A Conceptual Model for Representation of Taxi Trajectories

Ana Maria Amorim e Jorge Campos

Grupo de Aplicações e Análises Geoespaciais – GANGES
Mestrado em Sistemas e Computação – UNIFACS
Salvador, BA – Brazil
ctamp2000@yahoo.com.br, jorge@unifacs.br

Abstract. *The large-scale capture of data about the motion of moving objects has enabled the development of geospatial tools to analyze and present the characteristics of these objects' behavior in many different fields. Intelligent Transportation Systems, for instance, make intensive use of data collected from embedded in-vehicle devices to analyze and monitor roads conditions and the flow of vehicles and passengers of the public transportation system. The taxi fleet is an important transport modality complementary to the public transportation system. Thus, analysis of taxis' movements can be used to capture information about the condition of the traffic and to understand at a finer level of granularity the movement of people in an urban environment. This paper addresses the problem of mapping taxi raw trajectory data onto a more abstract and structured data model. The proposed data model aims to create an infrastructure to facilitate the implementation of algorithms for data mining and knowledge discovery about taxi movements and people's behavior using this means of transport.*

1. Introduction

With the evolution of technology, large-scale capture of data about the motion of moving objects has become technically and economically feasible. As a result, there are a growing number of new applications aiming at understanding and managing complex phenomena involving these objects.

Intelligent Transportation Systems (ITS) encompass new kind of applications designed to incorporate information and communication technologies to the transportation infrastructure. The main goal of such applications is to allow users to become more acquainted with the system functioning and to provide innovative services to enhance the system's coordination and maintenance. ITS make intensive use of data collected from sensors placed along the transportation network or embedded in-vehicle devices to analyze and monitor roads conditions and the flow of vehicles and users of the public transportation system. Although the taxi fleet cannot be considered as a component of the public transportation system, it is an important and complementary

transport modality. Thus, the analysis of taxis' movements can be used to capture information about the condition of the traffic and to understand at a finer level of granularity the movement of people in an urban environment.

In the ITS arena, data about vehicles' movements are usually stored in the form of tuples (identifier, location, time) describing the evolution of a vehicle position over time. This kind of data, however, does not meet the requirements of many applications interested in capturing the characteristics of the movement, patterns or anomalies in vehicles' behavior. These applications often enrich trajectory data with contextualized information about the environment, such as road conditions, landmarks or major cultural or sport events [Spaccapietra et al. 2011] [Bogorny et al. 2011] [Yan 2009] [Alvares et al. 2007]. Other kinds of contextualized information must be gathered during the data acquisition process and require special sensors or the direct interference of a human being. The latter case applies to the trajectory of taxis.

In order to illustrate a typical process of data acquisition about taxis movements, consider, for instance, that all taxis are equipped with a mobile device with embedded location mechanism and a mobile application capable of registering the path of all trips throughout the day and some relevant events. Once started, the application begins to collect and communicate data about vehicle's location and status (i.e., full or empty). Whenever the driver picks up a passenger, he/she should press the *pick-up* button and report the number of passengers that has boarded the vehicle. At the end of the trip, the driver must press the *drop-off* button indicating that the taxi has no passengers and it is available for a new trip.

The formidable and massive dataset generated by taxis movements, however, is barely used for analysis, data mining or knowledge discovery purpose. The major drawbacks of using these data are twofold. First, trajectory data lack a more abstract structure, have lots of redundant or inconsistent records, and carry little or no semantic information. Second, there are few algorithms tailored to analyze, mine and reveal patterns of this special kind of moving object. This paper addresses the problem of mapping the raw data about taxis trajectories onto a generic conceptual model. This model aims to facilitate queries and extraction of knowledge about the dynamics of this transport modality in major urban areas. By structuring taxis' raw trajectory data through more abstract entities, we intend to create the data infrastructure necessary to implement algorithms that identify patterns of people using taxi as a means of transport; show the characteristics of the movements of passengers by taxi from the city, its origins and predominant destinations; analyze the efficiency of the taxi system at different periods; among others.

The remainder of this paper is structured as follows: section 2 discusses related work. Section 3 presents some basic definitions used to define the conceptual model. Section 4 discusses the main entities of a model to represent the data trajectory of taxis. Section 5 discusses possible applications of the model and presents some conclusions.

