Analysis of Data from a Taxi Cab Participatory Sensor Network

Raghu Ganti, Iqbal Mohomed, Ramya Raghavendra, and Anand Ranganathan

IBM T. J. Watson Research Center,
19 Skyline Dr, Hawthorne, NY, USA

Abstract. Mobile *participatory* sensing applications are becoming quite popular, where individuals with mobile sensing devices such as smartphones, music players, and in-car GPS devices collect sensor data and share it with an external entity to compute statistics of mutual interest or map common phenomena. In this paper, we present an analysis of the data from a real-world city-scale mobile participatory sensor network comprised of about two thousand taxi cabs. Our analysis spans data collected from the taxi cab sensor network over the course of a year and we use it to make inferences about life in the city. The large scale data collection (size and time) from these taxi cabs allows us to examine various aspects about life in a city such as busy “party” times in the city, peak taxi usage (space and time), most traveled streets, and travel patterns on holidays. We also provide a summary of lessons learned from our analysis that can aid similar city-scale deployments and their analyses in the future.

1 Introduction

Participatory sensing applications [5,6,8,12,14,18,22] have become quite popular recently. Participatory sensing applications [3] rely on voluntary data collection and sharing within a community towards a common goal such as measurement or mapping of a common phenomenon. A large number of these applications [6,8,12,22] rely on data collected from mobile devices (e.g. in-car sensing devices, mobile GPS devices). For example, CarTel [12] collects location information from GPS sensors in cars to infer traffic conditions. GreenGPS [8] computes fuel-optimal routes from OBD-II sensors installed in cars. Such mobile sensor data sets are a rich source of information, that can be used to measure various global phenomena. For example, location and speed information from moving cars can be used to infer traffic conditions, air pollution levels at large scales can be collectively measured using individual measurements from cars, and potholes on roads can be detected using moving vehicles.

However, as we move from academic studies into real-world deployment and usage, the participatory system becomes more complex as it provides an end-to-end service. In such systems, not only does data get collected in near real-time, it must also be processed by applications expediently. Furthermore, real-world systems will not always be perfectly scalable, either due to architechtural limitations

A. Puiatti et al. (Eds.): MobiQuitous 2011, LNICST 104, pp. 197–208, 2012.

or lack of resources for overprovisoning. Thus, the vagaries of the real-world will affect the absolute number and mobility of nodes in the sensing network over time, and this leads to variation in the operational characteristics of the processing network.

In this paper, we present an analysis of data from a large scale participatory sensing deployment of GPS devices on a fleet of about two thousand taxi cabs in the city of Stockholm, Sweden. These cabs collect and relay periodic location information over the course of a year. We examined about 100 million data points (each point is a location sample from a particular cab) over the course of a year. In this paper, we take the first step of analyzing a large participatory sensing dataset to understand the implications on system deployment, urban management and city dwellers. In particular, we use this testbed to analyze taxi cab availability patterns and draw interesting life-in-a-city inferences from this rich data set.

We also consider the matter of spatio-temporal coverage of taxi cabs - a key issue if additional environmental sensors were installed in these vehicles. In this paper, we show that taxi cabs provide an excellent platform for mobile sensor data collection. They are wide spread, especially in large cities, and are constantly available (there are taxi cabs at any given time of the day, which we will show later). They can be used to determine patterns in the daily lives of people, for example we will show later that Thursday-Saturday is a peak taxi cab usage period, from which we can infer that there is a significant increase in traffic in this city on weekends.

The rest of the paper is divided as follows, we will begin by describing the system architecture and the implementation details in Section 2. We will then provide a detailed analysis of the data from our deployment in Section 3. We discuss the lessons learned and future research directions pertaining to our large scale deployment in Section 4. Related work is presented in Section 5 and the conclusions are summarized in Section 6.

2 System Description

The deployment described in this paper was carried out over time by a large number of individuals as part of a pilot project by the city government of Stockholm. One of the key goals of the deployment was to gain the capability to be able to do real-time analysis of traffic conditions in the city.

