Visual Analytics for Understanding Traffic Flows of Transport Hubs from Movement Data

Linfang Ding, Jian Yang, Liqiu Meng

Lehrstuhl für Kartographie, Technische Universität München

Abstract. Transport hubs such as airports, railway stations play an important role in the transportation system and urban development. Understanding the traffic flows of the transport hubs may help traffic engineers and policy makers to improve the transportation planning and support passengers to efficiently plan their trips. However, analyzing the movement data is very challenging due to their large data volume, implicit spatiotemporal relationship, and uncertain semantics. In this paper, we investigate effective visual analytics for exploring spatiotemporal, semantic patterns in the traffic flows in/out of the transport hubs. The contribution of this work is three fold: (1) we propose a visual analysis workflow incorporating data mining techniques and visualization methods to enable users visually explore the spatiotemporal and semantic patterns in the traffic flows in/out of the transport hubs; (2) we use a spatial clustering method to extract significant places and categorize them to different functional regions based on the POI information; (3) we design a novel visual interface that enables the visual exploration of movement data at both an aggregated and an individual level. The visual analytics framework is tested on a real world floating taxi data set in Shanghai. Preliminary experiments show that our proposed framework is feasible and can effectively support visual exploration and identify interesting traffic flow patterns.

Keywords: Visual analytics, Traffic flow, Transport hubs

1. Introduction

Transport hubs such as airports, railway stations play an important role in the transportation system and urban development. These terminals absorb and reflect huge amounts of traffic flows from/to the road network and have considerable social and economic impacts on the surrounding regions. Understanding the traffic flows of the transport hubs may help traffic engineers and policy makers to improve the transportation planning, and support passengers to efficiently plan their trips. However, analyzing the movement data is very challenging due to their large data volume, implicit spatiotemporal relationship, and uncertain semantics.

Many efforts have been made for the understanding of the complexity and the dynamics of cities using movement data. A systematic investigation of spatial trajectories from a wide spectrum of perspectives and disciplines, e.g. spatial database, mobile computing and data mining, can be found in (Zheng and Zhou 2011). Yuan et al. (2013) presented a recommender system for both taxi drivers and taker using the knowledge of both passengers' mobility patterns and taxi drivers' picking-up/dropping-off behaviors learned from the GPS trajectories of taxicabs. Liu et al. (2012) investigated the derived six traffic 'source-sink' areas from the temporal variations of both pick-ups and drop-offs, and their association with different land use features. Ding, Fan and Meng (2015) aimed to visualize and analyze the spatiotemporal driving patterns for two income-level groups, i.e. high-income and low-income taxis, when they are not occupied. Guo et al. (2011) and Andrienko and Andrienko (2011) explored the mobility patterns at an individual or an aggregated level. Interactive visualization systems are developed to support finding mobility patterns by visually exploring the pickup and drop off events and the statistics about the aggregates. HubCab[1] developed by MIT Senseable city lab allows users to get insight into the taxi mobility patterns at a fine granularity and supports future taxi sharing based on a model named "shareability networks" (Santi et al. 2014, Szell and Groß 2014). For example, the user can navigate to the places where his/her taxi trips start and end and to discover how many other people in his/her area follow the same travel patterns. NYC Taxi holiday visualization system[2] used 2013 NYC Taxi data to visualize traffic from JFK and LGA airports during the holiday season (Nov 15th to December 31st). Users can observe the traffic patterns and filter the visualization results by individual airlines or terminals.

Being inspired by many of the aforementioned studies, we investigate in this paper visual analytical methods for the exploration of spatiotemporal, semantic patterns of the traffic flows in/out of the transport hubs. The novelty of this work is that we propose a workflow that combines the data mining algorithms and visualization methods to inspect multivariate movement data of transport hubs at multiple levels. The workflow consists of four main steps. The preprocessing as the first step is targeted to reconstruct the

[1] http://hubcab.org/#13.00/40.7219/-73.9484

[2] http://taxi.imagework.com/

occupied trajectories from the GPS data and extract the pickup/drop-off events related to the transport hubs. The second step is to extract dense areas or significant places via spatial clustering method. The third step is to classify the significant places based on the POI information into different functional regions, e.g. public, commercial, residential and industrial regions. In the final step we design a graphic user interface to present the individual trajectories and events as well as the classification results and let users explore the mobility patterns in/out the transport hubs.