2. Related Work

The study of the movement of taxis aiming at understanding and improving urban mobility has become an active research field. This section discusses some works that address different aspects of this problem.

[Peng et al. 2012] presented an analysis of the taxi passengers' movement in Shanghai, China. This study found out that on weekdays people use the taxi mainly for three purposes: commuting between home and workplace, traveling between different business places, and going for other places for leisure purpose.

[Veloso et al. 2011] analyzed the movement of taxis in Lisbon, Portugal. In this work it is possible to visualize the spatiotemporal distribution of the vehicles, most frequent places of origin and destination at different periods of the day, the relationship between these locations and peaks of the system usage. This paper also analyzes the taxi behavior in what they called downtime (i.e., the time spent by the taxi driver looking for the next passenger) and conducts a study of predictability to locate the next passenger.

[Kamaroli et al. 2011] presented a methodology to analyze passengers' movement at the city of Singapore at different periods of the day. The main objective of this study was to quantify, visualize and examine the flow of taxis considering only information about origin and destination.

The objective of [Zheng et al. 2011] was to detect flaws in the Beijing urban planning based on information derived from the analyses taxis trajectories. As a result, they identified regions with significant traffic problems and diagnosed failures in the structure of links between these regions. Their findings can be used, for instance, by urban planners to propose the construction of a new road or a new subway line.

In [Yuan et al. 2011] was presented a recommendation system with suggestions for taxi drivers and passengers. Drivers use the system to identify locations in which the probability to pick-up passengers is high. Passengers use the system to identify places in a walking distance where they can easily find an empty taxi. These suggestions are based on the patterns of the passengers (i.e., where and when they usually get in and out of taxis) and the strategy used by most taxi drivers to pick-up passengers.

In [Ge et al. 2011] was presented an intelligent taxi system able to explore data collected from taxis in the city of San Francisco and New York for commercial purpose. The authors argue that the system increases the productivity of taxi drivers with routes recommendations, identifies fraud in the taxi system, and provides support for new business ideas.

In [Liu et al. 2009] was presented a methodology to analyze the behavior of taxi drivers in Shenzhen, China. They proposed a metric to measure drivers' skill, in what they called "mobility intelligence". Considering their income and behavior, taxis drivers are ranked as *top drivers* or *ordinary drivers*. The paper concluded that while ordinary drivers operate in fixed locations, the top drivers choose the places according to the most opportune time.

The goals of these works illustrate only some interesting possibilities of processing taxi trajectories data. The possibilities are endless, but they reveal a growing interest in the area. Considering the data used to support their analyses, all related work use raw trajectory data complemented with pick-up/drop-off information. In [Yuan et al. 2011], [Ge et al. 2011] and [Liu et al. 2009] the number of passengers is also considered. Considering the data models used to represent this dataset, however, all work use *ad hoc* data models to solve a specific problem or to carry out a particular analysis. At the best of our knowledge, no generic data model capable of supporting a wide range of analysis and knowledge discovery has been identified yet. The following sections present our contribution to this area.

3. Basic Definitions

[Spaccapietra et al. 2008] proposed the first model that treats trajectories of moving objects as a spatiotemporal concept. Spaccapietra conceptualized a trajectory as a space-time evolution of a traveling object to reach a certain goal. The trajectory is bounded by two instants of time (*Begin* and *End*) and an ordered sequence of pairs (point, time) representing the movement of the object. Semantically speaking, Spaccapietra considers a trajectory as an ordered list of *Stops* and *Moves*. A *Stop* is part of a trajectory that is relevant to the application in which the travelling object did not move (i.e., the object remains stationary for a minimal amount of time). The trajectory's *Begin* and *End* are not considered *Stops*, because their temporal extent is a single *chronon* (indivisible time unit). A *Move* is a sub-trajectory between two *Stops*, between the starting point of the trajectory (*Begin*) and the first *Stop*, or between the last *Stop* and the ending of the trajectory (*End*). The spatial representation of a *Stop* is a single point, while a *Move* is represented by a displacement function or a polyline built with trajectory's points.

Based on Spaccapietra's work, we present some relevant definitions to the model aimed to represent taxis daily trajectories.

Definition 1 *Working Trajectory* represents the evolution of the position of a taxi along the working hours of its driver.

