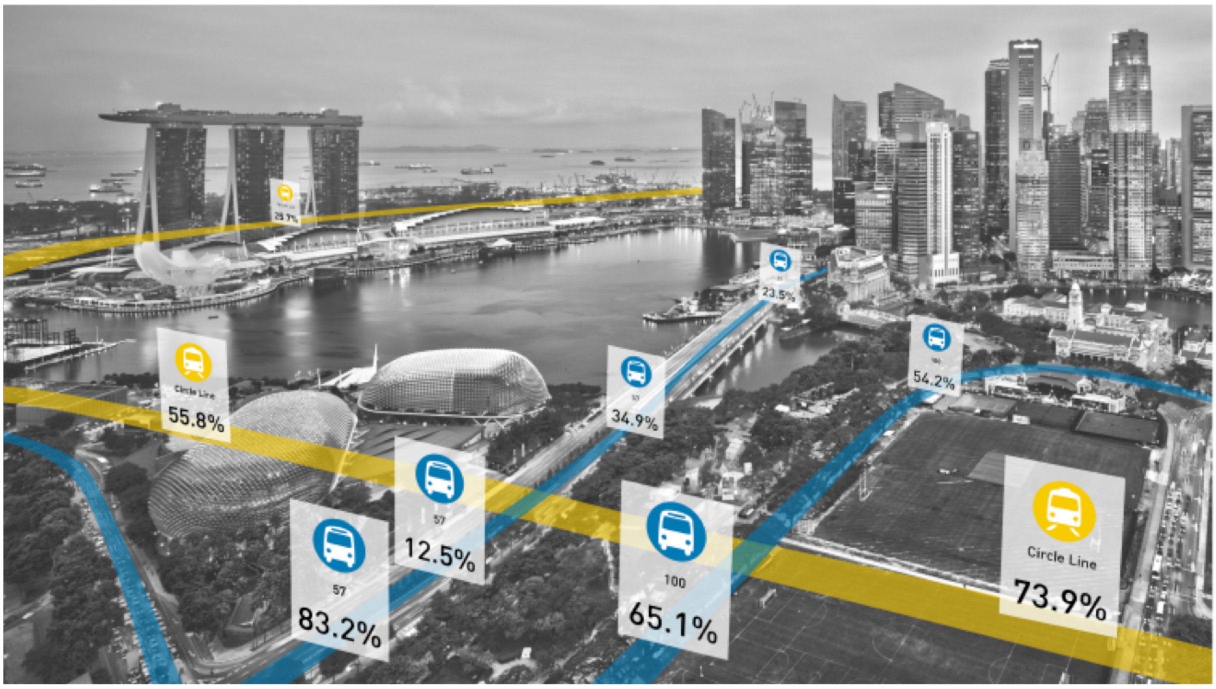

Transport Modelling in the Age of Big Data

Cuauhtémoc Anda

Pieter Fourie

Alexander Erath

Work Report **June 2016**

(FCL) FUTURE
CITIES
LABORATORY

(SEC) SINGAPORE-ETH 新加坡－ETH
CENTRE 研究中心

New Big Data sources such as mobile phone call data records, smart card data and geo-coded social media records allow to observe and understand mobility behaviour on an unprecedented level of detail. Despite the availability of such new Big Data sources, transport demand models used in planning practice still, almost exclusively, are based on conventional data such as travel diary surveys and population census. This literature review brings together recent advances in harnessing Big Data sources to understand travel behaviour and informing travel demand models that allow to compute *what-if* scenarios. From trip identification to activity inference, we review and analyse the existing data-mining methods that enable these opportunistically collected mobility traces inform transport demand models. We identify that future research should tap on the potential of probabilistic models as commonly used in data science. Those data mining approaches are designed to handle the uncertainty of sparse and noisy data as it is the case for mobile phone data derived mobility traces. In addition, data fusion approaches should be applied to integrate disparate but related datasets to blend Big Data with more granular information from travel diaries. In any case, we also acknowledge that sophisticated modelling knowledge has developed in the domain of transport planning and therefore we strongly advise that still domain expert knowledge should build the fundament when applying data driven approaches in transport planning. These new challenges call for a multidisciplinary collaboration between transport modellers and data scientists.

1 Introduction

Before the ubiquitous sensing of human mobility flows became possible through mobile phones, public transport smart card transactions or Global Positioning System (GPS)-enabled devices, it was difficult and expensive to generate large scale or even population-wide samples that allow developing travel demand models. The aim of such models is not only to replicate with relevant accuracy actual mobility flow but also the application of *what-if* scenarios to evaluate the impact of different infrastructure development decisions.

Despite the availability of such new Big Data sources, transport demand models used in planning practice still, almost exclusively, are based on conventional data such as travel diary surveys and population census. While the applied statistical models have become more sophisticated as computation power grew exponentially over the last decades, the most important change being the evolution from trip- to activity-based models, the basic modelling paradigm remained the same: mobility travel diary survey that only cover a small sample of the actual population is being used to synthesise transport flows of a representative population.

New Big Data sources such as mobile phone call data records, smart card data and geo-coded social media records allow us to observe and understand mobility behaviour on an unprecedented level of details. But simply observing is not particularly helpful for planning purposes. To allow for prediction in what-if scenarios, we need to understand and contextualise the information contained in such Big Data sources to inform models of travel behaviour and adapt them to be useful in travel demand modelling frameworks.

This literature review brings together recent advances[1] in the fields of harnessing Big Data sources to understand travel behaviour and inform travel demand models that allow to compute *what-if* scenarios. To this end, we first provide a primer on the latest advances in transport demand modelling including the latest agent-based approaches.

Our focus is then on the research that makes use of the relevant Big Data sources and directly ties in the methodological toolkit of travel demand models, hence excluding methods that primarily derive real-time analytics from such Big Data sources. The aim of the paper is to provide the reader with an overview how Big Data already improve the understanding of mobility flows and has been applied for transport demand models from a methodological angle. From this collection, we identify the advantages and disadvantage of the various methodologies and its applicability for being used in predictive transport models. Conclusions drawn from this literature review include the description of new modelling applications that the new data sources allow but also the specification of research gaps that need to be overcome to realise them.

[1] Mainly from 2010 to the first quarter of 2016

2 Travel demand models and mobility data

2.1 Transport Demand Models

Travel demand models have been developed to support decision making by forecasting the impacts of alternative transportation and land use scenarios. (Castiglione et al., 2015). The following two approaches for transport demand forecasting can be distinguished: first, aggregated models in which travel demand is specified as aggregated transport flow between zones, and second, agent-based models that maintain travel demand at the level of individuals throughout the model.

2.1.1 The classical four step model

’4-step’ demand modelling was introduced in the 1960s (de Dios Ortúzar and Willumsen, 2011). Originally specified as trip-based models they aim at predicting the number of trips for different travel modes and routes taken between any two origin and destination zones. The first trip generation component estimates the number of trips produced by and attracted to each zone. The second trip distribution step connects where trips are produced and where they are attracted to. The third mode choice step determines the travel mode, such as automobile or bus, used for each trip, while the fourth assignment step predicts the routes used for each trip and hence allows to model congestion induced traffic delays. Since such delays can influence the mode and route choice but also location choice behaviour, feedback loops including steps 2, 3 and 4 are usually introduced. Data requirement for four step modelling includes household travel survey information, census information, and a representation of the transportation network.

2.1.2 Activity-based models

Since the early 1990s, activity-based models have been promoted as a superior alternative to four-step models, avoiding some of the inherent limitations of the latter type of models. To appreciate the significance of activity-based models, Rasouli and Timmermans (2013) emphasise that four-step models are aggregate in nature – the unit of measurement is not an individual, but rather the number of trips emanating from any particular zone. In addition, the four-step model lacks consistency and congruence on how behavioural parameters are used across various submodels. Furthermore, the assumption of independency between the four modelling steps are regularly cited as key shortcomings when it comes to the the evaluation of travel demand management policies such as mobility pricing.

The fundamental principle of activity-based models is the understanding that travel is ultimately derived from the necessity to participate in activities. The aim of activity-based models is to predict for each individual the number, sequence and type of the activities are conducted over a certain time-period subject to a set of spatial, temporal and resources constraints. However, while activity-based models allow to generate spatially and temporally disaggregate description of travel demand, for route choice and traffic simulation this travel demand is often aggregated again to so-called origin-destination-matrices that describe how many trips are conducted between any two OD-pairs. This restriction was originally due to the lack of simulation models that are suitable to simulate traffic for a relevant spatial extent, i.e. entire cities or regions and across an entire day, but still apply today due to the computational requirements of agent-based transport simulation.

In addition to four step model data requirements, activity based models do require one additional type of input, a 'synthetic population' at the level of individual households and persons that is representative for the actual population of the area of interest. This 'synthetic population' includes a set of socio-demographic attributes which are then used for travel demand modelling processes. Moreover, for every agent in the synthetic population, a fully descriptive daily activity plan, including locations of daily activities such as work or education needs to be derived.

Agent-based transport models Agent-based transport models for strategic transport planning usually derive travel demand from activity-based modelling approaches but employ microscopic and completely time-dynamic traffic simulation of each agent's individual demand based on system constraints given by the transport network and its attributes. (Balmer, Axhausen and Nagel, 2006).

While the original development of Transims Smith et al. (1995) as the first large-scale agent-based transport simulator clearly focused on replacing aggregated transport assignment methods, later implementations of TRANSIM and more recent developments of agent-based model such as MATSim (Horni et al., 2016), SimMobility (Adnan et al., 2016), SimAGENT (Goulias et al., 2012), integrate to different degrees also mode, time, destination and activity scheduling processes into a single consistent modelling framework. Such an integrated modelling framework overcomes that travel demand can in disaggregated form throughout the whole modelling process. Besides the enhanced behavioural consistency, this also allows the modelling and analysis of modern travel demand management tools such as time- and demand-dependent pricing and new forms of mobility such as shared and autonomous vehicles.

Multi agent-based modelling is built upon a large scale of autonomous agents which perform their own decisions, interact with one another and with the environment. For each agent, an initial daily activity plan is assigned as a precise description of the activities' location, its durations, start and end time, and the trips connecting two activities, including mode and route.

Among the several agent-based transport models that are currently under continued development, MATSim takes a special role and can be considered the the currently most widely applied model. It can integrate a wide range of decision dimensions in a co-evolutionary learning loop, but due to its modular framework, it can also be used for traffic simulation only and integrated with other activity-based travel demand models.

