Research on Big Data for Mobile Computing

Ramakanth. Komati

PG Scholar, Assistant Professor, Christu Jyothi Institute of Technology and Science, India.

Abstract- This paper presents an overview of the Mobile Data Challenge (MDC), a large-scale research initiative aimed at generating innovations around smart phone-based research, as well as community-based evaluation of related mobile data analysis methodologies. First we review the Lausanne Data Collection Campaign (LDCC) {an initiative to collect unique, longitudinal smart phone data set for the basis of the MDC. Then, we introduce the Open and Dedicated Tracks of the MDC; describe the specific data sets used in each of them; and discuss some of the key aspects in order to generate privacy-respecting, challenging, and scientifically relevant mobile data resources for wider use of the research community. The concluding remarks will summarize the paper.

I. Introduction

Mobile phone technology has transformed the way we live, as phone adoption has increased rapidly across the globe [17]. This has widespread social implications. The phones themselves have become instruments for fast communication and collective participation. Further, different user groups, like teenagers, have started to use them in creative ways. At the same time, the number of sensors embedded in phones and the applications built around them have exploded. In the past few years smart phones remarkably started to carry sensors like GPS, accelerometer, gyroscope, microphone, camera and Bluetooth. Related application and service offering covers e.g. information search, entertainment or healthcare.

The ubiquity of mobile phones and the increasing wealth of the data generated from sensors and applications are giving rise to a new research domain across computing and social science. Researchers are beginning to examine issues in behavioral and social science from the Big Data perspective {by using large-scale mobile data as input to characterize and understand real-life phenomena, including individual traits, as well as human mobility, communication, and interaction patterns [11, 12, 9].

This new research, whose findings are clearly important to society at large, has been often conducted within corporations that historically have had access to these data types, including telecom operators [13] or Internet companies [6],or through granted data access to academics in highly restricted forms [12]. Some initiatives, like [1], have collected publicly available but in some extent limited data sets together. Clearly, government and corporate regulations for privacy and data protection play a fundamental and necessary role in protecting all sensitive aspects of mobile data. From the research perspective, this also implies that mobile data resources are scarce and often not ecologically valid to test scientific hypotheses related to real-life behavior.

The Mobile Data Challenge (MDC) by Nokia is motivated by our belief in the value of mobile computing research for the common good - i.e., of research that can result in a deeper scientific understanding of human and social phenomena, advanced mobile experiences and technological innovations. Guided by this principle, in January 2009 Nokia Research Center Lausanne and its Swiss academic partners Idiap and EPFL started an initiative to create large-scale mobile data research resources. This included the design and implementation of the Lausanne Data Collection Campaign (LDCC), an effort to collect a longitudinal smart phone data set from nearly 200 volunteers in the Lake Geneva region over one year of time. It also involved the definition of a number of research tasks with clearly specified experimental protocols. From the very beginning the intention was to share these research resources with the research community which required integration of holistic and proactive approach on privacy according to the of privacy-by-design principles [2].

The MDC is the visible outcome of nearly three years of work in this direction. The Challenge provided researchers with an opportunity to analyze a relatively unexplored data set including rich mobility, communication, and interaction information. The MDC comprised of two research alternatives through an Open Research Track and a Dedicated Research Track. In the Open Track, researchers were given opportunity to approach the data set from an exploratory perspective, by proposing their own tasks according to their interests and background. The Dedicated Track gave researchers the possibility to take on up to three tasks to solve, related with prediction of mobility patterns, recognition of place categories, and estimation of demographic attributes. Each of these tasks had properly defined experimental protocols and standard evaluation measures to assess and rank all contributions.

This paper presents an overview of the Mobile Data Challenge intended both for participants of the MDC and a wider audience. Section 2 summarizes the LDCC data, the basis for the MDC. Section 3 describes the MDC tracks and tasks in detail. Section 4 provides details on the specific data sets used for the MDC. Section 5 summarizes the schedule we have followed to organize the Challenge. Finally, Section 6 offers some final remarks.