Definition 2 *Full-Move Sub-Trajectory* corresponds to a segment of a *Working Trajectory* and represents the trajectory of the taxi while occupied by a passenger.

Definition 3 *Empty-Move Sub-Trajectory* corresponds to a segment of a *Working Trajectory* and represents the trajectory of the taxi in search of a new passenger.

Definition 4 *Pick-Up Point* indicates the time and location of the beginning of a *Full-Move Sub-Trajectory*, i.e., it represents the time and place of the start of a taxi's travel with passengers.

Definition 5 *Drop-Off Point* indicates the time and location of the end of a *Full-Move Sub-Trajectory*, i.e., represents the time and the location where the passenger leaves the taxi.

Definition 6 *Taxi Stop Point* is a known geographic location of a point where the taxicab remains stationary for a certain period of time waiting for passengers.

Working Trajectory is equivalent to Spaccapietra's concept of a travelling object trajectory. A *Working Trajectory* is split on semantically meaningful specialization of *Stops* and *Moves*. *Full-Move* and *Empty-Move Sub-Trajectory* correspond to *Moves* and a *Taxi Stop Point* corresponds to a *Stop*. *Pick-Up Point* and *Drop-Off Point* do not represent a *Stop*. They are equivalent to the endpoints of our sub-trajectories. Different from Spaccapietra's conceptualization, a *Working Trajectory* is not an alternate sequence of *Stop* and *Moves*. A *Working Trajectory* can have any combination of *Full-Move Sub-Trajectory*, *Empty-Move Sub-Trajectory* and *Taxi Stop Point*. These definitions are the basis for the understanding of a conceptual model aimed to represent the movement of taxis. This model will be presented in the next section.

4. A Conceptual Model for Taxi Trajectories

Before discussing the representation of taxi trajectory with high-level entities of a conceptual model, it is interesting to illustrate the process of capturing raw data by following a typical working day of John, a taxi driver. John begins his workday by logging to an application installed on a device with an integrated GPS and connected to a 3G network. The application sends the position of the vehicle at every minute and, eventually, allows the registration of some relevant events. After the initial setup, John

starts to drive his taxi in search of the first passenger of the day. After driving several blocks, a truck maneuvering forces John to stop and wait a few minutes. After the truck's maneuver, John continues his journey in search for passengers. Few miles away, three passengers take the taxi and ask John to go to the bus station in downtown. At this moment, John registers in his application the fact that three passengers have boarded. Near the bus station, a car accident forces John to wait a few minutes until the complete desobstruction of the road. At the bus station, the passengers exit the taxi and John records this fact in his application. Fortunately, at the same place there is a couple who immediately boarded the taxi. The couple is going to a meeting at a company near downtown. After drop-off the couple at their destination, John drives for a few minutes around downtown and decides to stop at a taxi stop to wait for the next passenger. After a few minutes, a passenger boards the taxi and asks for a trip to a suburban neighborhood. After drop-off the passenger, John goes to another taxi stop and stays there few hours waiting for passengers. Finding out that the strategy to wait in this taxi stop was not a good choice, John decides to search for passengers in the neighborhood. After a fruitless search, John decides that it is time to stop and finish his workday.

The raw data generated by the short working time of the taxi driver are shown in Figure 1. The cloud of points represents raw data captured by the GPS device. The continuous arrows indicate pick-up and drop-off events register by the driver using the mobile application. The dashed arrows indicate some external events experienced by the driver. These events were not reported, thus they are not part of the taxi trajectory raw data.

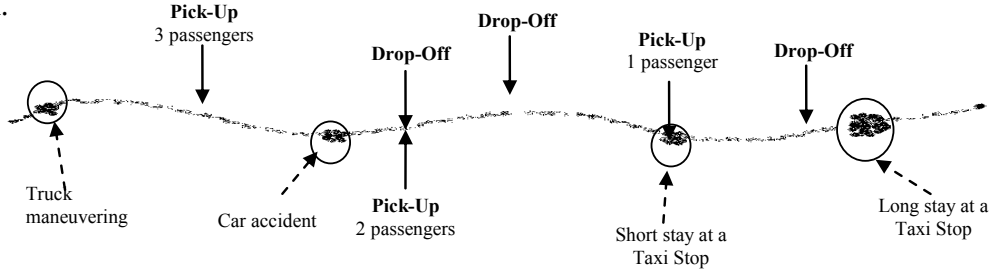

Figure 1. Raw trajectory data of a typical taxi driver working day.