We now provide a description of the entire system for data collection, processing, and analyzing the sensor data collected from the taxi based participatory sensor network. This system comprises of three main components, (i) mobile GPS devices equipped with cellular network connectivity installed in taxis, (ii) centralized data collection server, and (iii) IBM InfoSphere Streams [13] for data analysis. Our architecture is a layered one, depicted in Figure 1, which illustrates the various components of our system.

Taxi cabs (participating in our deployment) are installed with GPS devices that can upload acquired samples to the backend via a cellular data network. Due

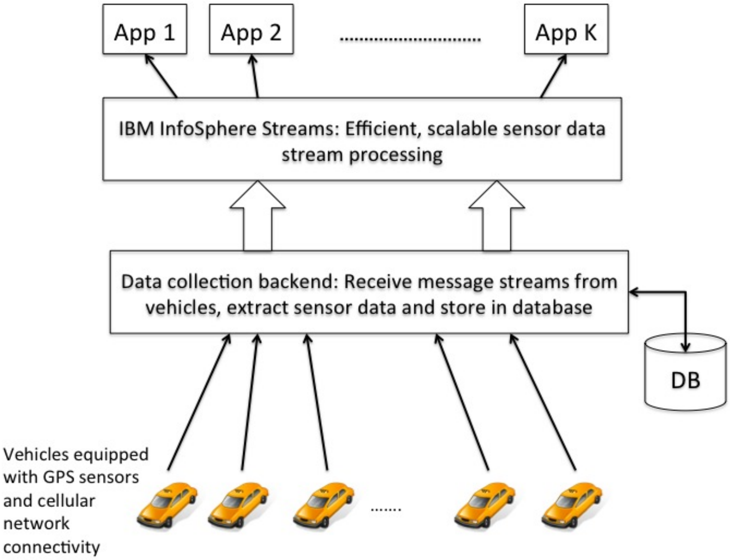

App 1
App 2
App K
IBM InfoSphere Streams: Efficient, scalable sensor data stream processing
Data collection backend: Receive message streams from vehicles, extract sensor data and store in database
DB
Vehicles equipped with GPS sensors and cellular network connectivity

Fig. 1. Architectural components of data collection and processing

to privacy constraints, these devices acquire GPS samples only when a passenger is not present in the taxi cab. Although, in our deployment, specialized GPS devices were used for data collection, we envision that future deployments will simply use smartphones such as the Apple iPhone, Google Nexus S, which are equipped with GPS sensors as well as cellular data networks.

The backend is composed of a data broker and a post processing module. The data broker receives data streams from mobile GPS devices installed in taxis and sends the received messages to the post processing module, which extracts the sensor data from the received stream and then stores it in a database.

The data analysis layer, IBM InfoSphere Streams, interacts with the database to analyze the aggregate sensor data to compute aggregate statistics or phenomena of common interest. Application developers use InfoSphere Streams and the corresponding tools to develop applications. In this paper, we do not use the online fast stream processing capabilities of Infosphere streams (in our previous work [2], we show the scalable and fast processing capabilities of Infosphere streams and use it for implementing a real-time traffic analysis application). We use Infosphere streams to analyze the data collected in an offline manner.

3 Analysis

This section describes and analyzes the sensor data collected from the experimental deployment. The goal of this analysis is to provide insights along two different dimensions: 1) utilize the data collected to infer societal scale phenomena and 2) utilize the data collected to showcase a taxi finding application. The sheer scale of this system enables us to draw meaningful conclusions about data collection and processing as well as societal phenomena.