In MATSim, a day is simulated multiple times and after each iteration a fraction of the agents is allowed to modify their plans (i.e. mutation/crossover phase). For instance, they can change their departure time, the travel mode of a sub-tour, location of a given type of activity, among others. At the end of each simulated day, the utility of the day is measured for each agent using a scoring function that rewards agents for performing activities, while penalising them for travelling, transferring between transport modes, waiting at transit stops and arriving late for activities, etc. In such way, agents seek to improve their utility over iterations until the system reaches an equilibrium where the generalised utility can not be longer improved (i.e. a steady-state is reached) (Balmer et al.,2009).

2.2 Big Data describing mobility

Precise geo-referenced location data represents a large and growing subset of Big Data as mobile devices and location-sensing technologies become ubiquitous. For the purpose of transportation planning, and accordingly to disaggregate activity-based approaches, the aim of this literature review paper will be limited only to mobility data sources emerging from individuals. We are interested in the digital trace left by individuals because it can provide us with more accurate and interesting insights on mobility patterns. Hence, out of the scope will remain mobility information generated by infrastructure sensors that document traffic flows at certain cross-sections (e.g. loop detectors, video vehicle detection systems, ERP systems).

Due to its wide-coverage in urban settings, the main focus of the survey review will be firstly on data generated through Smart Card Automatic Fare Collection (SC-AFC) systems, and mobile phone networks. Both of them can be classified as large-scale opportunistic human mobility sensors, which are able to provide insights on urban dynamics and human activities at an unprecedented scale and level of detail. Plus, the advantage that no additional infrastructure is needed to extract mobility information, since they were designed to collect for public transport fare and allow mobile communication network usage.

In addition, we will also cover research working with GPS data. Its high-resolution and accuracy has allowed the creation of preprocessing and data mining techniques as well as inference models that go beyond SC-AFC and mobile phone data. Plus, the recent drop in cost that allowed the widespread use of GPS sensors in mobile phones and other devices, outlooks an obvious potential for wide-spread application as urban mobility sensors in the future. Other supplementary datasets being reviewed are Points of Interest (POI) information, Census and Surveys, and Land use information. As commented on Calabrese et al. (2015), the purposes of the supplementary datasets are three-fold. 1) to validate findings extracted from the analysis of large-scale human mobility sensors. 2) to define scaling factors to extend results to the overall population. 3) to augment information about urban space, to be able to extract higher level patterns.

2.2.1 Mobile phone data

Purpose: Large-Scale Human Mobility Sensor

From the ubiquitous computing devices, mobile phones have the highest levels of penetration rate. While conventional mobile phones usually only sporadically exchange information with cell-towers, the widespread use of smart phones - the level of penetration rate for smart phones in Singapore reaches 88% of the population[2] - provide appealing new opportunities to inform travel demand model. Smart phones not only exchange data much more frequently with the mobile network provider allowing more continuous tracking, they also carry a series of additional sensors provide such as GPS that can be used to better understand mobility patterns. For instance, Airsage[3] is a company that process mobile phone data for transport planning applications.

Mobile phone networks Whether it is GSM, CDMA or LTE, mobile phone networks require regular and frequent handshakes (i.e. pings) between mobile phone devices and cellular communication antennas. In order to provide service to the users, mobile phone networks are constantly and frequently determining the location of the mobile phone devices even if it they are simply on standby. The user's location is calculated by determining the location of the cell antenna closest to the handset. This results in a precision equal to the size of the cell antenna coverage, which can range among few hundred metres in urban areas.

In order to understand the events that generate the user location updates we need first to know how the mobile phone network is constituted. The service coverage area of a given mobile phone

[2] *The Connected Consumer Survey 2014/2015*. https://www.consumerbarometer.com/

[3] www.airsage.com

network is divided into smaller areas of hexagonal shape, referred to as cells. Each cell is the area in which one can communicate with a certain base station antenna (also referred as cell tower). A Location Area (LA) is a geographic area covered by base stations antennas belonging to the same group. For users to be reached wherever they are in the network coverage area, we can divide the update location procedures in network-triggered and event-triggered.

Network-triggered location updates occur when a mobile phone is,

1. being switched on and connects to the cellularnetwork.
2. involved in a call and moves between two different cell areas (i.e. handover)
3. on standby and moves into a cell which belongs to a newLA.
4. polled by the network as its associated timer has come to an end (i.e. periodic location update, usually every 2 h)

Event-triggered updates happen in the following situations.

1. When a call is placed or received
2. When Short Message Service is used (sending and receiving).
3. When the user connects to the Internet (e.g. to browse the web, or through email periodically server check)

Most studies found in literature that use mobile network data analysing individual mobility patterns from mobile phone data make use of Call Detail Records, which is a subset from the mobile network data used for billing purposes, but also include event-triggered updates. However, there are also a series of studies that make use of the full spectrum of mobile network data, including both the event-triggered and network-triggered updates.

Moreover, precision of cell antenna location data can be improved between a dozen to hundred meters when information is triangulated with signals from other cell antennas. (International Transport Forum, 2015). By using the Timing Advance, which is a value that corresponds to the length of time a signal takes to reach the cell tower from a mobile phone, information can be triangulated from different cell antennas to have a more accurate estimation on the user's location. Other techniques used are based on the signal strength received by the mobile phone using known irradiation diagrams and propagation modelsfor localisation.

2.2.2 Smart card data

Purpose: Large-Scale Human Mobility Sensor

Smart Card Automatic Fare Collection (SC-AFC) systems are built on radio-frequency identification (RFID) technology. The objective of RFID is to use radio waves to exchange data between a reader and an electronic tag for the purpose of identification and transaction operations. Specifically, SCAFC systems use NFC (Near Field Communication) technology which ensures secure short range RFID transactions. Since 1990, the use of smart card has become significant in many sectors since it is perceived as a secure method of user validation and fare payment. (Trépanier et al., 2007).

For the case of public transportation, commonly, users have to tap in their smart card onto the reader device at the entrance of buses or metro stations and tap out at the alighting bus stop or metro station. Besides revenue collection, large quantities of individual detailed information are collected such as boarding times, boarding stations, alighting times, alighting stations, vehicle identification. This represents a huge potential in better understanding travel behaviour and improving current transport systems.

2.2.3 GPS

Purpose: Model Transferability

GPS information allows the collection of a detailed spatio-temporal trace describing the mobility of an individual. All smartphones are usually equipped with GPS sensors and also some conventional mobile phones nowadays include it. In order to calculate position, GPS uses information of at least four satellites in a method called *trilateration*. In open areas accuracy can be achieved up to 5 meters but it degrades, however, in areas where GPS signals are impaired by tall buildings or trees and inside of buildings.

Assisted-GPS (A-GPS) increases location accuracy in urban areas by combining GPS location signals with cellular location data providing an under 10 meters precision. Similarly, other forms of hybridised GPS location systems can include the use of Wi-Fi network signals through the tracking of media access control addresses (MAC addresses) within a network of Wi-Fi routers.

2.2.4 POIs

Purpose: Urban Space Augmentation

Points of Interests (POIs) are a list of business and important places to visit in a city, including their name, classification and location. There are many possible different sources: Yellow Pages, Google Places, Yahoo PlaceFinder, which might provide different information. For instance, in Google Places we can find in addition opening times, reviews and hourly-estimates on the crowdedness of a place. Furthermore, opportunistic POI datasets can be derived from crowdsourcing platforms and social networks (e.g. Foursquare, Flickr, Twitter, etc.). The importance of these datasets is the potential to serve as complementary datasets to augment urban space information and thus, improve activity and places inference estimations.

2.2.5 Census and Surveys

Purpose: Validation and Scaling Factors

Census and surveys provide datasets related to very different areas: demography, health, education government and security, communication and transport, etc. Such datasets can be used to: 1) validate home and working areas, 2) validate city patterns such as hotspots, commuting, traffic flows, 3) validate land use. The main advantage is the very refined spatial resolution which is often the census block. The main disadvantages are that they are updated usually only every 5 to 10 years.

Other use of Census information is that it provides the means to perform scaling expansion from information derived by large-scale human mobility sensors.

While the use of census and other conventional survey data has a long tradition in travel demand modelling and is well documented, the review of the related work does not fit the scope of this review. However, since census and survey data is still relevant to enrich and mobile phone data through data fusion, this review brings together pioneering work in those areas.

2.2.6 Land Use

Purpose: Urban Space Augmentation

Land use datasets offer access to various information that allows to characterise an area based on its planned and effective land use. Such land use data usually specifies for each plot the designated usage purpose in case of built up environment also the usage intensity. However, as different authorities use different land use classifications, the developed models are usually customised to local conditions which restricts the direct transferability between regions. Within the scope of this review, land use data is of interest due to its potential to impute the purposes of activities as identified from SC-AFC and mobile phone data.

3 Smart Card Data

Smart Card Automated Fare Collection (SC-AFC) systems are used in many public transportation systems around the world and continue to be adopted by public transport operators. The main objective of introducing smart cards for public transport is its ability for flexible and secure fare collection. However, the information generated with any transaction (time and location) has soon become recognised as a rich data source for transport and urban planning and science. From public transport ridership analysis to the creation of OD matrices, smart card data offers insights on urban dynamics and mobility patterns of a city's public transport. The following chapter presents the literature on ways to exploit smart card data for transport planning – from individual trip reconstruction to the estimation of OD matrices and its inclusion to agent-based simulations.

3.1 Individual Trip Reconstruction

Implementation of SC-AFC systems varies depending the city and its fare policies. Some cities like Amsterdam, Sydney, and Singapore charge public transport fares based on the total distance travelled, regardless if it is on a bus or train. This requires commuters to tap their smart card when they board and tap it again in the alighting stop or station. Nonetheless, other cities like London have non-ladder fares in which the fare is identical for the whole line regardless of where you get on or get off, thus commuters are only required to tap the smart card once. In any of the cases, to further analyse human mobility the main challenge of mining smart card data is to reconstruct the individual trips.

3.1.1 Estimating Alighting Stops

For SC-AFC systems in which it is only required to validate the boarding location, the first step is to estimate the alighting stop or station. Generally, to infer the alighting stop the *Trip-Chaining* algorithm is used based on two explicit assumptions by Barry, Newhouser, Rahbee and Sayeda (2002). The first one states that after a trip, users will return to the destination of the previous trip station; the second one, that at the end of a day, users will return to the station where they boarded for the first trip of that same day.