Raw trajectory data and events registering pick-up and drop-off points are useless for most applications interested in analyzing the movement of this transportation mode. Thus, a conceptual data model is essential to represent relevant aspects of the movement with more abstract and semantically meaningful entities.

The model to represent the movements of taxis is based on the entity *Working Trajectory*. This entity represents the movement of a taxi driver during his/her workday. A *Working Trajectory* (WT) has attributes identifying the driver, the vehicle and two

instants representing the beginning and end of the taxi driver workday. The combination vehicle-driver defines our moving object. This combination is required in order to identify everyday situations experienced by taxis fleet companies, in which many taxi drivers drives the same vehicle or in situations where the driver works for more than one Taxi Company. Besides the atomic attributes mentioned above, a *Working Trajectory* has also a composition of *Full-Move Sub-Trajectory*, *Empty-Move Sub-Trajectory* and *Taxi Stop Point*. Figure 2 shows the class diagram in UML style representing the relationships between entities of the model.

The entity *Full-Move Sub-Trajectory* (FMST) represents parts of a taxi trajectory while travelling with passengers. This entity has four attributes: an integer attribute to indicate the number of passengers who has boarded; two attributes of type *STPoint* indicating the start and end points of a taxi trip; and an attribute of type *STLine* to represent the path of the trip. The type *STPoint* represents a point in the space-time dimension. This type represents the position of a moving object and has an attribute to register the spatial location of the object and an attribute to associate the instant at which the object occupies that position. The type *STLine* represents an arbitrary non-empty collection of *STPoints*. Two operations of *STLine* type that deserve to be mentioned are length and boundingBox, which return the length and the wrap rectangle of the trajectory, respectively.

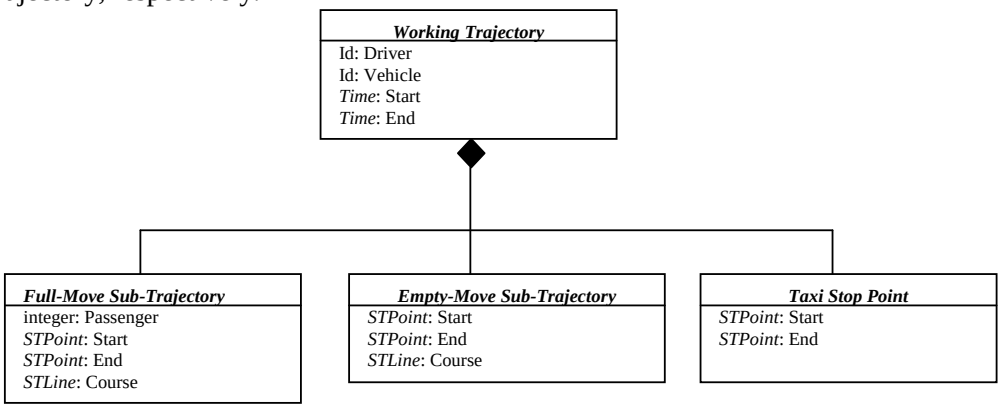

Figure 2. Class diagram with entities of the model to represent taxis trajectories.

There is no spatial dependence between the attributes *Start* and *End* of a *Full-Move Sub-Trajectory*, i.e., they can represent any point in space and may even be the same point in a hypothetical journey where the passenger returns to the same location in a round trip. In the temporal dimension, however, the final instant of the trajectory succeeds the initial instant.

The entity *Empty-Move Sub-Trajectory* (EMST) represents parts of a taxi trajectory while the vehicle is travelling without passengers. This entity is similar to

Full-Move Sub-Trajectory, differing only by the lack of the *Passenger* attribute. The starting point of an *Empty-Move Sub-Trajectory* can be spatially identical to the endpoint of a *Full-Move Sub-Trajectory* or a taxi stop location. Likewise, the end point of an *Empty-Move Sub-Trajectory* can be spatially identical to the starting point of a *Full-Move Sub-Trajectory* or a taxi stop location.