In Section 2, we present the proposed framework containing the workflow of data analysis. Section 3 is the experiment to test our proposed framework. A web-based interactive visual analytic system is developed to process and visualize massive taxi trajectories through transport hubs. In Section 4, we analyze and discuss our discoveries. Section 5 concludes the paper.

2. A Visual Analytics Framework

Movement data are complex in nature due to the inherent spatiotemporal, implicit semantic, and structural characteristics. In this paper, we propose a general framework and workflow to visually explore the mobility patterns of the flow in and out of the transport hubs. The focus is to allow users to explore at both individual and aggregated levels so that users can understand the movement patterns related to the transport hubs. The workflow is shown in Figure 1.

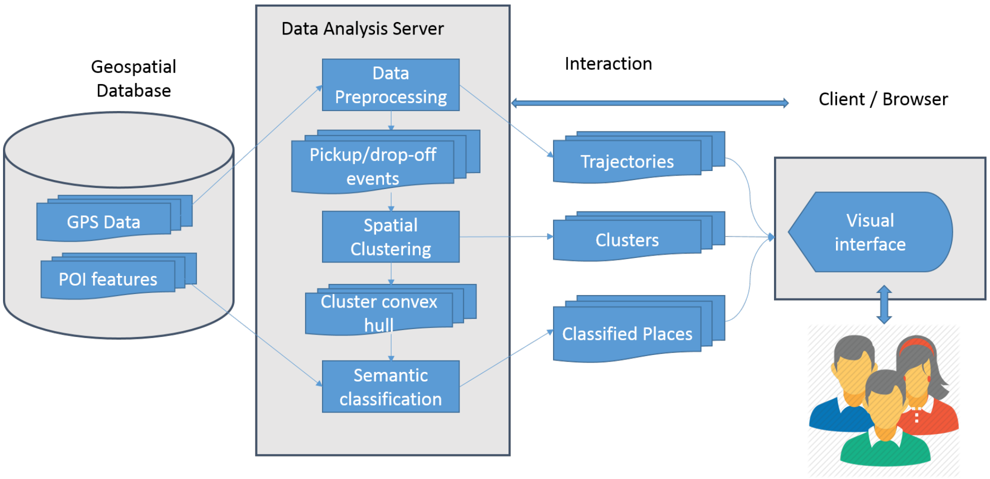


Figure 1. The framework of the visual analysis of traffic flow of transport hubs.

The workflow starts from the data preprocessing that aims to reconstruct from the GPS data in the geospatial database the occupied trajectories from/to the transport hubs and extract the pickup/drop-off events. The next step is based on the assumption that some places are more significantly connected, for instance through more events, with the transport hubs than the other places. Spatial clustering methods are applied to identify the dense areas or significant places from the extracted pickup/drop-off events. These places are represented by the convex hull polygons of the clusters. Furthermore, we classify the significant places into different categories of semantic meanings. This step is mainly based on the POI information inside the significant places. Finally a visual interface with interactive techniques is designed to enable the exploration of the spatiotemporal and semantic patterns. The following subsections describe in detail the individual steps.

2.1. Spatial Clustering

Spatial clustering can be used to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis (Han, Kamber and Tung 2001). In this work, we use spatial clusters to identify event-rich areas and as a preprocessing step to the semantic classification. Specifically, we use a hierarchical agglomerative clustering method (Kaufman and Rousseeuw 1990) to detect the significant areas or places where most of the pickup/drop-off events happen. The given set of data objects is hierarchically decomposed, forming a dendrogram - a tree that splits the database recursively into small subsets. The dendrogram can be formed either “bottom-up” or “top-down”. This work adopts the “bottom-up” way. The “bottom-up” approach, also called “agglomerative” approach, starts with each object forming a separate group. It successively merges the objects or groups according to some measures like the distance between the two group centers and this is done until a termination condition holds.

The reason we choose the hierarchical clustering method is that we can easily investigate the clustering results at multiple scales. In this work, the input of the clustering method is the individual pickup/drop-off events. The parameters of the methods are normally a distance function and a linkage criterion. The parameter setting is largely dependent on the applications or tasks. The output is the clustered events indicating which cluster the event belongs to. To filter out the significant places, we set a significant threshold value for the minimum number of the cluster. If a cluster has a number of elements larger than the significant threshold value, then this cluster is a significant cluster. The area bounded by the convex hull of the cluster is a significant place.