Several efforts have been done to improve the original idea by Barry et al. (2002). Zhao et al. (2007) expanded the idea to rail-to-bus sequences. Trépanier, Tranchant and Chapleau (2007) incorporated the possibility of looking at the next day, even observing weekly travel patterns to complete missing information for the bus system of Gatineau in Quebec. Munizaga and Palma (2012) proposed a multimodal public transport methodology using time constraints instead of distance constraint. For these studies, the reported success varies from 66% to a 80% of the individual trips reconstructed.

Furthermore, a different approach to reconstruct individual trips from smart card data which is based on a semi-supervised learning algorithm has been presented by Yuan, Wang, Zhang, Xie and Sun (2013). They proposed an integrated learning method in which they align the monetary, geospatial and temporal spaces to extrapolate a series of critical domain specific constraints. They incorporate those constraints in a semi-supervised conditional random field algorithm to infer the exact boarding and alighting stop even if there exist records on trips with unknown boarding and alighting information. Given only 10% trips with known alighting/boarding stops, they inferred more than 78% alighting and boarding stops from trips with missing information. The relevance of the work is not only the reconstruction of origin-only labelled trips, but a systematic way to recover individual mobility history from urban scale smart card transactions. This can be helpful as a pre-processing stage for later analysis or incorporation to transport demand models.

3.1.2 Stages, Trips and ODs

After the alighting location is known, the second step in the individual trip reconstruction is to infer whether the alighting location is the final destination (thus the trip is complete) or if it is only a stage of a multi-stage trip (i.e. a transfer). The common approach to identify stages is with a time-based rule. For instance, Munizaga and Palma (2012) used a 30 minute rule. If a person stays longer than 30 minutes in a particular point, then it is said to be the destination. For the case of London, Seaborn, Attanucci and Wilson (2009) recommended elapsed time thresholds depending on the transfer type — 20 min for underground-bus, 35 min for bus-to-underground, and 45 min for bus-to-bus.

Having information only on smart card data draws a limitation in identifying the spatio-temporal dimension of an individual performing an activity, since public transport is not exclusively used for all trips throughout a day. Chakirov and Erath (2012) describes the limitation with the concept of public transport trip consistency, in which consistency means that a person who arrived to the activity location by public transport, has to leave it after ending the activity also by public transport. Although smart card data do not record any other means of transport except from public transport, the most obvious cases of inconsistency can be identified by analysing the distances between the alighting location of the last journey and the boarding location of the following journey. This allows to identify if other means of transport such as taxi, car or walking must have been used in between.

For instance, on a typical workday in Singapore, Chakirov and Erath (2012) found that from persons with more than one journey recorded in the smart card data, 90% of the journeys, start less than 1 km away from the previous alighting location. This indicates firstly, that the majority of public transport users don't switch to other modes of transport between public transport journeys and therefore have consistent journey chains. And secondly, that with some degree of uncertainty an area can be limited for possible activity locations.

Once individual trips have been reconstructed to the point of knowing boarding locations and final destinations, a possible application is the calculation of a public transport OD matrix. It is important to take into account trips which could not be able to reconstruct. For this situation a typical solution is to build expansion factors. The work by Munizaga and Palma (2012) shows how to build expansion factors for smart card data trips associated with an origin but not with a destination, and for trips associated with no origin or transaction. For the former ones, it is assumed that the distribution of trips is the same as that of other trips with the same origin; whilst, for the latter case, the distribution of trips is associated only by their time disaggregation.

3.1.3 Activity Identification

Public transport consistent trips can be further studied to introduce semantic meaning to the inferred locations. Devillaine, Munizaga and Trépanier (2012) present a direct rule-based classification including information on the card type and the temporal attributes of the trips. *Work* is assigned for adult cards, for which the activity is longer than 2 hours and the trip before the activity is not the last of the day. Similarly, *study* is assigned for student or underage cards, with activities longer than 5 hours and the activity not being the last one of the day. Finally, *home* is assigned if the trip after the activity was the last of the day, and *others* was assigned to the rest of activities. The criteria used represents a generalisation on the working/studying and staying at home behaviours, which might not be accurate for an important percentage of the population. Shortcomings of such a rigid classification can be improved by the introduction of a probabilistic choice models, which can take into consideration land use information as well.

Such is the case of the work by Chakirov and Erath (2012). They proposed a multinomial logit model with activity duration, activity start time and land use as the the utility variables to match a discrete-choice space consisting of work activity, home activity and other activity as the target labels. Utilities for the model were constructed mainly using piecewise linear functions. For the case of activity duration and start time, the utility function was calibrated using information from the local travel diary survey, whilst for the land use information stemmed from the urban planning authority's Master Plan.

Another probabilistic approach was proposed by Isaacman, Becker, Cáceres, Kobourov, Martonosi, Rowland and Varshavsky (2011) who built a continuous-space model to detect home and work locations. They introduced a score function obtained from a logistic regression and calibrated from a set of trained users. Similarly to Chakirov and Erath (2012) the *home* and *work* location labels were identified mainly by the time factors associated to the events. However, the main differences between the two probabilistic approaches relies not in whether they have chosen discrete or continuous space for their methods, but that the calibration process diverges in using information from other data sources (i.e. household travel survey and land-use) in a transfer-learning scheme (Chakirov and Erath, 2012) against the traditional learning scheme, in which calibration comes from a labelled training subset (Isaacman et al.,2011).

After having obtained the home and work locations, a validation process is required to determine the model's accuracy. One possible validation process is the one used by Yuan et al. (2013). A 2D Kernel Density Estimation (KDE) was applied to identify hot spots that enabled a comparison against data derived from travel diary surveys.

3.2 Agent Based Transport Models and Simulation

State-of-the-art methods for travelling demand models are currently led by multi-agent simulation frameworks. This new paradigm was put forward to overcome some of the flaws of the classic four step model, namely to ensure behavioural consistency, include temporal travel dynamics in a continuous manner and maintain the disaggregate nature of transport demand through out the modelling process. The disaggregate nature of smart card data represents an appropriate input to such models. Moreover, by assuming that each unique card ID represents an agent, demand for a multi-agent based transport scenario can be directly derived from the smart card data.

The work of Bouman (2012) represents a first attempt to implement an agent-based micro-simulation of public transport for the cities of Amsterdam and Rotterdam. Based only on smart card data, the main challenge of the work is the generation of the agents' activity plans. For such task, they focus on the extraction of commuters' *home-work-home* pattern looking at several days of the same user. Work and home stations are identified as the two most visited stations during the weekdays, from which *home* is identified as the most visited station during the weekend. For smart cards ids whose data was not suitable to be fitted with such a pattern, but at least an activity chain could be reconstructed for one particular day, travel demand was specified by introducing dummy activities at the intermediate stations of the tour. Finally, for highly irregular travel patterns, a new agent for each of the remaining trips was generated.

The synthetic-population generation process suffers from different limitations that mainly stem from the various assumptions imposed in the modelling process. Opportunities for future research can be identified in a more accurate and validated representation of the actual travel demand including the inference of trip purposes and socio-demographic characteristics of the traveller. To this end, we consider the long observation periods of smart card data as an opportunity to apply modern data mining techniques to infer additional information. Along this line, (Bouman et al., 2013) for example explored how the concept of *eigenbehaviours* (Eagle and Pentland, 2009) can be applied to derive spatio-temporal patterns.

Another challenge when using smart card data for simulation is to model potential interactions of public transport vehicles with other transport modes (i.e. cars). A recent work by Fourie, Erath, Ordóñez Medina, Chakirov and Axhausen (2016) develops a simplified agent-based transport simulation for Singapore's public transport. In contrast to Bouman (2012), the interaction with private vehicles was accounted by introducing a stochastic model of the speed of buses between public transport stops and bus dwell time behaviour at stops. This allows not only to improve the simulation time substantially, but also to predict the operational stability of alternative public transport schedules, making simulations of system-wide network redesigns possible. For this purpose, they adapted three inputs to the MATSim environment.

Firstly, a reconstruction of bus trajectories from smart card data was developed. Given all boarding and alighting transactions of bus users, the position in space and time of the corresponding buses was estimated. They imputed the time it takes for a bus to travel between bus stops locations by grouping its transactions at each stop into sets that represent bus dwell operations. Then, from the reconstructed bus trajectories they determined the number of services and the time when the services start for every bus line in Singapore. For the particular case of train services, they took the start times from the Google Transit Feed Specification (GTFS) since especially during peak hours train trajectories cannot reliably inferred from public transport smart card data Sun et al. (2012).

Secondly, they needed to generate activity plans for each agent in the simulation. To that end, they established a 25 min threshold to identify the final alighting location of each multistage trip to not split journeys at transfer points. Since smart card records only document boarding times but not when a person actually arrived at the bus stop: given that average headway for most bus services is 10 minutes or shorter, they assumed a uniform arrival time distributions and randomly drew the actual arriving time from the bus stop with the only parameter being the corresponding headway between consecutive services of the specified line.

Lastly, as only public transport vehicles were simulated, a simplified network offered the opportunity to lower the computational demands of the simulation substantially. Instead of the MATSim queue model, a stochastic travel time model was introduced to model travel times between two subsequent stops. The model was fitted based on a multinomial regression model assuming that stop-to-stop follow a normal distribution Fourie (2014). As shown by Sarlas and Axhausen (2015) the parameters that determined the speed of vehicles in a network link were related not only to the level of demand on the link (taken from smart card data), but also to the topographical information contained in the network description. To account for dwell time variability in the simulation framework they included the model presented by Sun, Tirachini, Axhausen, Erath and Lee (2013).

As a case study to showcase the abilities of the model, they simulated the impact when splitting one of the longest bus lines in Singapore. The results suggested that incidences of bus bunching can be significantly reduced during the morning peak hour, and that headway reliability is also improved considerably.

The work of Fourie et al. (2016) shows one of the possible integrations of big data algorithms within an agent-based transport modelling framework. They enhanced the simulation results of a complex model by introducing what is called in the computer science literature as a machine learning *surrogate*. They substitute a compartment of the whole model with the behaviour obtained from the statistics of the smart card data. The results not only represent a more accurate representation of the real world, but also an improvement of the overall computation time.

However, still several limitations have to be addressed such as implementing the passing behaviour in the queue simulation, the reconstruction of train trajectories, a better representation of walking, waiting and transfer activities to better represent route and mode choice preferences. In addition, public transport smart card data obviously do not contain any information of motorised travel demand and active mobility. To tackle the aforementioned problems, additional datasets can be included.

4 Mobile Phone Data

Data derived from mobile phones network transactions constitute a potential source of information for models of daily activities and transportation. In contrast with household survey interviews, mobile phone data offers large sample sizes and long observation periods at negligible costs. However, one has to overcome the challenges of processing mobile phone traces for trip reconstruction — the information contain is such data streams is low in spatial resolution and sparse in time. Concretely, the precision on the location estimates depends on the cell tower distribution in a given area, whereas the frequency on the location updates is characterised by each user's usage.