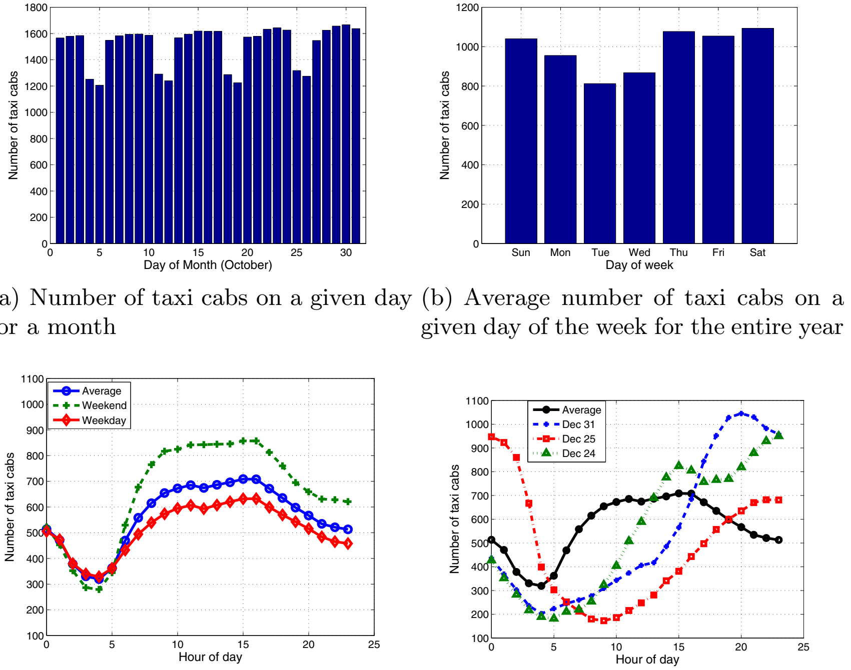
Before we dive into the detailed analysis, we will provide the reader with some details about the data we analyzed. The GPS data traces obtained are for about two thousand taxis for the entire year of 2008. In total, there were about 100 million GPS probe points for the whole year. We plot the number of vehicles collecting sensor data on a given day for a month in Figure 2(a). We observe from Figure 2(a) that the total number of taxicabs collecting sensor data on a given day varies between about 1400 and 1800. Further, we also observed a periodic pattern in the number of vehicles (not shown here) corresponding to the day of the week. A similar observation was made for other months of the year.

3.1 Societal Phenomena: Life in a City

Large scale participatory sensing systems are a platform for providing insights into interesting societal phenomena. In this section, we provide such insights into aggregate phenomena at a societal scale for this particular metropolitan city (with a note that these phenomena will differ based on the city and could change for this city over time). We plot the average number of cabs on a given day of the week for *all* the data of the entire year in Figure 2(b). We observe from Figure 2(b) that the cab traffic is the highest closer to the weekend (Thursday to Saturday). This suggests that there is an increased cab usage (e.g. people traveling to train stations or partying and getting back home) starting Thursday and all the way to Saturday. Further, we also observe that the traffic starts to decrease starting Sunday, with the minimum being on Tuesdays.

In order to understand the "times" when people are out on the streets in this large metropolitan city, we plot the average number of taxi cabs at a given hour of the day for the entire year in Figure 2(c). We also compare the historical average with that of the weekday and weekend averages. We observe from Figure 2(c) that the number of taxi cabs are least late in the night and early in the morning (00:00 to 05:00 hours) and peak at about 10:00 hours and remain more or less constant until 16:00 hours), this pattern is consistent on both weekdays and weekends. We note that the number of cabs is significantly higher on weekends compared to weekdays by a factor of about 1.5. Another observation is that the number of cabs on weekends is slightly lower during the late night/early morning hours (00:00 to 05:00 hours) when compared to weekdays.

Another aspect of the life in a city that we would like to highlight are the holidays and the traffic on these days. We plot the average number of cabs for a given hour of the day across our entire data set and compare it with the average number of cabs on three major holidays in this city, December 24^{th}, December 25^{th}, and December 31^{st}. We make some interesting observations during the holiday times in this city. First, we observe from Figure 2(d) that on December 31^{st}, the number of cabs are lesser than the average until 16:00 hours and it then increases over the next eight hours by a factor of 1.5. A similar observation is made for Christmas eve. On the other hand, the Christmas day itself has a



(a) Number of taxi cabs on a given day for a month

(b) Average number of taxi cabs on a given day of the week for the entire year

(c) Number of taxi cabs at a given hour of the day

(d) Average number of taxi cabs on three major holidays at a given hour of the day

Fig. 2. Perspectives on taxi cab availability

significant dip starting at 05:00 hours and the number of cabs are lesser by about factor of 1/3. We conclude that a participatory sensing system consisting of taxi cabs is likely to have uneven coverage. Such historical aggregate data can be used by taxi dispatching companies to efficiently manage their fleet by estimating the number of vehicles to dispatch and the time of dispatch as well. Further, they can be used to determine *hotspots*, where an individual can hail a taxi cab with a high probability (we will illustrate this aspect in the next section).