2.2. Semantic classification

With the significant places mined from long period taxi pick-up/drop-off data, semantics of these places can be further inferred. A few works have shed lights on mining semantics of the places. Yuan, Zheng and Xie (2012) introduced a topic modeling method that treats each functional region as a document and model the distribution of the vocabularies (i.e. POI) associated with certain topics. Zhu et al. (2013) shared their efforts on finding the best feature set for place classification using sensor data from mobile devices.

The challenges of place classification in our work is that 1) the spatial extension of the extracted places using spatial clustering is sensitive to the parameter setting of the clustering process, which often results in an imbalance pickup/drop-off distribution inside the places. 2) Only open source OSM POI data are used which lead to data quality bias to derive meaningful results. Being aware of these issues, we first tried Gaussian Mixture Model (GMM), which is a probabilistic model that intends to fit the data using a linear combination of Gaussian distributions. And our GMM-based method treats each place as a document and the type of POIs inside as vocabulary so as to infer the topics of the places. Unfortunately, GMM fails to produce consistent clustering results compared to the sample places we prepared for evaluation. Therefore, we adopt a simple rule-based classification based on the POI types (e.g. from the OSM tag information). The POI types are classified to four commonly used classes, namely public, commercial, residential and industrial. For example, POI types like “restaurant”, “hotel”, “company” are classified to “commercial”, and POI types like “embassy”, “station”, and “museum” belong to “public”. We examine the classified POIs inside each significant place and calculate which type is of the largest amount. Then, each place has a deterministic assignment to one of the classes with the largest POIs frequency.

2.3. Visual interface design

The visual interface is mainly based on a map view. It consists of three main components: (1) a map view showing the spatial distribution of the individual pickup/drop-off events, the reconstructed occupied trajectories, and the significant places inferred from the spatial clustering; (2) a circular histogram view located in the corresponding transport hub showing the semantics and the statistics of the classified significant places; (3) a clock chart view placed on the upper right corner of the interface to allow the visual exploration of different temporal patterns. It is worth to mention that the three components are interconnected and users can interactively click any of the components to simultaneously inspect the corresponding features highlighted in other components. For example, if the user selects a bar in the histogram view, the corresponding place will be highlighted so that the user knows where the significant place is and how the pickup/drop-off events distribute inside the place.

