps traces	**Ho-T-Drive**	9,987	65,117	02/02/08 – 02/08/08	{6, ..., 10}	379 MB
	Ho-Porto	442	1,668,859	07/01/13 – 06/30/14	{6, ..., 10}	370.2 MB
	Ho-Rome	315	5,873	02/01/14 – 03/02/14	{6, ..., 10}	16.6 MB
	Ho-Geolife	57	2,100	04/01/07 – 10/31/11	{6, ..., 10}	3 MB
visits	**Ho-Foursquare-Nyc**	1,083	49,983	04/12/12 – 02/16/13	{6, ..., 10}	12.3 MB
	Ho-Foursquare-Tky	2,293	117,593	04/12/12 – 02/16/13	{6, ..., 10}	29.3 MB
	Ho-Nyc-Taxi	N/A	2,062,554	01/01/16 – 06/30/16	{6, ..., 10}	341.4 MB

Table 2: Statistics of the higher-order mobility flow datasets that we provide (the original datasets are prefixed by 'Ho-').

Environmental Monitoring and Conservation. Another potential application of higher-order mobility data is in the form of assessing environmental impact (e.g., emissions, air quality, etc.). This helps evaluate the effectiveness of eco-friendly initiatives (e.g., relating to transportation) and supports sustainability planning.

Public Healthcare. Microscopic modeling of spatiotemporal epidemics dynamics research [2, 15] can benefit healthcare practitioners and researchers in surveying and monitoring disease spread through individual mobility behaviors. This can also inform public policy about the effectiveness of targeted actions that aim to mitigate the epidemic spread compared to horizontal measures.

5.2 Privacy and Ethics

The provided datasets have been anonymized to ensure privacy and ethical considerations. They have all been derived from publicly-available sources and have undergone necessary preprocessing to meet specific research requirements. Moreover, the original datasets are publicly available and are free of use (for research intent purposes), as outlined by their respective curators, and were deemed to be collected with proper informed consent and ethical approvals. We also followed all terms and conditions of use as specified by the dataset providers; this includes properly attributing and citing their works. By sharing these higher-order mobility datasets, we aim to encourage further research and innovation while upholding privacy and ethics standards. Datasets are to be used responsibly and adhere to relevant legal and ethical guidelines.

6 CONCLUSIONS

Many existing raw trajectory datasets, whether in the form of Gps traces or check-in visits, are noisy and sparse, and learning from these data directly using deep learning models can be challenging. The lack of available preprocessed and clean datasets has motivated us to contribute to this important resource dataset. The goal is to reduce sparsity and to express trajectories as statements/sentences. We do so by providing higher levels of abstraction of seven popular mobility datasets, called *higher-order mobility flow data.* This type of dataset(s) can find application in urban planning, transportation systems, environmental conservation, and public healthcare, to name a few. As such, we encourage its use, following ethical-observing guidelines. The code for generating these datasets has been well-documented, and the procedure followed is clearly explained, so anyone can generate their own higher-order mobility flow dataset. Our higher-order datasets are available for download on Zenodo, and our code and documentation are on GitHub.

REFERENCES

[1] G. Alix and M. Papagelis. 2023. PathletRL: Trajectory Pathlet Dictionary Construction using Reinforcement Learning. In *Proc. of the 31st ACM SIGSPATIAL.* 1–12.

[2] G. Alix, N. Yanin, T. Pechlivanoglou, J. Li, F. Heidari, and M. Papagelis. 2022. A Mobility-based Recommendation System for Mitigating the Risk of Infection during Epidemics. In *23rd IEEE MDM.*

[3] M. Alsaeed, A. Agrawal, and M. Papagelis. 2023. Trajectory-User Linking using Higher-order Mobility Flow Representations. In *24th IEEE MDM.* In Press.

[4] C. P. Birch, S. P. Oom, and J. A. Beecham. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Eco. modelling* 206, 3-4 (2007), 347–359.

[5] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi. 2022. CRAWDAD roma/taxi.