In what follows, we will provide some insights into the cab drivers lives. The GPS location samples acquired from the taxi cabs are map matched using a simple algorithm. The map matching algorithm does the following, it finds a few candidate nearest links (on the map) to a point, then for each of the points, it finds the shortest path from the previous point to the current point and verifies if the speed of travel from the previous point to the current point is within a certain threshold. We are aware of more sophisticated map matching algorithms [21,16], which we plan on incorporating in future analyses of the data from our deployments. We plot the top one hundred frequently traveled roads (based on the map matched data) for a month across all the cabs in Figure 3 within the downtown area of the metropolitan city. We observe from Figure 3 that the majority of roads used by the taxi cabs correspond to either highways or major arterial roads (which we verified by comparing against the actual map of the city). We conclude that taxi cab drivers adapt their behavior based on the customer demand they expect.

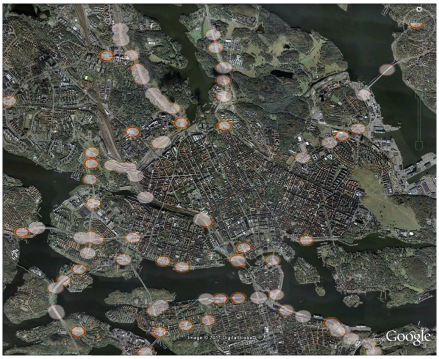
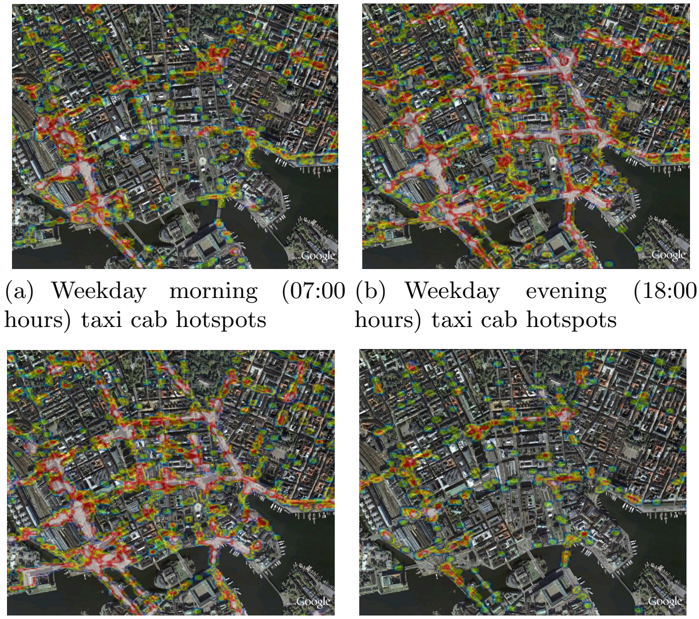


Fig. 3. Most traveled roads by taxi cabs in the downtown area of the city

3.2 Taxi Distribution in Geographical Space and Time

Many taxi cab customers have long wait times, especially at peak times and high passenger traffic areas (e.g. train stations, airports). For example, a taxi cab wait time survey [20] shows that the wait times for cabs at a major junction can be as high as 60 minutes. A useful application to taxi cab customers would be one that provides a spatio-temporal heatmap of available taxis, such that they can determine the closest junction to which they can walk towards. We built one such application using InfoSphere Streams and illustrate a few results. Figure 4 plots the taxi cab availability at different times of the day and the day of the week near a major train station in the big city. The availability is color coded with green being the least number of cabs and white being the most number of cabs (red indicates a fairly reasonable number of cabs). We observe from Figure 4 that the train station is a hotspot for taxi cabs at any time of the day or day of the week.