3. Experiment

3.1. Test data and data pre-processing

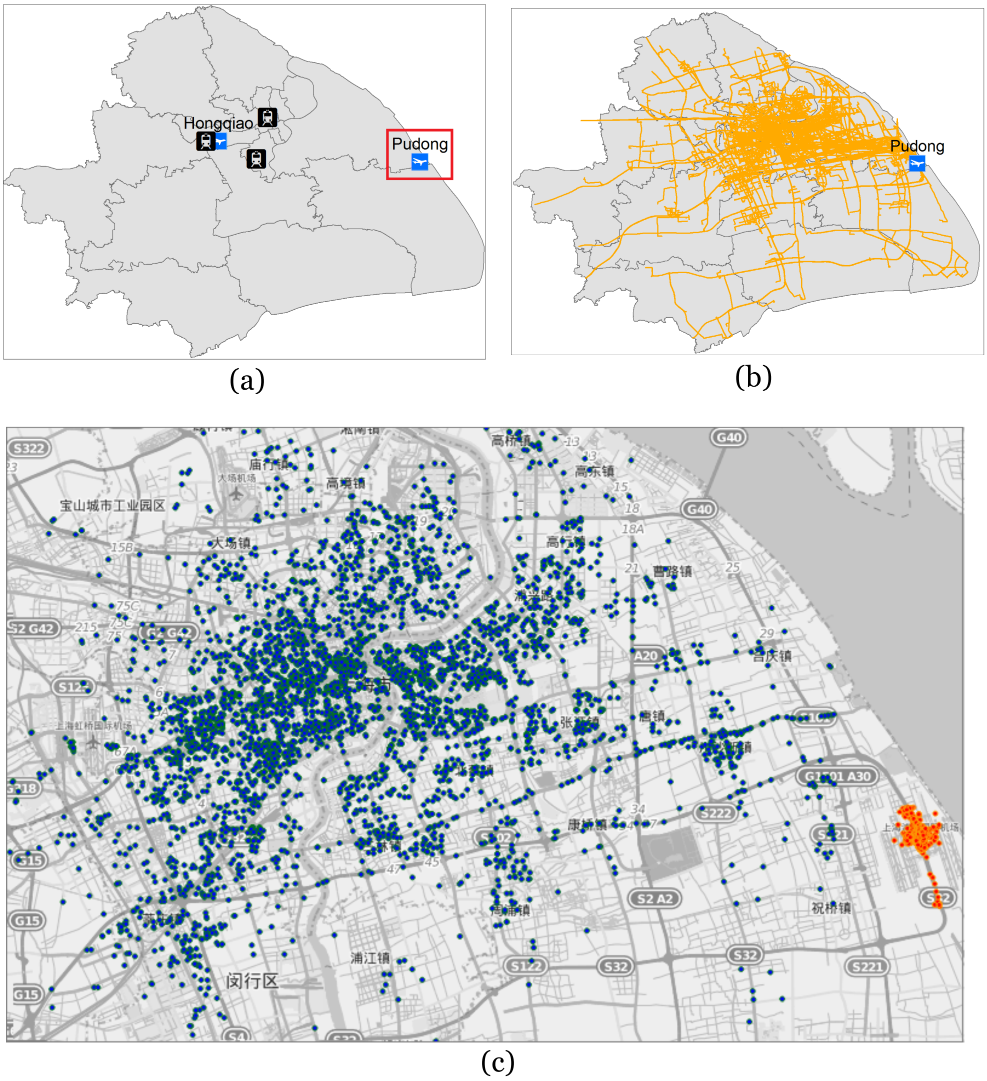
The test dataset are temporally ordered position records collected from about 2000 GPS-enabled taxis within 47 days from 10th May to 30th June 2010, in Shanghai. The temporal resolution of the dataset is 10 seconds and thus theoretically around 8000 GPS points of each car would be recorded in one day (24 hours) given the GPS device effective. Each position record has nine attributes, i.e. car identification number, company name, current timestamp, current location (longitude, latitude), instantaneous velocity, and car-status (meaning taxi occupied or empty). The detailed description of the fields is shown in Table 1. The data are stored in a MongoDB database, which provides geo-spatial features.

Field	Example field value	Field description
Date	20100517	8-digit number, yyyymmdd
Time	235903	6-digit number, HHMMSS
Company name	QS	2-digit letter
Car identifier	10003	5-digit number
Longitude	121.472038	Accurate to 6 decimal places, in degrees
Latitude	31.236135	Accurate to 6 decimal places, in degrees
Velocity	16.1	In km/h
Car status	1/0	1-occupied; 0-unoccupied

Table 1. Description of the fields of floating car data.

This paper focuses on the occupied trajectories (with passengers) from and to the transport hubs. Occupied trips can be reconstructed by connecting the temporal sequences of GPS points with the “car status” attribute value of 1. Pickup and drop-off points correspond to the starting and ending points of the occupied trips respectively.

We extracted from the original dataset several important transportation hubs. In this work, Pudong international airport is chosen as our study area. Figure 2a shows the location of the Pudong international airport. Figure 2b shows about 6000 occupied trajectories extracted from the floating car data originating from Pudong airport. Figure 2c shows the distribution of the pick-up and drop-off events of these trajectories. The orange and blue dots indicate respectively the pickup and drop-off points.



(a)

(b)

(c)

Figure 2. (a) The location of Pudong international airport in Shanghai. (b) The occupied trajectories starting from Pudong airport. (c) The corresponding pickup (in yellow) and drop-off (in blue) events.

3.2. Hierarchical clustering of the drop-off events

In this section we use a hierarchical agglomerative clustering method to extract dense regions where the drop-off events from Pudong airport (shown in Figure 2(c)) very often happen.

The two parameters as the input of the clustering method are a distance function and a linkage criterion. Based on our experiment and empirical knowledge, a Euclidian distance threshold of 500m and an average linkage criterion are applied. Since the clustering method will assign each point to a cluster and here we are interested only in clusters with higher density or a larger number of cluster elements, we select clusters with more than 10 points as salient or significant clusters. The number of representative clusters is 122. Figure 3 illustrates the resulted clusters.