The entity *Taxi Stop Point* (TSP) represents parts of the taxi trajectory in which the taxi driver had stopped at a known location to wait for the next passenger. This entity does not have an associated trajectory. Thus, only two *STPoint* attributes are enough to record the location and time of this event. The spatial information stored in the *start* and *end* attributes must be the same geographic location of a known taxi stop point. On the temporal domain, the initial and final moments indicates duration of the wait. On the spatial domain, the distance between the start and end points gives a rough idea of the length of the queue.

In addition to the spatial and temporal constraints already mentioned, the composition of entities of a *Working Trajectory* has an additional restriction, that is, there are no two consecutives *Empty-Move Sub-Trajectory*. An *Empty-Move Sub-Trajectory* must be intermingled with *Full-Move SubTrajectories* or *Taxi Stop Points* or be the first or last entity of a *Working Trajectory*.

The next step is to convert raw trajectory data (Figure 1) into entities of the conceptual model (Figure 2). An instance of a *Working Trajectory* is created to represent John's workday. At this point only atomic attributes are filled with the identity and time duration of the trajectory. Details about the trajectory are built upon the processing of trajectory raw data and pick-up and drop-off events. Before creating the instances of the composite entities, the raw data of the trajectory goes through a cleaning process to eliminate redundant information and keep only the information needed for the representation of the trajectory of the vehicle [Bogorny et al. 2011]. At this stage, some points that indicate the vehicle *Stops* and *Moves* are also identified [Bogorny et al. 2011] [Palma et al. 2008].

According to [Spaccapietra et al. 2008], the fact that the position of the object be the same for two or more consecutive instants does not define that position as a *Stop*. A *Stop* is a relevant situation to the application. In our model, we are interested in *Stops* indicating where and when a taxi driver stops at a known location (i.e., a taxi stop) waiting for a passenger. Therefore, clusters of points that occur during a period when the cab was busy are simply discarded. This is the case when John stops because a car accident (Figures 1 and 3.a). Moreover, *Pick-up* and *Drop-off Points* are also not

represented as *Stops* (i.e., modeled as first-class entities). These entities represent the end points of a *Full-Move Sub-trajectory* and are represented by two attributes of the type *STPoint* in the entity *Full-Move Sub-Trajectory*. *Stops* that occur during the period where the taxi is empty are different. They can be either a stop in a taxi stop point or a stop due to an external event. The former is of our interest and the latter will be also discarded. Thus, *Stops* that occur while the taxi is empty are marked as candidates to represent *Taxi Stop* entities (Figure 3.a). The decision whether these candidates are actually a *Taxi Stop* is done in a next step.