2. The Lausanne Data Collection Campaign (LDCC)

LDCC aimed at designing and implementing a large-scale campaign to collect smartphone data in everyday life conditions, grounding the study on a European culture. The overall goal was to collect quasi-continuous measurements covering all sensory and other available information on a smartphone. This way we were able to capture phone users' daily activities unobtrusively, in a setting that implemented the privacy-by design principles [2]. The collected data included a significant amount of behavioral information, including both personal and relational aspects. This enables investigation of a large number of research questions related to personal and social context - including mobility, phone usage, communication, and interaction. Only content, like image _les or content of the messages, was excluded because content capturing was considered too intrusive for the longitudinal study based on volunteering participation with selfless drivers. Instead log-_les with metadata were collected both for imaging and messaging applications. This section provides a summary on the LDCC implementation and captured data types. An initial paper introducing LDCC, its data types and statistics early 2010 appeared in [14]. Part of the material in this section has been adapted from it.

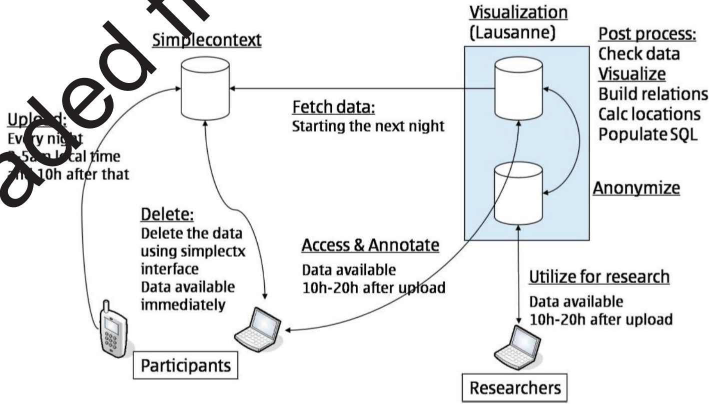


Figure 1: LDCC data flow, progressing from mobile data from volunteers to anonymized data for research [14]).

2.1 LDCC design

Nokia Research Center, Idiap, and EPFL partnered towards LDCC since January 2009. After the implementation and evaluation of the sensing architecture, and the recruitment of the initial pool of volunteers, the data collection started in October 2009. Over time, smartphones with data collection software were allocated to close to 200 volunteers in the Lake Geneva region. A viral approach was used to pro mote the campaign and recruit volunteers. This resulted in a great proportion of the members of the campaign population having social connections to other participants, as well as to the demographical representativeness. A key aspect of the success of LDCC was the enthusiastic participation of volunteers who agreed to participate and share their data mainly driven by selfless interest. The campaign concluded in March 2011.

Data was collected using Nokia N95 phones and a client-server architecture that made the data collection invisible to the participants. A seamless implementation of the data recording process was a key to make a longitudinal study feasible in practice { many participants remained in the study for over a year. Another important target for the client software design was to reach an appropriate trade-off between quality of the collected data and phone energy consumption. The collected data was first stored in the device and then uploaded automatically to a Simple Context server via WLAN. The server received the data, and built a database that could be accessed by the campaign participants. The Nokia Simple Context backend had been developed already earlier by the Nokia Research Center in Palo Alto. Additionally data visualization tool was developed which offered a “life diary" type of view for the campaign participants on their data. Simultaneously, an anonymized database was being populated, from which researchers were able to access the data for their purposes. Fig. 1 presents a block diagram of the collection architecture.

2.2 Data characteristics

The LDCC initiative produced a unique data set in terms of scale, temporal dimension, and variety of data types. The campaign population reached 185 participants (38% female, 62% male), and was concentrated on young individuals (the age range of 22-33 year-old accounts for roughly two thirds of the population.) A bird-eye's view on the LDCC in terms of data types appears in Table 1. As can be seen, data types related to location (GPS, WLAN), motion (accelerometer), proximity (Bluetooth), communication (phone call and SMS logs), multimedia (camera, media player), and application usage (user-downloaded applications in addition to system ones) and audio environment (optional) were recorded. The numbers themselves reflect a combination of experimental design choices (e.g., every user had the same phone and data plan) and specific aspects of the volunteer population (e.g., many participants use public transportation).

Data type	Quantity
Calls (in/out/missed)	240,227
SMS (in/out/failed/pending)	175,832
Photos	37,151
Videos Application events	2,940
Calendar entries	8,096,870
Phone book entries	13,792
Location points	45,928
Unique cell towers	26,152,673
Accelerometer samples	99,166
Bluetooth observations	1,273,333
Unique Bluetooth devices	38,259,550
WLAN observations	498,593
Unique WLAN access points	31,013,270
Audio samples	560,441

Table 1: LDCC main data types and amounts for each type.