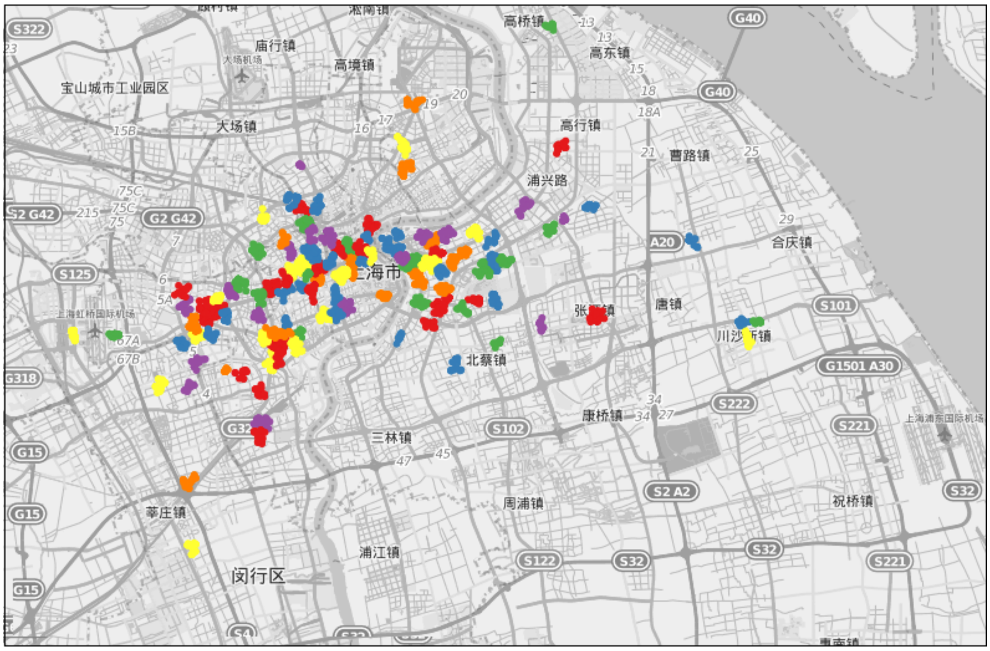


Figure 3. The representative clusters of drop-off events resulting from the clustering method.

Furthermore, for each cluster, we calculate its convex hull and extend the convex hull polygon with a 100m buffer to represent the spatial extent of the cluster and is a significant place.