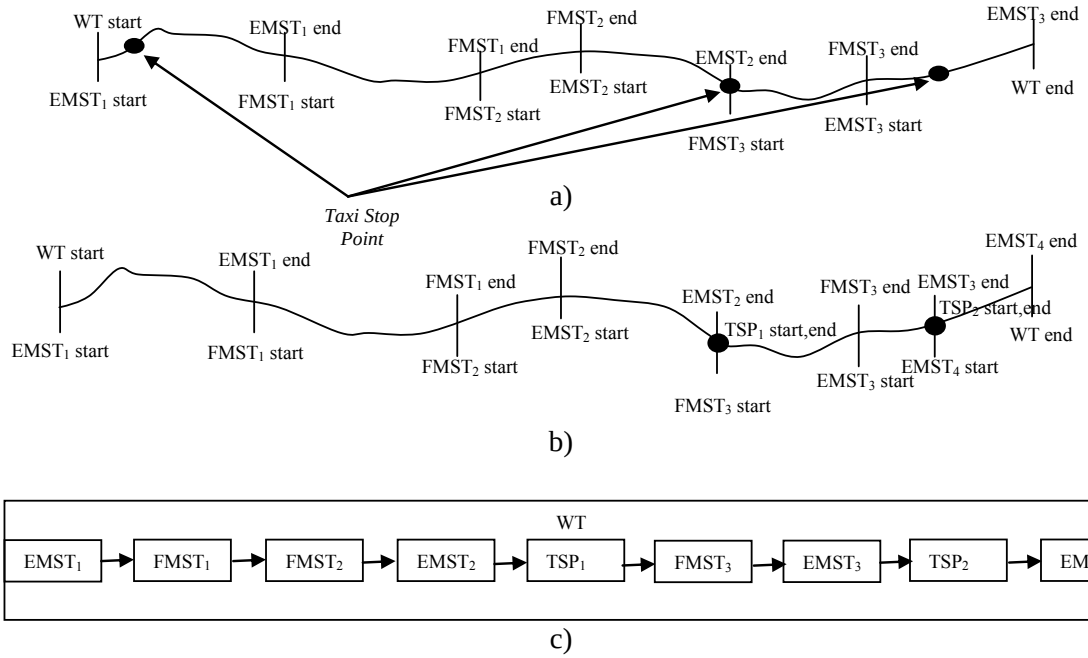

Figure 3. Steps in the process of creating entities of taxi trajectory conceptual model. a) identification of taxi stop points candidates and a first attempt of sub-trajectories entities; b) identification of real taxi stop locations; and c) object diagram with all entities of John's workday.

The *Moves* identified in this phase create either an *Empty-Move Sub-trajectory* or a *Full-Move Sub-trajectory*. This decision is based on instants and locations of pick-up and drop-off events reported by the driver. At this point the collection of raw trajectory that forms the course of each sub-trajectory is also captured by the model's entities. Thus, based on data captured during John's journey, three *Empty-Move Sub-trajectory*, three *Full-Move Sub-trajectory* and three candidates for *Taxi Stop Point* were created (Figure 3.b).

The last step in the process of creating entities of the conceptual model is the identification of what is really a *Taxi Stop Point*. The main problem in identifying this

entity is the distinction between *Stops* at a known taxi stop location and *Stops* that occur during an *Empty-Move Sub-trajectory* caused by external facts. The latter type of stop may be caused, for example, by a traffic jam or mechanical problem on the vehicle and it is not of our interest, thus it is not explicitly represented in the model. For this purpose, we use the approach developed by [Yuan et al. 2011]. They use the concept of point of interest and known taxi point location to discard *Stops* candidates that are not a real taxi stop.

For the data used in our example, the first *Taxi Stop Point Candidate* was rejected and the last two was identified as a true *Taxi Stop Point* (Figure 3b). With the creation of the last *Taxi Stop Point*, the third *Empty-Move Sub-Trajectory* was divided into two *Empty-Move Sub-Trajectory* entities with a *Taxi Stop Point* in-between. At the end of the raw data processing, a *Working Trajectory* composed by an ordered list of entities of the type *Empty-Move Sub-trajectory*, *Full-Move Sub-trajectory* and *Taxi Stop Point* is created (Figure 3.c).

We choose to not consider an event indicating when the driver stops at Taxi Stop. We believe that different from our example, all information reported by the driver can be completely automated. A taximeter connected to a data network, an embedded GPS device, and a load sensor, for instance, can send *pick-up* and *drop-off* information and an estimated number of passengers in the vehicle with no driver intervention. The stops at taxi stops points, however, cannot be determined using this technology.

5. Conclusion and Future Work

This paper introduces a conceptual model to represent taxi raw trajectories. Unlike the bus, train and subway systems that have pre-defined routes and stop points, taxis pick-up and drop-off passengers wherever they want. This capillarity allows a precise determination of people's origin and destination.

The taxi conceptual model aims to facilitate the task of querying, analyzing, data mining and performing knowledge discovery about this transport mode. It was shown a technique to create entities of the conceptual model based on raw trajectory data. The conceptual model is quite broad and can be used in many types of applications. Public managers, for example, may be interested in identifying a pattern in the behavior of users of the taxi system or to identify a need for a new bus itinerary. Fleet managers may be interested in measuring the efficiency of a taxi driver through the time spent without passengers. Users may be interested in know places with high probability of finding an empty taxi.

Entities of the conceptual model carry semantic information about taxis' movements, which facilitate the implementation of data mining and knowledge discovery algorithms at different levels of granularity. At a low level of granularity, historical data of taxis movement can be analyzed through the whole course of their trajectory. Analyzing the courses of all *Full-Move Sub-Trajectories*, for example, it is possible to know if the taxi took the shortest route from the origin to its destination, the time taken to complete the trip, and the traffic conditions along the route. The courses of all *Full-Move Sub-Trajectories*, *Empty-Move Sub-Trajectories*, and *Taxi Stop Points* of a certain driver can be used to highlight the driver strategy and efficiency. The efficiency of a taxi driver can be measured, for instance, by the ratio between the sum of the duration of all *Empty-Move Sub-Trajectories* and *Taxi Stop Points* over the duration of the entire journey of the driver or by the ratio between the sum of the length of the courses of all *Full-Move Sub-Trajectories* and all *Empty-Move Sub-Trajectories*. The former mechanism uses temporal information to measure the taxi efficiency, while the later uses spatial information. These indices can be combined to produce a spatial-temporal index of efficiency.