Due to space limitations, it is not possible to visualize multiple data types here. A compelling example, however, is presented in Fig. 2, which plots the raw location data of the LDCC on the map of Switzerland for the volunteer population after 1 week, and then after 1, 3, 6, 12, and 18 months of campaign. When seen in detail, the geographical coverage of the LDCC allows a reasonable tracing of the main routes on the map of Suisse Romande -French-speaking, western part of Switzerland - and gradually also of other regions of the country.

In addition to contributing phone data, participants of the LDCC also agreed to fill a small number of surveys during the data recording process. We would like to highlight two types of survey data which were important for the later development of the MDC - (1) a set of manual semantic labels for frequently and infrequently visited places for each user and (2) basic demographic attributes. The relevant places were first detected automatically with a method discussed in [15]. After that the campaign participants specified place categories from a fixed list of tags (home, work, leisure places, etc.). In sense of demographics, participants self-reported their attributes like gender, age group, marital status and job type etc.

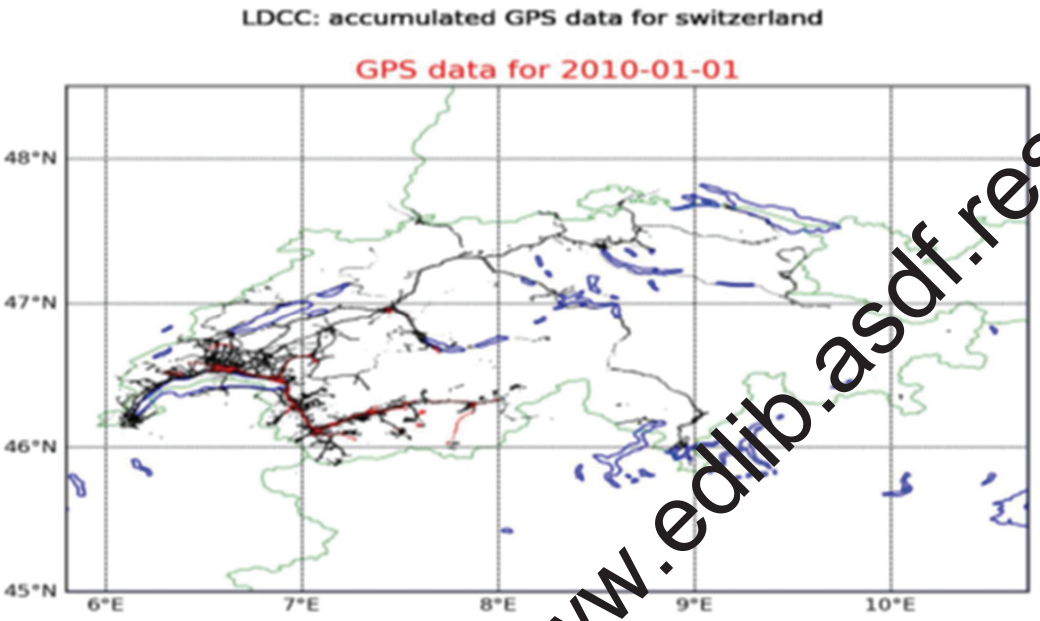
2.3 Privacy

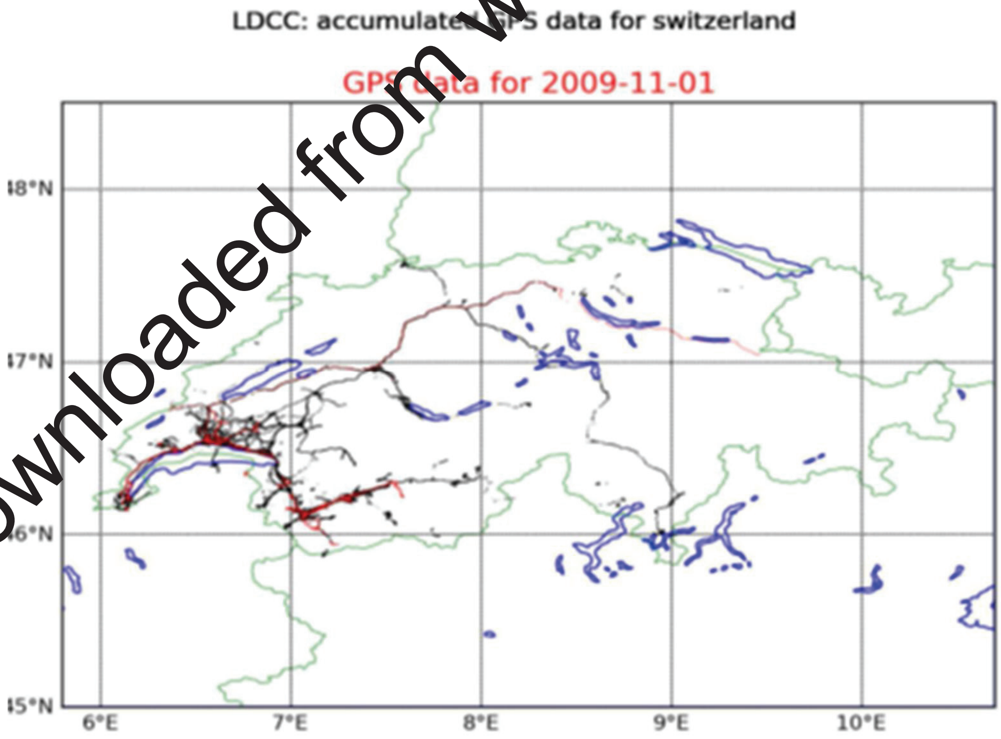
Privacy played an essential role in the design and implementation of the LDCC, given the nature and scale of the data shared by the participants of the initiative. In order to satisfy the ethical and legal requirements to collect data while protecting the privacy of the participants, the LDCC research team implemented an approach based on multiple strict measures. The approach can be summarized as follows (more details can be found in [14]):

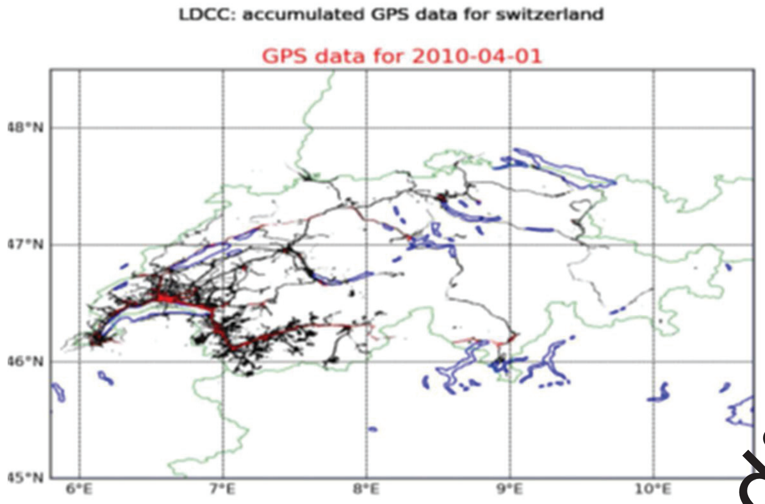