Moreover, for the different studies found, the quality of the data also depends on the type of dataset provided by the mobile network operator. While some studies work with the full spectrum of mobile phone management signals (e.g. location area (LA) handles, device updates), others work with the Call Detail Records (CDR) subset, or with already preprocessed data (e.g. triangulated location estimates). Thus, the general challenge in using mobile phone data is how to robustly extract people's trip sequences from sparse and noisy measurements and enrich the extracted trips with semantic meaning (i.e. trip purpose) (Widhalm et al., 2015).

In the following, we present a series of studies that aimed to reconstruct individual trips from mobile phone data in order to extract mobility patterns. Thereby, we focus on approaches that allow the generation of OD-matrices and applications with potential to be adopted for agent-based simulations. For this purpose, we have divided the relevant literature based on their methodological approach and their scope. The first group of studies introduced resembles the traditional trip-based approach, a second group focuses on extracting stay locations from noisy mobile phone traces, and the third group attempts to infer activities performed at the extracted locations. At the end of the chapter, we finalise presenting the studies that have been made using mobile phone data specifically for the generation of agent-based simulations.

4.1 Trip-based OD reconstruction

The first approaches found in literature that aim at understanding mobility flows based on mobile phone refer to attempts to generate OD matrices. Caceres, Wideberg and Benitez (2007) performed a pilot study using CDR simulated data to match origin and destination areas with the LAs used for the management of the mobile phone network. The idea behind the study was to demonstrate the correlation between the movements of anonymous mobile phones and vehicle displacements.

This idea was then formalised by Wang, Hunter, Bayen, Schechtner and González (2012) with the definition of the *transient* OD matrix. The concept recognises that even if segments of the trip are unobserved in the CDR (e.g. real origin and destination locations) due to mobile phone inactivity, still a large portion of the actual ODs are retain and can be use to analyse road usage patterns. The *transient* OD matrices are constructed by simply counting trips for each pair of consecutive calls made within the same hour from two different towers, and then the OD trips are assigned to the road network by a shortest path algorithm.

Similarly, Iqbal, Choudhury, Wang and González (2014) demonstrate the development of OD matrices using CDRs from Dhaka, Bangladesh, and traffic counts from a video vehicle detection system. Firstly, tower-to-tower *transient* OD matrices are generated and then associated with corresponding nodes of the traffic network converting them to node-to-node *transient* OD matrices. Then, the *transient* OD matrices are scaled up to match the traffic counts. To determine the scaling factors, an optimisation-based approach is used which minimises the differences between observed and simulated traffic counts at the points where the traffic counts are available. Lastly, for the estimation of the final OD matrix they introduce correction factors to account for the mobile phone market penetration rates and mobile phone usage.

As discussed by Jiang et al. (2015) one of the problems of the trip-based approach is that it can introduce biases when CDR data are low in spatial resolution. In addition, the former methods are not able to handle noisy measurements from raw mobile phone traces. In order to avoid these issues, the approach of the next group of studies presented is based on parsing the trajectories observed into stay-locations.

4.2 Stay location-based OD reconstruction

The fundamental premise of activity-based travel models is that travel demand derives from people's needs and desires to participate in activities. In the following, we present a set of methodologies that have been proposed in literature to identify activity locations from mobile phone data and therefore are potentially relevant also for building activity-based transport demand models. Activity locations are identified by filtering out *passing by* points, and estimating arrival and departure times from raw mobile phone data. In contrast with trip-based models, individual trips are obtained from the flows between the identified stay-locations. However, in order to obtain such information, sophisticated mobile phone data processing algorithms are required.

Additional to a lower spatial resolution in comparison with GPS traces, mobile phone data suffers from a phenomenon called *supersonic jumps* or *signal jumps* (i.e. outliers). These are events that suddenly occur kilometres away within a short period of time. Although such jumps usually are system inherent data noise, some jumps might be triggered by external mechanisms in aims to protecting the privacy of the users (Horn et al., 2014). To use mobile phone data for accurate traffic modelling, these shortcomings must be considered in order to derive realistic trajectories.

Temporal-based clustering Schlaich, Otterstätter and Friedrich (2010) work with Location Area (LA) updates from a region of southwest Germany. The algorithm proposed is built on the principle that if a user remains a considerably longer time in a location area than the time required for directly traversing the area, the user potentially starts or ends a trip in the respective location area. For this purpose they suggested a 60 min rule in which if the time period between the first login and the last logout of a multiple visited LA is 60 min or more, then it is considered to be a stay location. In addition, as a strategy to deal with signal jumps, they calculated a *jumpiness factor* and deleted user entries that exceeded a given threshold.

Certainly, the approach suffers from several limitations due to the resolution of the trips extracted being at the broad LA level and not the cell-tower area level. For instance, in their preprocessing step, the decision to delete consecutive data points from users with the same location areas, disables them to estimate arrival times and activity durations. Also, since their method requires a minimum of three LAs, they deleted users that show records of less than three different LAs. Hence, a more robust methodology would be desirable that can handle noisy signals and outlier points without the need to delete any entries, and that can be able to estimate trips at the cell-tower resolution.

Distance-based clustering Calabrese, Lorenzo, Liu and Ratti (2011) proposed a method to identify trips at the cell-tower level based on Call Detail Records (CDRs) generated from phone calls, messages and internet usage. They also included a strategy to handle noisy traces employing different clustering techniques. For the preprocessing step they first characterised the individual calling activity and verified that it was frequent enough to allow monitoring the user's movement over time with a fine enough resolution. Then, they applied a low-pass filter with a 10-minute resampling rate and a clustering technique to identify minor oscillations around a common location. As for the extraction of stay-points, they performed a distance-based clustering to fuse points within a 1 km area. The centroid of the cluster was defined to be a virtual location and in a final step, individual trips were reconstructed by connecting the paths from the identified virtual locations. However, the methodology lacks to robustly filter out *passing by* events.

Frequency-based clustering An alternative to identify important places (i.e. stay-locations) from temporally sparse and spatially coarse CDRs is presented in Isaacman, Becker, Cáceres, Kobourov, Martonosi, Rowland and Varshavsky (2011). They assume that the most visited cell-towers are connected to important places in a person's life. Instead of using temporal or spatial clustering algorithms to obtain those locations, they use the frequency of cell-tower visits. The method consists in sorting cell towers based on the total number of days they were contacted and applied a cluster leader algorithm. After obtaining the cell tower clusters, they used a logistic regression model trained on the activity behaviour of 18 volunteers. The explanatory variables included on the model were the number of days during which any cell tower in the cluster was contacted, the span time between the first and last contact with any cell tower in the cluster, the working hour events (between 1pm and 5pm) and the home hour events (between 7pm and 7am). While the proposal performs well with the identification of primary activity locations (e.g. home, work), it is not designed to extract secondary activity places for which they can be confused with en-route events.

At this point, we can recapitulate and reformulate the main ideas towards extracting places from raw mobile phone data according to the data mining pipeline suggested by Jiang et al. (2015). First, the need to eliminate outlier noise and signal jumps between towers. Secondly, the need to cluster points that are spatially close and temporally adjacent into a single location. And thirdly, the need to agglomerate points that are spatially close but not necessarily adjacent in temporal consecutive sequence, since we are interested in the unique stay locations that a user frequents. In addition, an estimation on activity start times and durations is also needed. The following studies represent the latest endeavour (from 2013 to 2015) to mine location points from mobile phone data.

Preprocessing techniques For the first goal, Horn, Klampfl, Cik and Reiter (2014) perform an evaluation on three different types of filters to detect outliers on mobile phone traces: a Recursive Naive Filter, a Recursive Look-Ahead Filter, and a Kalman Filter. On the one hand, the first two basically act as low-pass filters (Calabrese et al., 2011). They smooth out large positioning errors by introducing an upper bound constraint on the travel speed. Hence, the speed is calculated for each each consecutive pair of points (Recursive Naive Filter), or each triad of points (Recursive Look-Ahead Filter) and compared to a certain threshold. On the other hand, the Kalman Filter is a probabilistic approach that reconstructs the trajectory. The results demonstrated that the Recursive Look-Ahead Filter performed better as it eliminated the outlier points, and in addition maintained the accuracy of the trajectories. Although the Kalman Filter also eliminated the outlier points, the trajectories lose accuracy. However, the satisfactory results of Ficek and Kencl (2012) to extend the spatial resolution of the Reality Mining Dataset (Eagle and , Sandy) using a Gaussian Mixture Model, suggest that given the low-resolution of CDRs, more complex probabilistic filters are needed in order to outperform the naïve approaches.

Time-distance clustering For the second and third goals (i.e. location extraction), Jiang et al. (2013, 2015), Toole et al. (2015), and Alexander et al. (2015) used time and distance clustering techniques to filter out *passing by* points. Firstly, they grouped points that are spatially close by measuring the distance between two consecutive points and comparing them to a distance threshold (e.g. roaming distance of 300 m). Then, the clusters obtained are considered to be potential stays if the time between the first and the last observation in the cluster are separated by a time greater than a time threshold (e.g. 10 min). Then, the geographic location of the potential stay is set to be at the centroid of all points within the cluster. Due to noise in locations, multiple potential stays that are actually the same place may be estimated at a slightly different geographic coordinate on different observation days. To account for this, a final agglomerative clustering algorithm is used to consolidate candidate stays to a single semantic location regardless of the temporal sequence of the records.

Trip Validation It is important to verify for the algorithms proposed, that users with more phone activity do not have systematic differences in travel behaviour. For instance, that there does not exist a correlation between the number of places detected and the mobile phone usage. Jiang et al. (2015) segmented users into 5 groups according to the frequency of total number of phone usage observations. Then they examined for each group the daily travel patterns, including daily number of trips, and daily number of unique destinations. Finally, they compared the frequency distributions of both the number or trips and the daily number of unique locations and conclude that they follow similar patterns.

Activity start times and durations After having identified the stay locations, Widhalm et al. (2015) continue the study by estimating the arrival time as the average between the earliest record in the arrival activity (i.e. the upper bound of the arrival time), and a lower bound estimate, calculated as the sum of the latest record at the previous location plus the travel time between the previous and present location. The travel time was determined as the distance between the consecutive centroid of clusters divided by an assumed travel speed. The same process was performed for the expected time of departure, and the activity duration calculated by subtracting both estimates.