[6] Z. Fan, X. Yang, W. Yuan, R. Jiang, Q. Chen, X. Song, and R. Shibasaki. 2022. Online trajectory prediction for metropolitan scale mobility digital twin. In *Proc. of the 30th ACM SIGSPATIAL.* 1–12.

[7] N. Han, S. Qiao, K. Yue, J. Huang, Q. He, T. Tang, F. Huang, C. He, and C.-A. Yuan. Algorithms for Trajectory Points Clustering in Location-Based Social Networks. *ACM TIST* 13, 3, Article 43 (Mar 2022), 29 pages.

[8] L. Li, R. Jiang, Z. He, X. M. Chen, and X. Zhou. Trajectory data-based traffic flow studies: A revisit. *Transp. Research Part C: Emerg. Techn.* 114 (2020), 225–240.

[9] T. Li, J. Gao, and X. Peng. 2021. Deep Learning for Spatiotemporal Modeling of Urbanization. In *Proc. of the 35th NeurIPS 2021* (Sydney, Australia).

[10] D. Luxen and C. Vetter. 2011. Real-time routing with OpenStreetMap data. In *Proc. of the 19th ACM SIGSPATIAL.* 513–516.

[11] C. Miao, J. Wang, H. Yu, W. Zhang, and Y. Qi. 2020. Trajectory-User Linking with Attentive Recurrent Network. In *Proc. of the 19th AAMAS.* 878–886.

[12] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas. 2015. *ECML/PKDD 15: Taxi Trajectory Prediction Porto Dataset.*

[13] M. Musleh and M. Mokbel. 2023. A Demonstration of KAMEL: A Scalable BERT-based System for Trajectory Imputation. In *ACM SIGMOD.*

[14] P. Newson and J. Krumm. 2009. Hidden Markov Map Matching through Noise and Sparseness. In *Proc. of the 17th ACM SIGSPATIAL.* 336–343.

[15] T. Pechlivanoglou, G. Alix, N. Yanin, J. Li, F. Heidari, and M. Papagelis. 2022. Microscopic Modeling of Spatiotemporal Epidemic Dynamics *(SpatialEpi '22).* ACM, 11–21.

[16] A. Strzelecki. The Apple Mobility Trends Data in Human Mobility Patterns during Restrictions and Prediction of COVID-19: A Systematic Review and Meta-Analysis. *Healthcare* 10, 12 (2022).

[17] Taxi and N. Limousine Commission. 2023. *TLC Trip Record Data.*

[18] Z. Wang, C. Long, and G. Cong. 2021. Trajectory Simplification with Reinforcement Learning. In *37th IEEE ICDE.* 684–695.

[19] H. Xue, B. P. Voutharoja, and F. D. Salim. 2022. Leveraging Language Foundation Models for Human Mobility Forecasting. In *Proc. of the 30th ACM SIGSPATIAL.*

[20] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE SMC* 45, 1 (2015), 129–142.

[21] J. Yuan, Y. Zheng, X. Xie, and G. Sun. 2011. Driving with Knowledge from the Physical World. In *Proc. of the 17th ACM SIGKDD.* 316–324.

[22] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. 2010. T-Drive: Driving Directions Based on Taxi Trajectories. In *Proc. of the 18th ACM SIGSPATIAL.* 99–108.

[23] J. Yuan, Y. Zheng, C. Zhang, X. Xie, and G.-Z. Sun. 2010. An Interactive-Voting Based Map Matching Algorithm. In *11th IEEE MDM.* 43–52.

[24] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. 2008. Understanding Mobility Based on GPS Data. In *Proc. of the 10th UbiComp '08.* 312–321.

[25] Y. Zheng, X. Xie, and W.-Y. Ma. GeoLife: A Collaborative Social Networking Service among User, location and trajectory. *IEEE Data(base) Eng. Bulletin* (2010).

[26] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. 2009. Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *Proc. 18th ACM WWW '09.* 791–800.