1. Communication with volunteers about privacy. Following Nokia's general privacy policy, we obtained written consent from each individual participating the LDCC. We explicitly stated that data would be collected for research purposes. All participants were informed about their data rights, including the right to access their own collected data and to decide what to do with it (e.g. to delete data entries if they opted to do so). The participants had also opportunity to opt-out at any moment.
2. Data security. The data was recorded and stored using best industry practices in this domain.
3. Data anonymization. By design, the LDCC did not store any content information (e.g. no photo _les or message content were recorded). The major portion of the collected data consisted of event logs, and when sensitive data beyond logs was collected, it was anonymized using state-of-the-art techniques and/or aggregated for research purposes [5]. Examples include the use of pseudonyms instead of identifiable data and the reduction of location accuracy around potentially sensitive locations. The researchers do have access only to the anonymized data.
4. Commitment of researchers to respect privacy. Privacy protection of such a rich data only by automatic anonymization techniques is not possible so that research value and richness of the data can be simultaneously maintained. In addition to technical means also agreement based counter-measures are necessary. Trusted researchers have been able to work with the LDCC data after agreeing in written form to respect the anonymity and privacy of the volunteering LDCC participants. This practically limited the access to the LDCC data to a small number of authorized partners and their affiliated researchers. After our initial experience with the LDCC, the next step was to outreach the mobile computing community at large, which motivated the creation of the Mobile Data Challenge, discussed in detail in the next sections.