Another alternative to infer arrival/departure times of activities is the one of Alexander et al. (2015). They proposed to use probability density functions of activity durations derived from the National Household Travel Survey. They constructed six hourly distributions for weekdays and weekends and the following trip purposes: home-based work (HBW), home-based other (HBO), and non-home based (NHB). Then, they randomly generated the departing time within the time window of observations, using the distribution that corresponded to the day (weekday, weekend) and the trip purpose (HBW, HBO, NHB).

4.2.1 Expansion Factors

After locations have been extracted from mobile phone data, trips can be generated by connecting consecutive locations and OD matrices can be potentially determined. However, in order to project the mobile phone sample to the entire population, expansion factors need to be carefully calculated, avoiding any socio-economical biases, and correctly transforming the individual trips recorded from mobile phone data into vehicle flows while taking into account shared rides (and also trips made by public transport).

Several authors (Jiang et al., 2015; Alexander et al., 2015; Toole et al., 2015) have proposed the expansion factor calculation by identifying the tower a user is connected to while being at home. Since the ratio of mobile phone users to the population is not uniform within the research region, each user is assigned a home census area, and expansion factors are computed foreach area by measuring the ratio of the total population living in the area, and the sum of users whose home-tower was identified inside the area. General OD matrices are built, and if a certain mode of transport wants to be considered, the vehicle OD matrix is approximated by weighting the total number of user trips by the vehicle usage rate in the home census zones. However, the limitation of the methodology is the generalisation of travel patterns by census areas which might reflect biases in the vehicle tripsODs.

Alternatively, Zhang, Qin, Dong and Ran (2010) build an expansion factor based on a probabilistic approach. They firstly looked at the problem on how to avoid multiplication of trips when there exist more than one mobile phone probe in a vehicle. They calculate a conversion factor from mobile phone probe flows into equivalent vehicle flows, using the following assumptions: —
1) Mobile phones in close proximity (i.e. the same car) generate signal transition events at exactly the same time. 2) There is a very small probability that some parallel travelling cars are crossing at least two LA boundaries at two same timestamps. 3) Within the saturation headway (typically 2 seconds) there is only one vehicle crossing an LA boundary in each lane. Secondly, they computed a conditional probability for mobile phone ownership including mobile phone market penetration data, market shares for a given mobile phone carrier, and age and income distributions from census data. Finally, for the purpose of projecting to the population vehicle OD matrix, they calculated the scaling factor through a Horvitz-Thompson estimator, in which the conversion factor from mobile phone to vehicle flows was included, as well as the posterior probability of mobile phone ownership.

In contrast with the home-based approach, the model takes into account the conversion from individual mobile phone flows to vehicle flows (i.e. shared rides) not as an *a posteriori* step, but integrated in their model. Plus, the fact of characterising the population mobile phone ownership by age and income, instead of zones, avoids socio-economical biases more accurately in the projection process.

4.3 Activity-based ABM

So far we have obtained locations, start times and activity durations from raw mobile phone traces. However, to derive an activity-based travel demand description to be used for for agent-based simulations, there is a need to identify the trip purposes, in order to construct the agents daily plans. To this extent, we begin by looking at the existing activity inference methods for mobile phone data, followed by studies who have built a mobile phone data driven agent simulation.

4.3.1 Activity Inference Methods

In traditional survey data, activity purposes are revealed by individuals who answer the travel diaries; whereas in the mobile phone data, activity types are latent. Furthermore, none of the data sources (travel survey or mobile phone data) pinpoint the exact locations of the trip destinations, but areas surrounding these precise locations. Generally, we can find in the literature two different approaches for Activity Inference — time-frequency rules, and probabilistic models.

Activity inference by time-frequency rules One of the direct ways to infer contextual information such as location's function or trip purpose, is by time-frequency rules. Several authors (Toole et al., 2015; Alexander et al., 2015; Jiang et al., 2015) have improved the general idea shown in Wang et al. (2012) and Iqbal et al. (2014) in using both visit frequencies and temporal data to identify *work*, *home*, and *other* locations. A user home location is defined as the stay point most frequently observed between 8 pm and 7 am during weekdays and weekends. Whereas, work location is defined as the stay point, other than home, that users visits the most between 7 am and 8 pm on weekdays. Since some individuals do not work, the work location is left blank if the candidate location is not visited more than once per week or if the location is less than 500 m away from the home location (to avoid work identification through signal noise rather than a distinct location). Another variation (Alexander et al., 2015) is that work location is identified as the stay to which the user travels the maximum total distance from home location, to identify evening and night shift jobs. All remaining non-home or work stay points are designated as other.

Activity inference by probabilistic modelling Time-frequency rules to infer activity context are a straightforward method, but may not be useful for certain groups in the population. More robust inference methods can be applied using probabilistic models, in which other datasources can be naturally included (semantic-enriched geographic data, POIs, etc.). Thus, we can reference the spatial information and characteristics of destinations to build probabilistic models to infer activity types at different destinations in space and time.

Jiang et al. (2013) formulate the target probability as the probability that an individual performs a certain activity at a certain time depending (conditional) to her/his destination information and her/his extracted unlabelled activity chain (i.e. motif, (Schneider et al., 2013)). Once the problem has been reformulated as a target probability, the next step is to choose a probabilistic inference model suitable to handle the dependencies on the random variables and to calibrate the parameters of the model. For the latter problem, two approaches exist based on the availability of training data: — Supervised learning, and Unsupervised learning.

Activity Classification Activity classification is done through supervised learning algorithms and it requires a labelled training set. It addresses the problem of identifying the activity type for each stay, given the extracted stay sequences and labelled travel survey data and requires to predefine the categories that the classification algorithm will be trained to detect. For instance, Yang et al. (2015) work with 5 different activity categories: *home*, *work*, *leisure*, *shopping*, and *other*. For their feature vector they use activity start times, activity duration and the location. They employed a Bayesian classification algorithm which recognises the transition probability between activities from the labelled training data. Given that home and work have more distinct start time, duration and location distribution, they performed a stepwise classification, firstly for the work and home labels, and secondly to distinguish between leisure and shopping using the transition probabilities.

Activity Clustering When there is no available labelled training data, we can still reveal the spatial-temporal structure of activities by performing unsupervised clustering. Widhalm et al. (2015) and Yang et al. (2015) clustered the stays into meaningful categories based on stay start time, duration, sequence of stay locations, and a vector of land use shares. To model the dependencies between the explanatory variables and construct the joint distribution they used a Relational Markov Network (based on the work of (Liao et al., 2005)), which is an extension of undirected graphical models known as Markov Random Fields. To compute the joint posterior distribution they used Rejection Sampling, which is a technique that takes samples from the posterior distribution. Finally, to train the network in an unsupervised way, they used Expectation Maximisation (EM) algorithm and applied this methodology for case studies in Boston and Vienna.

The resulting clusters reflect trip chains and activity scheduling patterns that agree well with data obtained from traditional surveys. Moreover, the comparison between both cities showed similarity in their clusters. Still, several future improvements are suggested. Firstly, to study the relationship between the automatically discovered activity clusters and conventional activity types used in traditional surveys more closely. Secondly, to examine and include the interaction between land use and travel behaviour (e.g. density, regional accessibility, roadway connectivity, etc.) in the modelling, for instance how the inclusion of point-of-interest (POI) databases allows to further improve the accuracy of the results. And thirdly, the inclusion of the results in a simulation model (i.e. agent-based model) to compare the resulting traffic flows to actual traffic measurements as a validation step.

4.3.2 Mode Inference

Transportation mode inference from ubiquitous computing devices is a common challenge in the literature. However, the majority of the methods proposed are based on mobile phone sensors, such as GPS, accelerometers, and gyroscopes, in which fine-grained sampling is available. On the light of getting large-scale observations for transport planning purposes in urban areas, a broader classification can still be made based only on CDRs. These methods infer the transportation mode by estimating the mobile phone's speed and associating it with a transport mode. For instance, Wang, Calabrese, Lorenzo and Ratti (2010) used the information on a trip's origin and destination, as well as its travel time to classify the travel mode in three group: car, public transit, and walking. Firstly, they filtered their dataset to include only trips with distances more than 3 km, and users with update location frequency of more than 1 per hour. They followed by grouping trips according to their origins and destinations, and performed k-means clustering to differentiate between the modes. Finally, the results were validated against Google Maps travel times information.

4.3.3 Synthetic population from Call Data Records

As we have seen on the lines of this chapter, many studies have been made to mine individual travel behaviour from mobile phone data for the purpose of having more frequently-updated and opportunistic collected data to enhance four-step submodels (e.g. trip generation, distribution and travel mode), and creating dynamic OD matrices. However, one of the challenges still remaining consists of using mobile phone data in the context of activity-based transport modelling frameworks and agent-based simulations.

The work of Zilske and Nagel (2014) represents a pilot study of such an approach. They seek to replace travel diaries with sets of CDRs as input data for agent-oriented traffic simulation. For this purpose, they generated synthetic CDR data from a MATSim simulation using a plug-in that introduces cell coverage and a mobile phone usage model for the agents. CDR information is generated on arrival at or departure from activity locations. After CDRs are generated, they identified every observed person with a MATSim agent and converted every call information into an activity. The main limitation on the study is the simplification of the mobile phone data generation process. The full-spectre of mobile phone data is not represented (e.g. cell handovers, automatic location updates, mobile internet usage), plus CDR information was not generated for instances outside the activity locations (no mobile phone activity on-the-go). These resulted in an underrepresentation of the traffic simulation in MATSim.

However, in their next paper on this topic, Zilske and Nagel (2015) propose to mitigate the underrepresentation and reduce the spatio-temporal uncertainty by fusing the CDR dataset with traffic counts. They computed an expansion factor of the population to compensate for the underestimate demand and matched it to traffic counts. From each trace they created several agents and used the expanded population as a buffer (with the introduction of a stay-at-home plan probability) to steer the demand towards matching the known link volume counts. Additionally, in order that the modelled travel demand matches with the traffic counts, they introduced a parameter to alter probability by which a particular activity plan is chosen. Intuitively, the offset was a calculation based on how much a specific choice of a plan contributes to the whole traffic system fitting to the traffic counts.