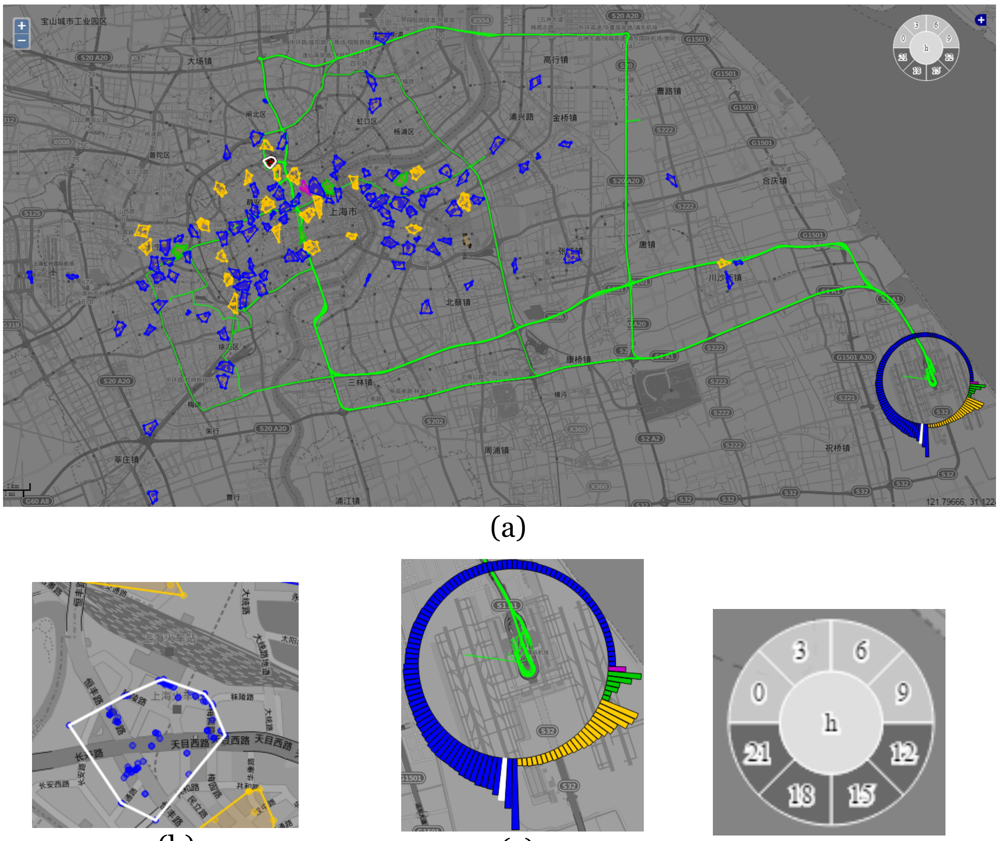
3.3. Semantic classification

According to the third step of the framework, firstly, we extract the current POI data from OpenStreetMap[3] within the convex hull polygons from the previous step. We then classify the POI types to four large categories, namely, commercial, public, residential and industrial. For each of the 122 polygons, we calculate which category with the largest proportion and assign to our polygons or significant places the category. Therefore, the places now are in four categories.

[3] www.openstreetmap.org

3.4. Interactive visual interface

A web-based graphic user interface (shown in Figure 4(a)) is designed based on the aforementioned idea and consists of three interlinked components, namely a map view, a histogram view (bottom right) and a clock view (upper right). The interface is implemented using OpenLayers library[4].



(a)

(b)

(c)

(d)

Figure 4. (a) The screenshot of the proposed graphic user interface; (b) the highlighted significant place; (c) the circular histogram view; (d) the clock view.

On the map view, the irregular colourful polygons represent the classified significant places. The color of a polygon indicates the semantic meaning of this place. Blue, yellow, green and magenta represent the categories of public, commercial, residential and industrial respectively. For instance, if a polygon is in yellow, the place is regarded as a commercial place. The dots

[4] http://openlayers.org/

inside each polygon represent the individual drop-off events. The light green lines represent the taxi trajectories from Pudong airport to the railway station.

The circular histogram view is superimposed around the Pudong airport, consisting of compact individual bars. Each bar represents a significant place mined from the origin/destination locations. Its color indicates its corresponding category, while the height represents the amount of drop-off events inside each cluster. The bars are ordered by their categories and their height values. This structured way allows the user to easily perceive the relative amount of significant places in each category and the significance level of each place.

The clock chart view with eight sections in the upper right corner shows a three-hour time slot. The user can switch from one time slot to another to observe the temporal patterns. The default visualization result shows the all-day long data.

The three views are interconnected and have thus maximized the freedom of exploring the spatiotemporal and semantic information. For example, when the user selects a bar in the circular histogram view, on the map view the corresponding polygon and the points inside would be highlighted. The user can zoom into the place and examine in detail the distribution of the actual drop-off events. Meanwhile, the movement trajectories from the Pudong airport and the significant places (i.e. destinations) will be shown on the map, enabling the user to get insight into the movement traces. Similarly, when the user selects one interesting place, the corresponding bar on the histogram view could be highlighted. The user then can get a general idea about the significance level of the place and the approximate amount of the drop-off points inside this place.

4. Analysis and Discussion

With help of the visualization, we can immediately perceive on the map the overall spatial distribution of the significant places, and their categories and the associated statistics from the circular histogram view. The spatial clusters or significant places are mainly distributed around the city centre with a few scattered far from the centre. With regard to the circular histogram view, the first impression is that there are two large groups of bars in blue and yellow, meaning that the majority of the significant places are “public” and “commercial”. A small group of dark green bars represent the residential category and there is only one significant place in magenta classified as “industrial”.

The interactive operations empower the visual exploration of the spatial and semantic relationships between the significant places and the transport hub. In Figure 4, a bar of the category “public” in the histogram view is highlighted in white. The corresponding significant place, i.e. Shanghai railway station in the upper middle of the screenshot, is highlighted as well. The trajectories from Pudong airport to the highlighted Shanghai railway station are shown in light green. The relative brightness of the green lines reflects the relative amount of trajectories on the road segments. The user can thus obviously perceive the frequent driving routes from the airport to the railway station.