3. MDC Tracks

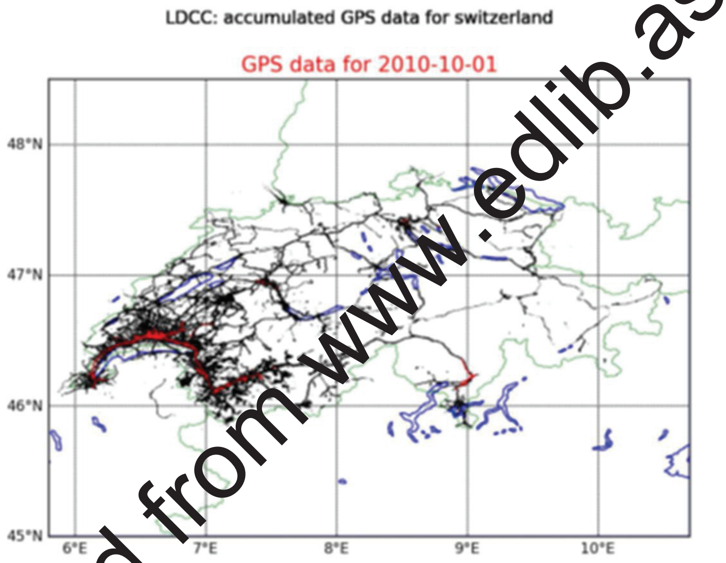
MDC's original intention was to be inclusive at a global scale. Other previously successful evaluation initiatives in computing, like those organized by NIST in several areas [16, 18] or the Netix challenge [8, 7] focused on either one or at most a small number of tasks with objective evaluation protocols. This was also a guiding principle for MDC. On the other hand, the nature of mobile data is highly exploratory, so there was a clear benefit in encouraging and welcoming also open ideas. Learning from these past experiences, we decided that MDC would feature both open and pre-defined options to participate. The Open Track was defined to receive self-defined ideas from the community. On the other hand, the concrete options were given in the Dedicated Track, which defined three classification/prediction tasks. These tasks covered several key aspects of mobility and mobile users. The Open Track. This track enabled participants to propose their own Challenge task based on their own research interests and background.



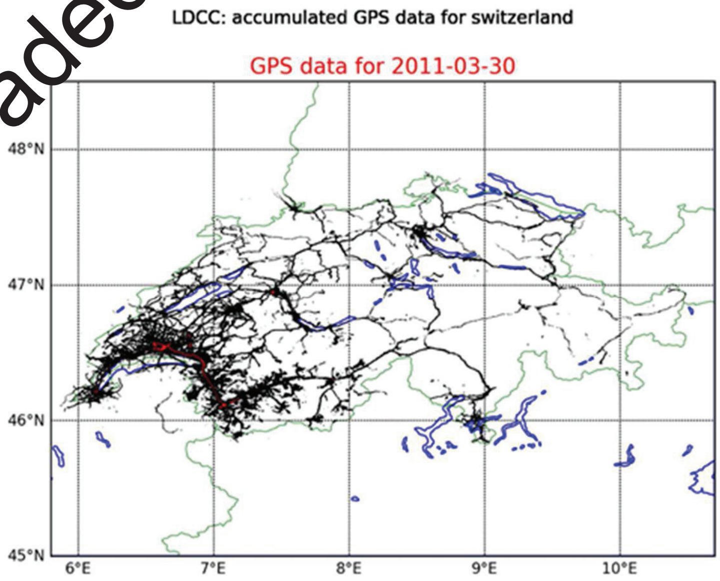



LDCC: accumulated GPS data for switzerland

GPS data for 2010-04-01


LDCC: accumulated GPS data for switzerland

GPS data for 2010-10-01


LDCC: accumulated GPS data for switzerland

GPS data for 2011-03-30

4. MDC Data

This section presents an overview of the MDC datasets and the corresponding preparation procedures. We first describe the division of the original LDCC data that was needed in order to address the different MDC tasks. We then summarize the data types that were made available. We finalize by discussing the procedures related to privacy and data security.