4.3.4 Mobile phone data driven MATSim

The latest work in building a fully mobile phone data driven agent-based simulation is the *Smartbay area* project which is presented in a paper by Pozdnoukhov (2015). They introduced to the MATSim environment anonymised CDRs recorded at the spatial resolution of the deployed mobile phone towers or antennas. Each individual's travel behaviour is modelled using a hidden semi-Markov model (HSMM), assuming that a user's (hidden) activity is influenced by temporal factors (i.e. day of the week and hours of the day), the type of the previous activity, and observed factors related to the current activity. After having inferred activity patterns, daily plans for the virtual population are generated and a MATSim traffic simulation scenario is run. It is important to note that the simulation output preserves user anonymisation since CDR trajectories are not involved in the micro-simulation. The agent-based simulation computes the route selection for each user based on the individual improvement of the utility function and not directly from CDR traces, avoiding the possibility to recover one specific mobile phone customer from the output of the simulation.

5 Supplementary Datasets: POIs and GPS

5.1 POIs

Although public transport smart card and mobile phone data have wide coverage across a city's population, the lack of purpose for the identified trips needs to be inferred in order to be useful for activity-based modelling approaches. Temporal features such as activity start time and duration may be enough to determine primary locations (i.e. work, home) for the majority of the population. However, for secondary locations in which individuals perform a range of activities from dining, shopping, and different types of leisure activities, enrichment of the spatial feature space is needed from the establishments surrounding a detected stop area from the mobility traces. Such information can be derived from Points of Interest datasets.

In terms of supplementing mobile phone data with POI information, Noulas and Mascolo (2013) compute the most popular activity within an area delimited by cellular antennas, using CDRs and Foursquare check ins and places data. They built a set of features to exploit semantic annotations of Foursquare data and test different supervised learning classifiers for the following classes: Arts and Entertainment, College and Education, Food, Work, Nightlife, Parks and Outdoors, Shops, Travel spots. Results show better accuracy for Nightlife, Arts and Entertainment, followed by Shops and Parks and Outdoors, whilst, the classifiers did not show good results for College and Education. These results encourage the use of POI as a complement on helping with secondary activities inference, in which, temporal features are not enough.

A similar work was done by Phithakkitnukoon, Horanont, Di Lorenzo, Shibasaki and Ratti (2010) for which they used POIs extracted from Yahoo maps API. They cluster four different types of activities (eating, shopping, entertainment, and recreational) using k-means algorithm, and estimate the most probable activity in each cluster using Bayes theorem. Using the most probable activity in a region and work place identification through temporal features (Calabrese et al., 2011), they infer a daily activity pattern for each of the users. As a result, they found a strong correlation in daily activity patterns within groups of people who share a common work area's profile.

It is important to note that both studies focus on identifying the most probable/popular activity in a delimited area, and not the most probable activity given a daily individual tour behaviour. For the latter, it is important to include individual and temporal variables in the activity recognition process such as the previously visited places and the starting time and duration of the activities, so that the region-based approach can be changed to an individual-based activity inference process. While approaches to do so with mobile phones and fare collection smart cards are limited and have been presented in previous chapter, in the next section we will present the extensive efforts done to annotate individual trip purposes using GPS traces.

5.2 GPS

GPS consist of a spatio-temporal high-resolution track (i.e. usually one reading per second, depending on configuration and application purpose). The importance of GPS technology or mobility applications is growing steadily: it is widely used for location-based smart phone application, to monitor public transport operations (GPS-enabled buses and taxis), and in the automobile industry as core technology for various applications such as route guidance but also is applied for toll collection and travel diary surveys.

However, in this chapter, we focus on applications that use GPS data to derive information with a semantic meaning (i.e. trip purpose). Depending on their methods, the following taxonomy according to Huang, Li and Yue (2010) is presented.

5.2.1 Distance-based

The basic idea is to assign the closest POIs to the raw trajectory's clusters. Commonly, a minimisation of the Euclidean distance is used between the location of nearby POIs and the identified GPS stay point (Bohte and Maat, 2009); Xie, Deng and Zhou (2009) present a more elaborated procedures for which they construct a Voronoi diagram using POIs as Voronoi sites and then select the POIs closest to the polyline geometry of the trajectory.

Although distance-based methods are easy to implement, they are only suitable for when there exists a high-accurate GPS trace and therefore are not ideally suited for application to mobile phone and smart card data.

5.2.2 Attractiveness-based

Different to the distance-based methods, attractiveness-based approached are designed to also include POI-related information to assign a particular POI given a GPS-point cluster. Huang et al. (2010) measured the spatio-temporal POI attractiveness based on statical factors (e.g. size of POI, popularity, and category) and a dynamic function for the attractiveness variation along the hours of the day. From this information they constructed an attractiveness prism, and selected the POI for which the mobility trajectory intersects the prism. Furletti, Cintia, Renso and Spinsanti (2013) present a similar approach, but apply a gravitational model to identify the most probable activity from a list of ranked POIs.

One of the advantages of the attractiveness-based approach is that it is also suitable for other types of mobility traces to a certain extent. The CDR + POI examples covered in the previous section (Noulas and Mascolo, 2013; Phithakkitnukoon et al., 2010) can be fitted in this category, for the reason that they seek the most attractive/popular activity in an area.

5.2.3 Probabilistic-based to handle uncertainty

Probabilistic-based methods are ideally suitable if certain GPS information is lacking or uncertain. Typical use cases are when GPS traces are only available for a vehicle a person used but do not cover walking legs leading to the actual destination or when surrounding high-rise buildings decrease the signal accuracy. Moreover, probabilistic approaches can handle uncertainty and capture interdependencies between explanatory variables, which make them a promising alternative to exploit different streams of information along with low resolution human mobility sensors. For these reasons, we present in the following an overview of such approaches in a dedicated section.

5.3 Probabilistic Graphical Models for activity inference

Probabilistic Graphical Models (PGM) as described in Koller and Friedman (2009) are graphical representations that efficiently encode and manipulate probability distributions over high-dimensional spaces. Variables are represented by nodes, and the probabilistic dependency (i.e. causality) is represented by edges that connect two variables. These graph models can be regarded as a compact or factor representation of a set of independences that hold in the specific distribution.

Learning The structure of a graphical model can be learned from data automatically or pre-defined by human knowledge. Graphical models usually contain hidden variables to be inferred (i.e. activity / trip purpose). The learning process of graphical models is to estimate the probabilistic dependency between different variables given the observed data. Expectation and Maximisation (EM) algorithms are commonly used methods.

Inference The inference process is to predict the status of hidden variables, given the values of observed variables and learned parameters. The inference algorithms can include deterministic approaches, such as variational methods, and stochastic algorithms like Gibbs Sampling.

In the following, we present representative examples for both generative and discriminative classification graphical models.

5.3.1 Generative models: Hidden Markov Models

Generative models explicitly model the dependencies between the observations and the class labels through the joint probability distribution $p(x, y)$. Hidden Markov Models (HMM) are generative models represented mainly by an entire observation sequence and a hidden high-level state sequence connected through a state transitions probability matrix, a matrix representing the conditional probability between states and observations, and an initial probability distribution.

Different extensions of HMMs have been developed to compute activity likelihoods and probabilistic estimates of the purpose behind the stop. For instance, Liao et al. (2007b) proposed a hierarchical HMM trained in an unsupervised manner using Expectation Maximisation (EM) to learn the parameters of the models, and Rao-Blackwellised particle filter for the inference task. However, the model still needed to be expanded in order to include information about time of day and day of the week. Such modification was included in Duong et al. (2005). They introduce a Switching Hidden Semi-Markov Model (S-HSMM) to exploit both the inherent hierarchical organization of the activities and their typical durations.

More recently, Yan, Chakraborty, Parent, Spaccapietra and Aberer (2011) developed a framework that enriches trajectories with any kind of semantic data provided by POI datasets based on a Hidden Markov Model. To define the initial probabilities they used the percentage of POI samples belonging to each category. Then they use information on transitions between regions (annotated with land use information) to construct the state transition. Finally, to infer the hidden states they maximised the likelihood of the HMM. One of their main contributions is the effective use of both land use information with POIs to infer the activity purpose. Additionally, Baratchi, Meratnia, Havinga, Skidmore and Toxopeus (2014) propose an extension of the hierarchical HSMM that captures spatio-temporal associations in the locational history in both stay-points and trips connecting the stops.

5.3.2 Discriminative models: Conditional Random Fields and Relational Markov Networks

In contrast with Generative models, Discriminative models avoid making independence assumptions among the observations, instead they model directly the discriminative boundary between the different class labels, namely, the model learns the conditional probability distribution $p(y|x)$. Thus, one of the advantages is that all sorts of rich overlapping features can be incorporate without violating any independence assumption. (Sutton and McCallum, 2006).

Conditional Random Fields (CRFs) are an example of discriminative models suitable for classification tasks with complex and overlapped attributes or observations. Liao et al. (2007a) show a holistic approach using hierarchical CRFs to extract places and activities from GPS traces. The main objective of their work is to segment a user's day into everyday activities such as *working*, *visiting*, or *travel* and to recognise and label significant places such as *workplace*, *friend's house*, or *bus stop*. To determine activities, the model relies on temporal features, such as duration or time of day, and geographic information such as locations of restaurants, stores, and bus stops. They use maximum pseudo-likelihood estimation to learn the parameters of the model, and belief propagation for the inference task.

Relational Markov Networks (RMN) are extensions of CRFs that provide a relational language for describing clique structures and enforcing parameter sharing at the template level. Liao et al. (2005) trained a RMN for labelling the following activities: *at home*, *at work*, *shopping*, *dinning out*, *visiting*, and *other*. They incorporate global features (e.g. number of home locations), temporal information (e.g. duration, time of day), and spatial information (POIs) in the clique templates. For the inference and learning task they develop a technique based on the Markov-Chain Monte-Carlo algorithm in a supervised manner. They showed that it is possible to learn the parameters of a complex model using less data by using priors extracted from other people's data.

6 Conclusion

6.1 Data driven Agent Based Modelling for Transport Planning

Traditional data sources for transport forecasting, i.e. household travel surveys, are of undeniable value. They not only cover detailed data on individual and household mobility patterns but also include relevant information on travel modes and purposes. Yet, stand-alone they are not able to exploit the full benefits of the agent-based transport modelling paradigm. Two main limitations can be identified. Firstly, they represent only a small sample of the population (normally around 1%). Secondly, they are usually only updated every five to ten years.[4]

Opportunistic human mobility sensors tackle these drawbacks and become a promising path to continue developing agent based models for transport planning. The tradeoff of using such opportunistic widely-collected information is its raw nature. An additional analytic effort has to be done to identify trips and trip purposes so that they can be integrated in the agent-based simulations. Thus, the key challenge is the development of robust algorithms that can extract daily individual schedules from sparse mobility traces. Specifically, as mentioned in Jiang et al. (2013), developing effective techniques to link the association rules of semantic land use and POI information of the diverse areas that individuals visit is an open challenge for estimating the activity types that individuals engage.