We plot a similar map for a predominantly residential neighborhood near the downtown. The hotspot pattern in this neighborhood is quite different from the one near the train station. For example, we observe that the weekday evenings and nights are better times to get cabs in residential neighborhoods as opposed to weekday mornings or weekend mornings (at around 07:00 hours). Real-time hotspot availabilities can be provided by this application (when receiving real-time GPS streams) and thus enable customers to find taxi cabs efficiently (in space and time).



(a) Weekday morning (07:00 hours) taxi cab hotspots

(b) Weekday evening (18:00 hours) taxi cab hotspots

(c) Weekday late night (23:00 hours) taxi cab hotspots

(d) Weekend morning (07:00 hours) taxi cab hotspots

Fig. 4. Taxi hotspots near the main train station for one day

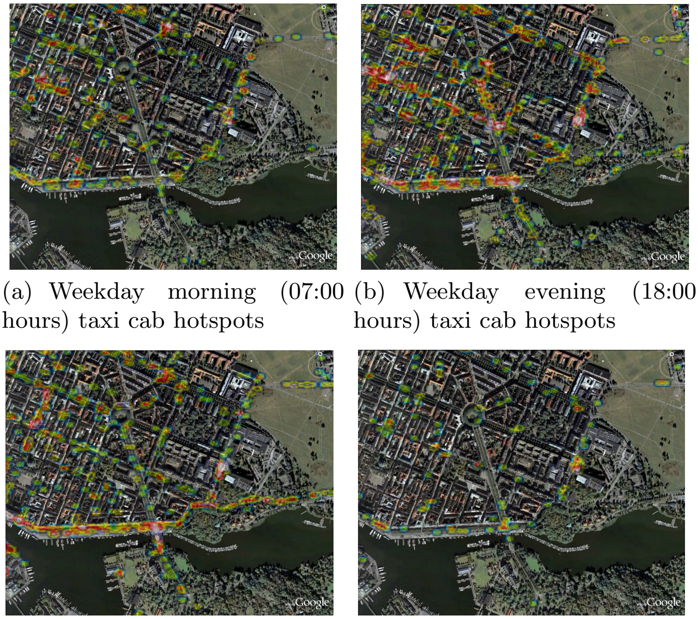
4 Lessons Learned

In the course of our analysis of the data from the deployment, we gleaned insights into several potential challenges that would be relevant to similar deployments in the future. We are currently exploring solutions to these problems and provide a research roadmap for the deployment (data collection, processing, and analysis) of large scale participatory sensing applications.

4.1 Taxi Cabs as a Deployment Platform

While the taxi cabs in our deployment only consist of GPS location sensors, additional sensors can be placed in the taxi cabs. For instance, previous research projects have demonstrated the benefits of placing accelerometers in vehicles in order to detect pot holes on roads [12]. We also postulate that air quality sensors placed in vehicles would provide important data for researchers. The benefit of placing such environmental sensors on taxi cabs is that the constantly moving taxi cabs would provide wide geographical coverage.

Our analysis in Section 3 suggest that there is significant unevenness in the spatial distribution of taxi cabs. We observed persistent "coverage" at major transit hubs (train stations, airports, etc.) and highways, while some neighborhoods (upscale residential areas, for instance) exhibit significant variability over