Besides, we can also explore the significant places in detail by combining with the map context. For example, by zooming into the highlighted polygon of the railway station in Figure 4, the user may find that the drop-off events inside are unevenly distributed. Moreover, the user can examine the drop-off events based on the map context, for example, in combination with the spatial distribution of the inner POIs (e.g. hotels, restaurants) and road networks. Based on the observations, the user might be aware of the underlying structure and refine their understanding of the deterministic classification results. For instance, inside the polygon of the railway station, there are a large number of public POIs, e.g. stations, but also a certain amount of commercial POIs, e.g. hotels. Even the deterministic classification result of the polygon is “public”, it contains a large portion of features belonging to the “commercial” category.

5. Conclusion and Future Work

In this paper, we propose a general workflow of visual analysis of traffic flows of transport hubs. It mainly consists of a pre-processing step, a spatial clustering, a semantic classification procedure, and a visual interactive system. We test our framework using the movement data of a transport hub Pudong international airport in Shanghai. The experiment results show that our proposed method is feasible. Users can get a general idea of where are the most significant places of the drop-off events, what are their semantic meanings, and which routes are frequently taken from the airport to a significant place.

The following improvements are anticipated in the future. Firstly, the system will be elaborated to allow users to interactively set spatial clustering parameters. For instance, currently a distance of 500m is set in the hierarchical clustering method, which is chosen according to our empirical knowledge. An interactive selection of the parameter could help adjust the parameters and generate clusters with a reasonable spatial extent. Secondly, we need high quality POI information. In this work, the classification method suffers largely from the bad quality of the POI data, mainly due to the incompleteness and inaccuracy of the current OSM data in Shanghai. Besides, the user interface design and the visualization method can be improved. An iterative refinement of the whole process is needed. Moreover, we will include the temporal analysis and investigate whether there are hourly or daily patterns of the traffic flows in/out of the transport hubs. We believe that this refinement is necessary and will be helpful for planners to understand the individual and collective behaviors of interacting with the urban space in metropolitan areas.

Acknowledgement

The first two authors are funded by the China Scholarship Council (CSC). We would like to thank Prof. Chun Liu for sharing with us the Shanghai Taxi FCD dataset.

References

Han, J., M. Kamber & A. K. H. Tung. 2001. Spatial Clustering Methods in Data Mining: A Survey. In *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS,* eds. H. J. Miller & J. Han. Taylor and Francis.

Andrienko, N. & G. Andrienko (2011) Spatial generalization and aggregation of massive movement data. *Visualization and Computer Graphics, IEEE Transactions on,* 17, 205-219.

Ding, L., H. Fan & L. Meng. 2015. Understanding taxi driving behaviors from movement data. In *AGILE Conference on Geographic Information Science*. Lisbon.

Guo, H., Z. Wang, B. Yu, H. Zhao & X. Yuan. 2011. TripVista: Triple Perspective Visual Trajectory Analytics and its application on microscopic traffic data at a road intersection. 163-170.

Kaufman, L. & P. Rousseeuw. 1990. *Finding Groups in Data*. Wiley & Sons, Inc., New York.

Liu, Y., F. Wang, Y. Xiao & S. Gao (2012) Urban land uses and traffic ‘source-sink areas’: Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning,* 106, 73-87.

Santi, P., G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz & C. Ratti (2014) Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences,* 111, 13290-13294.

Szell, M. & B. Groß. 2014. Hubcab – Exploring the Benefits of Shared Taxi Services. In *Decoding the City How Big Data Can Change Urbanism.*

Yuan, J., Y. Zheng & X. Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 186-194. ACM.

Yuan, N. J., Y. Zheng, L. Zhang & X. Xie (2013) T-finder: A recommender system for finding passengers and vacant taxis. *Knowledge and Data Engineering, IEEE Transactions on,* 25**,** 2390-2403.

Zheng, Y. & X. Zhou. 2011. *Computing with spatial trajectories*. Springer Science & Business Media.

Zhu, Y., E. Zhong, Z. Lu & Q. Yang (2013) Feature engineering for semantic place prediction. *Pervasive and Mobile Computing,* 9**,** 772-783.