6.2 Model transferability from GPS to CDRs

One of the directions encountered, when using sparse CDRs to extract activities, is the adaptation of approaches that originally were developed to be applied with GPS data. For instance, Widhalm et al. (2015) adopted for CDRs the Relational Markov Networks used in Liao et al. (2005) originally for GPS traces; or Yuan et al. (2013) and Pozdnoukhov (2015) who adapted Conditional Random Fields for smart card data and Hidden Markov Models for CDRs, respectively. One of the reasons is that GPS traces have been the subject of a wider number of studies regarding activity identification. Hence, one important research question is if those models are suitable for lower resolution mobility traces, such as the ones provide by CDR and smart card data. Besides the discrepancy in the level of granularity of the traces, GPS-based studies usually have a controlled sample with activity labels. This is used to train the parameters of the model. For

[4] However, a few authorities have started with continuous surveys also using smart phones to lower the response burden and increase data quality, especially with regards to capturing activities that only last over a short duration.

CDRs it is likely that such training sample would not be easy to obtain. One of the possible options is to use information from travel surveys and design a feature-space in which CDRs and travel survey information can work along in a transfer-learning paradigm. Finally, another important issue to note, specifically, about inference models designed from GPS traces, is that usually they are trained and validated for small samples of the population (e.g. 4 persons in Liao et al. (2007a)). This certainly rises questions about their performance when scalable to the city size, in which a wider set of behavioural and mobility patterns from individuals might invalidate the results achieved by the models.

6.3 Unlocking the knowledge of different datasets

In order to be able to extract mobility and activity behaviours from human mobility sensors, and obtain similar quality to travel surveys, we need to combine the information across available datasets. For instance, if the interest is in finding the transport mode from CDRs, a viable option is to leverage public transit smart card data and available GPS traces from taxi services in a probabilistic trajectory matching approach. For the case of activity detection, the inclusion of POIs to help the inference models from human mobility sensors is a path that has not being explored yet.

As mentioned in Calabrese et al. (2015), there are some challenges in comparing different datasets, even if they are related. The main one being the different collection periods and different spatial units. For instance, census data is usually available at the block level, while mobile phone data relates to individual cell towers. However, we also see the main advantage of using different human mobility sensors and supplementary datasets such as travel diary data to complement underlying the importance of data fusion approaches.

From the literature review on the use of smart card and mobile phone data in transport modelling, we have seen mainly the use of census or household travel surveys as a mean to validate or to build expansion factors from the algorithms proposed (Alexander et al., 2015), or approaches that simply concatenate the features in a classification algorithm (Noulas and Mascolo, 2013; Chakirov and Erath, 2012). In the Big Data era, however, the aspiration will be to unlock the power of knowledge from multiple disparate, but potentially connected datasets (Zheng, 2015). Therefore, we expect the most promising approaches will be applications will stem from data science domains such as machine learning and data mining.

Cross-Domain Data Fusion Zheng (2015) presents a good survey that analyses, classifies and exemplifies methodologies for Cross-Domain Data Fusion. From a transport planning perspective, the promising methods for activity and transport inference are the probabilistic semantic-meaning data fusion methods, namely, the Probabilistic Graphical Models reviewed in the previous chapter. As mentioned before, these models are able to capture the dependencies and correlations between the features in order to produce better estimates. They also constitute an essential tool to reason coherently from limited and noisy observations (Koller and Friedman, 2009).

6.4 Data Privacy

While travel diary data and census information are generally surveyed by government agencies with consent of the relevant data protection authority, data privacy is an important matter due to the pervasiveness and level of detail how both smart card and mobile phone data document an individual's travel patterns. Even if CDRs are anonymised (smart card data normally is not linked to a particular person by default), individual mobility patterns are pseudo-identifiers. For instance, de Montjoye, Hidalgo, Verleysen and Blondel (2013) showed that even with the spatial resolution given by the mobile phone antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. Efforts have been taken to obfuscate the location in a way not to be able to re-identify a user and still being able to extract useful mobility patterns. These algorithmic efforts to preserve privacy are currently lead by the emerging *Differential Privacy*. Generally, we expect that addressing those concerns in convincing manner will be crucial towards the development and practical adoption of data-driven, agent-based simulation for transport planning.

Differential Privacy Differential Privacy (DP) is a mathematical requirement on the results of interaction with data (Mir et al., 2013). By adding controlled noise, DP formalises the idea that results of a query should be almost the same whether or not an individual is in the database. DP hides the participation of a user in a database by choosing a budget parameter that represents the trade-off between the level of privacy and precision. Andrés et al. (2013) extended the concept of DP for the protection of location data. Although DP has been proved to be effective with certain location-based services (Andrés et al., 2013), and with aggregate location information (Mir et al., 2013), when applied to individual mobility traces it seems that a trade-off between privacy and precision has not been able to obtain with state-of-the-art techniques (Primault et al., 2014; Hu et al., 2015).

New Models of Data Ownership Location obfuscation might not be the viable mechanism for sharing individual mobility traces. Other suggestions, aside from the algorithmic perspective, are related to changing the current data ownership paradigm. A popular idea is giving people ownership of their data (Pentland, 2009). With the creation of *data vaults* (Mun et al., 2010) or by means of a *data trust* (Lawrence, 2016), each individual would have the right to dispose or distribute their personal information. However, a short-term implementation doesn't seem plausible since there exist conflict of interests with big data-driven companies. Still, a transparent use of personal information is certainly a relevant way to facilitate access for researchers to both develop better anonymisation methodologies and showcase the societal benefits of using anonymised mobility data.

7 Summary and Research Agenda

7.1 Summary

In the first part of the paper, we provide a primer on the latest advances in transport demand modelling. We acknowledge that activity-based models and agent-based simulations tie in well with new Big Data sources documenting human mobility since both directly stem from the concept of individual travel patterns rather then aggregate traffic flows. In order to fully exploit the capabilities of agent-based simulations, there is tremendous potential to not only use conventional data inputs (e.g. travel surveys, population census), but also include new opportunistically collected mobility traces, namely, public transit smart card and mobile phone data, which can document travel behaviour at an unprecedented scale and level of detail. However, an additional analytic effort has to be done to identify trips and trip purposes so that they can be integrated into activity-based travel demand framework and to be used to its full potential in agent-based simulations.

In the second part of the paper, we present a literature review on the methodologies needed to extract mobility behaviour from such Big Data sources. From trip identification to activity inference, and their application for transport demand models, we review the efforts in a step-by-step manner both for the public transport smart card and mobile phone data. We also cover the relevance of other datasets such as POIs to infer trip purposes and also document how GPS-based data collection and Probabilistic Graphical Models can be included for model refinement.

Finally, we discuss the findings of the literature review and also identify a set of future challenges, in particular with regards to data privacy implications.

7.2 Research Agenda

The biggest potential for future research we identify on the application of big data sources (e.g. smart card and mobile phone data) in combination with traditional data sources (e.g. household travel surveys) to inform activity-based models and agent-based simulations for transport planning. To this end, the following challenges need to be added to the research agenda.

1. Further exploration of probabilistic approaches that can handle the uncertainty of Big data mobility traces in the modelling process and showcasing the relevance of such approaches for instance through validation with independent data sets such as loop-detector data.

2. Integration of different available datasets in a data-fusion scheme. For instance, smart card data and CDRs for mode inference or POI datasets and semantic information in social networks to inform in the activity inference process of CDR traces.

3. Exploration on Transfer Learning approaches to cover up the lack of training samples when using opportunistic collected data sources.

4. Seek new ways to validate the results for every step of the data mining pipeline.

5. Further exploration of preprocessing techniques for sparse and noisy mobility traces. For instance, demonstrate the effectiveness of non-linear Kalman filters, or more complex probabilistic filters (e.g. Gaussian processes).

6. Measure and guarantee privacy in the output of agent-based simulations.

In any case, we also acknowledge that sophisticated modelling knowledge has developed in the domain of transport planning and therefore we strongly advise that still domain expert knowledge should build the fundament when applying data driven approaches in transport planning. These new challenges call for a multidisciplinary collaboration between transport modellers and data scientists

8 Acknowledgements

We would like to thank Seungjae Lee for organising the Seoul Metropolitan Forum 2016 at the University of Seoul where an earlier version of this paper was presented. This research has been conducted at the Singapore-ETH Centre for Global Environmental Sustainability (SEC), co-funded by the Singapore National Research Foundation (NRF) and ETH Zurich.“

9 References

Adnan, M., F. C. Pereira, C. M. Lima Azevedo, K. Basak, M. Lovric, S. Raveau, Y. Zhu, J. Ferreira, C. Zegras and M. E. Ben-Akiva (2016) SimMobility: A Multi-scale Integrated Agent-Based Simulation Platform, paper presented at the *Transportation Research Board 95th Annual Meeting*.

Alexander, L., S. Jiang, M. Murga and M. C. González (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data, *Transportation Research Part C: Emerging Technologies*, **58**, 240–250.

Andrés, M. E., N. E. Bordenabe, K. Chatzikokolakis and C. Palamidessi (2013) Geo-indistinguishability: Differential privacy for location-based systems, paper presented at the *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 901–914.

Balmer, M., K. Axhausen and K. Nagel (2006) Agent-based demand-modeling framework for large-scale microsimulations, *Transportation Research Record: Journal of the Transportation Research Board*, (1985) 125–134.

Balmer, M., M. Rieser, K. Meister, D. Charypar, N. Lefebvre, K. Nagel and K. Axhausen (2009) MATSim-T: Architecture and simulation times, *Multi-agent systems for traffic and transportation engineering*, 57–78.

Baratchi, M., N. Meratnia, P. J. M. Havinga, A. K. Skidmore and B. A. K. G. Toxopeus (2014) A Hierarchical Hidden semi-Markov Model for Modeling Mobility Data, paper presented at the *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’14, 401–412, New York, NY, USA, ISBN 978-1-4503-2968-2.

Barry, J., R. Newhouser, A. Rahbee and S. Sayeda (2002) Origin and Destination Estimation in New York City with Automated Fare System Data, *Transportation Research Record: Journal of the Transportation Research Board*, **1817**, 183–187, January 2002, ISSN 0361-1981.

Bohte, W. and K. Maat (2009) Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands, *Transportation Research Part C: Emerging Technologies*, **17** (3) 285–297.

Bouman, P. (2012) Recognizing demand patterns from smart card data for agent-based micro-simulation of public transport, Ph.D. Thesis, Department of Decision and Information Sciences, Erasmus University Rotterdam, The Netherlands.