(a) Weekday morning (07:00 hours) taxi cab hotspots

(b) Weekday evening (18:00 hours) taxi cab hotspots

(c) Weekday late night (23:00 hours) taxi cab hotspots

(d) Weekend morning (07:00 hours) taxi cab hotspots

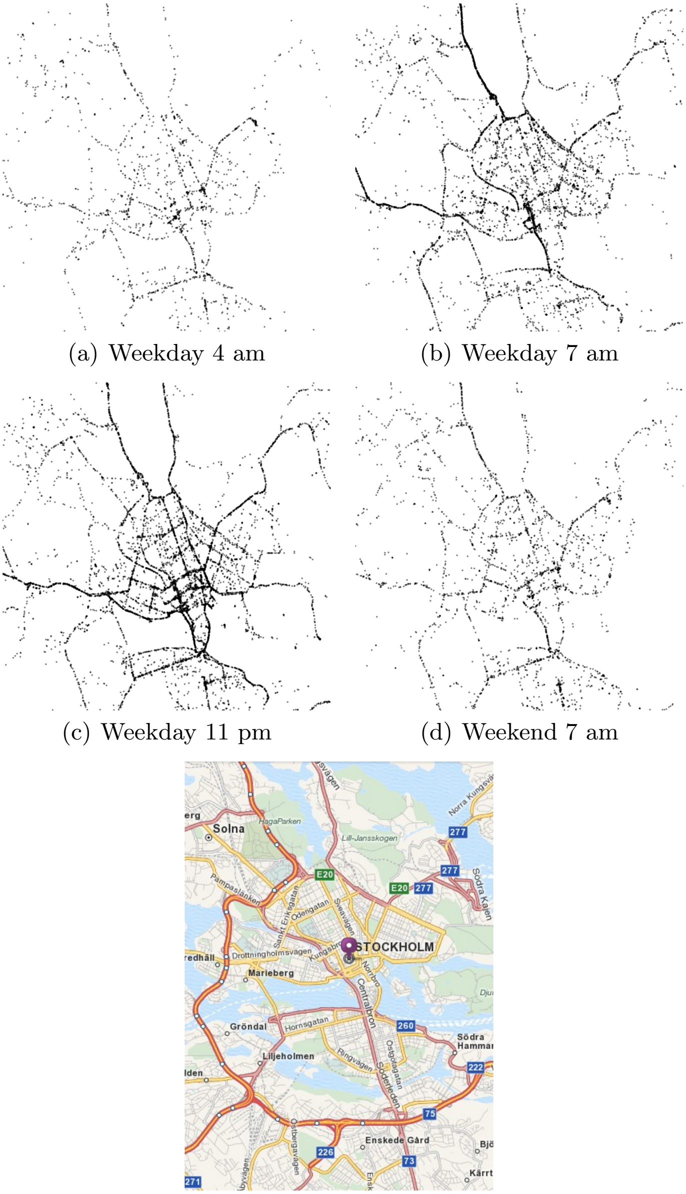
Fig. 5. Taxi hotspots in a predominantly residential neighborhood for one day

time. The basis for this variability is that taxi cabs go to where they expect their customers to be.

A trivial application we considered was whether the data received from cabs in a given hour of one day could be used to plot the major arteries of the city? Figure 6 shows the resulting image for 4 hours: three hours on a weekday and one hour on a weekend. We had assumed that if we considered normal hours when taxicabs would be used, we would be able to see a clear picture. We can see this in panels (b) and (c) of the figure. We also plot the road network map of Stockholm in Figure 6(e), which acts as a baseline for comparison.

However, our intuition was that if we looked at very early hours (4am on a weekday and 7am on a weekend - shown in (a) and (d) panels of the figure), we would not be able to make out much. We were surprised to see that this was not the case. First off, even early in the day, there are a significant number of cabs in operation. Second, these cabs are moving (at car speeds, no less) during this hour, which means that the number of geographical points we collect as well as their spatial coverage will be significant. This is a somewhat special feature of building a sensing platform from taxicabs - a sensing system built on top of private cars or pedestrians may not share this feature.

An application that requires coverage in spatial regions where there is none, could attain this by directing an empty taxi cab to go to a particular address. Such an arrangement would be agreeable to taxi cab drivers as they would still get paid. The benefit for the application creators is that they would have to pay drivers only when there happens to be no opportunistic coverage of a neighborhood. This raises a number of interesting questions, such as what address to pick



(a) Weekday 4 am

(b) Weekday 7 am

(c) Weekday 11 pm

(d) Weekend 7 am

(e) Stockholm road map

Fig. 6. Taxi presence for one day

to minimize cost for the application creator while maximizing coverage, etc. We plan to explore this matter in future work.