4.1 Division of the Dataset

The datasets provided to the participants of the MDC consist of slices of the full LDCC dataset. Slicing the data was needed in order to create separate training and test sets for the tasks in the Dedicated Track, but was also useful to assign the richest and cleanest parts of the LDCC dataset to the right type of challenge. Four data slices were created for the MDC:

Set A: Common training set for the three dedicated tasks.
Set B: Test set for demographic attribute and semantic place label prediction tasks.
Set C: Test set for location prediction task.

Open set. Set for all open track entries.

The overall structure of the datasets is given in Figure 3. The rationale behind this structure was the following. First, the participants of the LDCC were separated in three groups, according to the quality of their data according to different aspects. The 80 users with the highest-quality location traces were assigned to sets A and C. Set A contains the full data for these users except the 50 last days of traces, whereas set C contains the 50 last days for which location data is available for testing. In order to maximize the use of our available data, we reused Set A as a training set for the two other dedicated tasks. A set of 34 further users was selected as a test set for these tasks and appeared as Set B. In this way, models trained on the users of Set A can be applied to the users of their most visited locations.

4.2 Data Types

For both Open and Dedicated Tracks, most data types were released in a raw format except a few data types that needed to be anonymized. There are two main differences between the Open Track data and the Dedicated Track data. First, the physical location (based on GPS coordinates) was avail- able in the Open Track but not in the Dedicated Track. Instead, we released a preprocessed version of the location data in the form of sequences of visited places for the Dedicated Track. This allowed studying performance of algorithms in location privacy-sensitive manner. The second main dfference was in the availability of relational data between users. This included both direct contacts (e.g., when a user calls another user) and indirect contacts (e.g. if two users observe the same WLAN access point at the same time then they are in proximity). We decided to keep this data in the Open Track but removed it in the Dedicated Track since it could have potentially revealed the ground truth to be predicted.

In the anonymization algorithm, a common encryption password was used for the users selected to the Open Track data sets. On the other hand, we used a different password for each user in the Dedicated Track.

Common data types. Each data type corresponds to a table in which each row represents a record such as a phone call or an observation of a WLAN access point. User IDs and timestamps are the basic information for each record.

Data types for Open Track only. Geo-location information was only available in the Open Track. In addition to GPS data, we also used WLAN data for inferring user location. The location of WLAN access points was computed by matching WLAN traces with GPS traces during the data collection campaign.

5. MDC Schedule

The plans to organize MDC started in summer 2011. We targeted to organize the final MDC workshop within one year. We decided to keep the challenge open for all the researchers with purely academic affiliation. The prospective participants of the Open Track had to submit a short proposal with their concrete plan, and the participants of the Dedicated Track had to agree to participate at least one task. While the MDC was by nature open, a series of important steps were established for participant registration. Importantly, this included signature of the Terms and Conditions agreement.

6. Conclusions

This paper described a systematic flow of research, targeting to create and provide unique longitudinal smartphone data set for wider use by the research community. In this paper we gave motivation for this initiative and summarized the key aspects of the Lausanne Data Collection Campaign (LDCC) in which the rich smartphone data was collected from around 200 individuals over more than a year. We also described in further details the Mobile Data Challenge (MDC) by Nokia which was a data analytics contest making this data widely available to the research community.

7. References

1. http://crawdad.cs.dartmouth.edu/.
2. http://privacybydesign.ca/.
3. http://en.wikipedia.org/wiki/MAC_address.
4. http://research.nokia.com/mdc.
5. I. Aad and V. Niemi. NRC Data Collection and the Privacy by Design Principles. In PhoneSense, 2011.
6. L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In Proc. World Wide Web Conf. (WWW), Apr. 2010.
7. R. Bell, J. Bennett, Y. Koren, and C. Volinsky. The million dollar programming prize. Spectrum, IEEE, 46(5):28{33, 2009.
8. J. Bennett and S. Lanning. The netflix prize. In Proceedings of KDD Cup and Workshop, volume 2007, page 35, 2007.
9. G. Chittaranjan, J. Blom, and D. Gatica-Perez. Mining large-scale smartphone data for personality studies. Personal and Ubiquitous Computing, published online Dec. 2011.
10. T. Do and D. Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In Proc. ACM Int. Conf. on Ubiquitous Computing, Pittsburgh, Sep. 2012.