At a high level of granularity, the movement of the taxis can be used to identify, for instance, mostly wanted origin and destination places along the day and places where taxis are in great demand. By analyzing the start and end points of all *Full-Move Sub-Trajectories*, it is possible to map all pick-up and drop-off points and to identify where these hot spots are likely to occur along the day.

The importance of studying taxis movement is not restricted to the analysis of the historical data. Considering that the information about taxis' position, speed and status is published in real time, it can be used to identify empty taxis in a given neighborhood. This information can be used by a taxi company to dispatch the closest taxi in response to passenger call or by passengers viewing all available taxis on a map displayed on the screen of their Smartphone. Moreover, the average speed of thousand of vehicles crossing the city gives an excellent overview of traffic conditions at different locations along the road network, serving for any driver looking for the best uncongested route and improving urban mobility.

The examples discussed above require historical data covering a significant amount of time. Thus, the volume of data to be processed is expected to be huge. As future work, we are planning to apply the ideas presented in this paper in a real case scenario, that is, work with real data from a cooperative or Taxi Company and to develop a tool to support different kind of analysis.

References

Alvares, L.O., Bogorny, V., Kuijpers, B., Fernandes, J.A., Moelans, B., Vaisman, A., (2007), "A Model for Enriching Trajectories with Semantic Geographical Information". ACM-GIS'07.

Bogorny, V., Avancini, H., de Paula, B., C., Kuplich, C., R., Alvares, L., O., (2011), "Weka-STPM: a Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization". Transactions in GIS.

Ge, Y., Liu, C., Xiong, H., Chen, J., (2011), "A Taxi Business Intelligence System". Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining

Kamaroli, N.Q.B., Mulianwan, R.P., Kang, E.P.X., Ru, T.J., (2011), "Analysis of Taxi Movement through Flow Mapping". IS415 – Geospatial Analytics for Business Intelligence.

Liu, L., Andris, C., Biderman, A., Ratti, C., (2009), "Uncovering Taxi Driver's Mobility Intelligence through His Trace". SENSEable City Lab, Massachusetts Institute of Technology, USA.

Palma, A.T., Bogorny, V., Kuijpers, B., Alvares, L.O., (2008), "A Clustering-based Approach for Discovering Interesting Places in Trajectories", ACM Symposium on Applied Computing (SAC'08), Fortaleza, Ceará, Brazil.

Peng, C., Jin, X., Wong, K-C., Shi, M., Liò, P., (2012), "Collective Human Mobility Pattern from Taxi Trips in Urban Area", PLoS ONE 7(4):e34487. doi:10.1371/journal.pone.0034487.

Spaccapietra, S., Parent, C., Damiani, M.L., Macedo, J.A., Porto, F., Vangenot, C., (2008), "A Conceptual View on Trajectories", Data and Knowledge Engineering, 65(1): 126 – 146, 2008.

Spaccapietra, S., Chakraborty, D., Aberer, K., Parent, C., Yan, Z., (2011), "SeMiTri : A Framework for Semantic Annotation of Heterogeneous Trajectories", EDBT 2011, March 22–24, 2011, Uppsala, Sweden.

Veloso, M., Phithakkitnukoon, S., Bento, C., (2011), "Urban Mobility Study using Taxi Traces", Proceedings of the 2011 international workshop on Trajectory data mining and analysis TDMA 11 (2011).

Yan, Z., (2009), "Towards Semantic Trajectory Data Analysis : A Conceptual and Computational Approach". VLDB'09, Lyon, France.

Yuan, J., Zheng, Y., Zhang, L., Xie, X., Sun, G., (2011), "Where to find my next passenger", Proceedings of the 13th international conference on Ubiquitous computing.

Zheng, Y., Liu, Y., Yuan, J., Xie, X., (2011), "Urban Computing with Taxicabs", Proceedings of the 13th International Conference on Ubiquitous Computing, pages 89-98.