4.2 Server Provisioning

In recent years, the idea of elastic computing (dynamically scaling the number of processing nodes based on demand) has received considerable positive attention. As the sensing applications built on top of the taxicab network process data as soon as it is received, the computational resources required by these applications directly depends on the amount of sensor data flowing into the system. Of course, this is a function of the number of taxicabs in operation at that time. We also observed that the average number of taxicabs varies across the day. This suggests an opportunity for savings if we reduce machines during early morning hours. Additionally, our data tells us that there are far more cabs in the city on weekends (Figure 2(c)). As such we may want to have more machines on these days.

An interesting observation we make from our data was that special days, such as holidays, will drastically change the number of cabs being driven around. For instance, Figure 2(d) shows a peak around new years eve and very significant drops at other times. Clearly, occasions such as holidays and special events (e.g. concerts) must be taken into account if we start scaling the number of machines based on history. Another observation is that a steep ebb in the number of taxicabs occurs just before the peaks. We see this pattern on each of the three special days we considered. This is somewhat intuitive because cab drivers may decide to shift their working hours based on the occasion in order to improve their incomes.

5 Related Work

In this section, we describe related work. CarTel [4,7,12] is a participatory sensing deployment where twenty seven cars were equipped with in-vehicle sensors (e.g. GPS, OBD-II) and collected traces for about 290 hours. The CarTel project developed a system to collect continuously streaming sensor data from in-vehicle sensors using WiFi hotspots. They developed traffic prediction, pothole detection, and automotive diagnostics applications using the data collected. Further, algorithms for map matching were also developed [21]. The movement of about fifteen thousand taxis over a 21 month period in Singapore was analyzed in [1]. The paper focused on obtaining traffic information from the deployment. In [23], taxi cab mobility data is used to detect flawed “urban planning” by identifying traffic bottlenecks. In contrast to the above deployments, our paper describes a generic system for data collection from a mobile participatory sensing system and provides analysis of societal phenomena and discusses various lessons learned.

The Berkeley mobile millenium project [11] deployed GPS enabled cell phones on one hundred cars and collected location data on a ten mile stretch near Union city, California for eight hours. The collected data was used for extensive traffic analysis including traffic prediction. Other traffic analysis applications such as [9,17] focus on privacy issues when sharing location data within a community. Nericell [15] uses mobile phones for monitoring traffic flows in cities of the developing regions, where the traffic tends to be more chaotic and there are a large number of vehicles including 2-wheelers, 3-wheelers, cars, and buses. The StarTrack [10] system provides a scalable middleware and API for building continuous tracking application. The location data that is collected from users can be leveraged by multiple applications via a high-level API that aggregates individual location samples into “tracks”. A cooperative transit tracking application using GPS enabled smartphones where individuals share GPS samples from their daily bus rides to track in-transit buses was developed in [22]. GreenGPS, a participatory sensing application that predicts fuel efficient routes is presented in [8]. In the paper, the authors deploy GPS and OBD-II sensors on sixteen personal cars and collect fuel consumption sensor data tagged with location data. These data are then used to construct fuel consumption models to predict the fuel efficient route. Our goal in this paper is to analyze data from a taxi cab participatory sensor network and provide insights into lessons learned and societal phenomena.

We are also aware of an iPhone application called *Cab Sense* [19] that provides the user with the best corner to catch a taxi in New York City. As far as we are aware, we do not know of any related publication. In this paper, we analyze the data collected and present various insights and claim no novelty in regard to the application itself.

6 Conclusions

In this paper, we analyzed data from a real-word deployment of a system that continuously collects and processes GPS data from taxicabs. We made several observations pertaining to the system’s operational characteristics. For instance, we found that processing delays in this system are a function of the number of taxi cabs that are actively collecting data at that time. We also made use of the data to examine societal phenomenon and taxi distributions across geographical space and time. Finally, we presented operational insights gleaned from our analyses and suggested avenues for future research.

References

1. Balan, R.K., Nguyen, K.X., Jiang, L.: Real-time trip information service for a large taxi fleet. In: Proc. of ACM MobiSys, pp. 99–112 (2011)
2. Biem, A., et al.: Ibm infosphere streams for scalable, real-time, intelligent transportation services. In: Proc. of ACM SIGMOD, pp. 1093–1104 (2010)
3. Burke, J., et al.: Participatory sensing. Workshop on World-Sensor-Web, co-located with ACM SenSys (2006)
4. Bychkovsky, V., et al.: A measurement study of vehicular internet access using in situ wi-fi networks. In: Proc. of ACM MobiCom, pp. 50–61 (2006)
5. Davis, M., et al.: Mmm2: Mobile media metadata for media sharing. In: CHI Extended Abstracts on Human Factors in Computing Systems, pp. 1335–1338 (2005)

6. Eisenman, S.B., et al.: The bikenet mobile sensing system for cyclist experience mapping. In: Proc. of SenSys (November 2007)
7. Eriksson, J., et al.: The pothole patrol: Using a mobile sensor network for road surface monitoring. In: Proc. of ACM MobiSys, pp. 29–39 (2008)
8. Ganti, R.K., et al.: GreenGPS: A participatory sensing fuel-efficient maps application. In: Proc. of ACM MobiSys, pp. 151–164 (2010)
9. Ganti, R.K., Pham, N., Tsai, Y.-E., Abdelzaher, T.F.: Poolview: Stream privacy for grassroots participatory sensing. In: Proc. of SenSys 2008, pp. 281–294 (2008)
10. Haridasan, M., Mohomed, I., Terry, D., Thekkath, C.A., Zhang, L.: Startrack next generation: A scalable infrastructure for track-based applications. In: Proc. of OSDI, pp. 409–422 (2010)
11. Herrera, J.C., et al.: Evaluation of traffic data obtained via gps-enabled mobile phones. Transport Research, Part C 18(4), 568–583 (2009)
12. Hull, B., et al.: Cartel: a distributed mobile sensor computing system. In: Proc. of SenSys, pp. 125–138 (2006)
13. IBM. Infosphere streams, `http://www.ibm.com/software/data/infosphere/streams/`
14. Lu, H., et al.: Soundsense: Scalable sound sensing for people-centric applications on mobile phones. In: Proc. of ACM MobiSys, pp. 165-178 (2009)
15. Mohan, P., Padmanabhan, V.N., Ramjee, R.: Nericell: Rich monitoring of road and traffic conditions using mobile smartphones. In: Proc. of ACM SenSys, pp. 323–336 (2008)
16. Newson, P., Krumm, J.: Hidden markov map matching through noise and sparseness. In: ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 336–343 (2009)
17. Pham, N., Ganti, R.K., Uddin, Y.S., Nath, S., Abdelzaher, T.: Privacy-Preserving Reconstruction of Multidimensional Data Maps in Vehicular Participatory Sensing. In: Silva, J.S., Krishnamachari, B., Boavida, F. (eds.) EWSN 2010. LNCS, vol. 5970, pp. 114–130. Springer, Heidelberg (2010)
18. Reddy, S., et al.: Image browsing, processing, and clustering for participatory sensing: Lessons from a dietsense prototype. In: Proc of EmNets, pp. 13-17 (2007)
19. Sense Networks. Cab sense, `http://www.cabsense.com/`
20. Singapore Government. Average hourly passenger wait time for taxi cabs, `http://www.lta.gov.sg/public_transport/doc/Website-Feb11.pdf`
21. Thiagarajan, A., et al.: Vtrack: Accurate, energy-aware traffic delay estimation using mobile phones. In: Proc. of ACM SenSys, pp. 85–98 (2009)
22. Thiagarajan, A., et al.: Cooperative transit tracking using smart-phones. In: Proc. of ACM SenSys, pp. 85–98 (2010)
23. Zheng, Y., Liu, Y., Yuan, J., Xie, X.: Urban computing with taxicabs. In: Proc. of ACM UbiComp (2011)