Bouman, P., E. Van der Hurk, L. Kroon, T. Li and P. Vervest (2013) Detecting activity patterns from smart card Data, paper presented at the *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence, Delft, The Netherlands, November 7-8, 2013*.

Caceres, N., J. P. Wideberg and F. G. Benitez (2007) Deriving origin destination data from a mobile phone network, *IET Intelligent Transport Systems*, **1** (1) 15–26, March 2007, ISSN 1751-956X.

Calabrese, F., L. Ferrari and V. D. Blondel (2015) Urban sensing using mobile phone network data: a survey of research, *ACM Computing Surveys (CSUR)*, **47** (2) 25.

Calabrese, F., G. D. Lorenzo, L. Liu and C. Ratti (2011) Estimating Origin-Destination Flows Using Mobile Phone Location Data, *IEEE Pervasive Computing*, **10** (4) 36–44, April 2011, ISSN 1536-1268.

Castiglione, J., M. Bradley and J. Gliebe (2015) *Activity-based travel demand models: a primer*, Transportation Research Board, Washington, DC, ISBN 978-0-309-27399-2.

Chakirov, A. and A. Erath (2012) Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport, Toronto, June 2012.

de Dios Ortuzar, J. and L. G. Willumsen (2011) *Modelling transport*, John Wiley & Sons.

de Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen and V. D. Blondel (2013) Unique in the crowd: The privacy bounds of human mobility, *Scientific reports*, **3**.

Devillaine, F., M. Munizaga and M. Trépanier (2012) Detection of Activities of Public Transport Users by Analyzing Smart Card Data, *Transportation Research Record: Journal of the Transportation Research Board*, **2276**, 48–55, December 2012, ISSN 0361-1981.

Duong, T. V., H. H. Bui, D. Q. Phung and S. Venkatesh (2005) Activity recognition and abnormality detection with the switching hidden semi-Markov model, paper presented at the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, vol. 1, 838–845 vol. 1, June 2005.

Eagle, N. and A. S. Pentland (2009) Eigenbehaviors: identifying structure in routine, *Behavioral Ecology and Sociobiology*, **63** (7) 1057–1066, April 2009, ISSN 0340-5443, 1432-0762.

Eagle, N. and A. (Sandy) Pentland (2006) Reality Mining: Sensing Complex Social Systems, *Personal Ubiquitous Comput.*, **10** (4) 255–268, March 2006, ISSN 1617-4909.

Ficek, M. and L. Kencl (2012) Inter-call mobility model: A spatio-temporal refinement of call data records using a Gaussian mixture model, paper presented at the *INFOCOM, 2012 Proceedings IEEE*, 469–477.

Fourie, P. J. (2014) Reconstructing bus vehicle trajectories from transit smart-card data, *Working paper*, **986**.

Fourie, P. J., A. Erath, S. A. Ordóñez Medina, A. Chakirov and K. W. Axhausen (2016) Using smartcard data for agent-based transport simulation: the case of Singapore, in J.-D. Schmoecker and F. Kurauchi (eds.) *Public Transport Planning with Smart Card Data*, Taylor & Francis.

Furletti, B., P. Cintia, C. Renso and L. Spinsanti (2013) Inferring human activities from gps tracks, paper presented at the *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 5.

Goulias, K. G., C. R. Bhat, R. M. Pendyala, Y. Chen, R. Paleti, K. C. Konduri, T. Lei, D. Tang, S. Y. Youn, G. Huang and others (2012) Simulator of activities, greenhouse emissions, networks, and travel (SimAGENT) in Southern California, paper presented at the *91st annual meeting of the Transportation Research Board, Washington, DC.*

Horn, C., S. Klampfl, M. Cik and T. Reiter (2014) Detecting Outliers in Cell Phone Data, *Transportation Research Record: Journal of the Transportation Research Board*, **2405**, 49–56, July 2014, ISSN 0361-1981.

Horni, A., K. Nagel and K. Axhausen (2016) *The Multi-Agent Transport Simulation MATSim*, Ubiquity Press, June 2016, ISBN 978-1-909188-75-4.

Hu, X., M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, H. Guan and J. Zeng (2015) Differential privacy in telco big data platform, *Proceedings of the VLDB Endowment*, **8** (12) 1692–1703.

Huang, L., Q. Li and Y. Yue (2010) Activity identification from GPS trajectories using spatial temporal POIs' attractiveness, paper presented at the *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, 27–30.

International Transport Forum (2015) Big Data and Transport - Understanding and assessing options, *Technical Report*, OECD, Paris.

Iqbal, M. S., C. F. Choudhury, P. Wang and M. C. González (2014) Development of ori- gin–destination matrices using mobile phone call data, *Transportation Research Part C: Emerging Technologies*, **40**, 63–74, March 2014, ISSN 0968-090X.

Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland and A. Varshavsky (2011) Identifying Important Places in People's Lives from Cellular Network Data, in K. Lyons, J. Hightower and E. M. Huang (eds.) *Pervasive Computing*, no. 6696 in Lecture Notes in Computer Science, 133–151, Springer Berlin Heidelberg, June 2011, ISBN 978-3-642-21725-8 978-3-642-21726-5. DOI: 10.1007/978-3-642-21726-5_9.

Jiang, S., J. Ferreira Jr and M. C. González (2015) Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore, paper presented at the *Int. Workshop on Urban Computing*.

Jiang, S., G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli and M. C. González (2013) A review of urban computing for mobile phone traces: current methods, challenges and opportunities, paper presented at the *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, 2.

Koller, D. and N. Friedman (2009) *Probabilistic graphical models: principles and techniques*, MIT press.

Lawrence, N. (2016) Data trusts could allay our privacy fears, June 2016, `http://www.theguardian.com/media-network/2016/jun/03/data-trusts-privacy-fears-feudalism-democracy`.

Liao, L., D. Fox and H. Kautz (2005) Location-based Activity Recognition Using Relational Markov Networks, paper presented at the *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, 773–778, San Francisco, CA, USA.

Liao, L., D. Fox and H. Kautz (2007a) Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields, *The International Journal of Robotics Research*, **26** (1) 119–134, January 2007, ISSN 0278-3649, 1741-3176.

Liao, L., D. J. Patterson, D. Fox and H. Kautz (2007b) Learning and inferring transportation routines, *Artificial Intelligence*, **171** (5–6) 311–331, April 2007, ISSN 0004-3702.

Mir, D. J., S. Isaacman, R. Caceres, M. Martonosi and R. N. Wright (2013) Dp-where: Differentially private modeling of human mobility, paper presented at the *Big Data, 2013 IEEE International Conference on*, 580–588.

Mun, M., S. Hao, N. Mishra, K. Shilton, J. Burke, D. Estrin, M. Hansen and R. Govindan (2010) Personal data vaults: a locus of control for personal data streams, paper presented at the *Proceedings of the 6th International COnference*, 17.

Munizaga, M. A. and C. Palma (2012) Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile, *Transportation Research Part C: Emerging Technologies*, **24**, 9–18, October 2012, ISSN 0968-090X.

Noulas, A. and C. Mascolo (2013) Exploiting foursquare and cellular data to infer user activity in urban environments, paper presented at the *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, vol. 1, 167–176.

Pentland, A. (2009) Reality Mining of Mobile Communications: Toward A New Deal On Data, in *Social Computing and Behavioral Modeling*, 1–1, Springer US, ISBN 978-1-4419-0055-5 978-1-4419-0056-2. DOI: 10.1007/978-1-4419-0056-2_1.

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki and C. Ratti (2010) Activity-aware map: Identifying human daily activity pattern using mobile phone data, in *Human Behavior Understanding*, 14–25, Springer.

Pozdnoukhov, A. (2015) Activity Based Travel Demand Modelling with Cellular Data, October 2015, `http://www.ucconnect.berkeley.edu/workforce-development/ucconnect-transportation-planning-workshops/role-big-data-transportation`.

Primault, V., S. B. Mokhtar, C. Lauradoux and L. Brunie (2014) Differentially private location privacy in practice, *arXiv preprint arXiv:1410.7744*.

Rasouli, S. and H. Timmermans (2013) Activity-based models of travel demand: promises, progress and prospects, *International Journal of Urban Sciences*, **0** (0) 1–30, ISSN 1226-5934.

Sarlas, G. and K. W. Axhausen (2015) Localized Speed Prediction with the use of Spatial Simultaneous Autoregressive Models, paper presented at the *Transportation Research Board 94th Annual Meeting*.

Schlaich, J., T. Otterstätter and M. Friedrich (2010) Generating trajectories from mobile phone data, paper presented at the *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*.

Schneider, C. M., V. Belik, T. Couronné, Z. Smoreda and M. C. González (2013) Unravelling daily human mobility motifs, *Journal of The Royal Society Interface*, **10** (84) 20130246, July 2013, ISSN 1742-5689, 1742-5662.

Seaborn, C., J. Attanucci and N. Wilson (2009) Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data, *Transportation Research Record: Journal of the Transportation Research Board*, **2121**, 55–62, December 2009, ISSN 0361-1981.

Smith, L., R. J. Beckman, D. Anson, K. Nagel and M. E. Williams (1995) TRANSIMS: TRansportation ANalysis and SIMulation System, paper presented at the *5th TRB National Transportation Planning Methods Applications Conference*, Seattle, April 1995.

Sun, L., D.-H. Lee, A. Erath and X. Huang (2012) Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System, paper presented at the *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, 142–148, New York, NY, USA, ISBN 978-1-4503-1542-5.

Sun, L., A. Tirachini, K. W. Axhausen, A. Erath and D.-H. Lee (2013) Models of Bus Boarding/Alighting Dynamics and Dwell Time Variability.

Sutton, C. and A. McCallum (2006) An introduction to conditional random fields for relational learning, *Introduction to statistical relational learning*, 93–128.

Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff and M. C. González (2015) The path most traveled: Travel demand estimation using big data resources, *Transportation Research Part C: Emerging Technologies*, **58**, 162–177.

Trépanier, M., N. Tranchant and R. Chapleau (2007) Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System, *Journal of Intelligent Transportation Systems*, **11** (1) 1–14, April 2007, ISSN 1547-2450.

Wang, H., F. Calabrese, G. D. Lorenzo and C. Ratti (2010) Transportation mode inference from anonymized and aggregated mobile phone call detail records, paper presented at the *2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 318–323, September 2010.

Wang, P., T. Hunter, A. M. Bayen, K. Schechtner and M. C. González (2012) Understanding Road Usage Patterns in Urban Areas, *Scientific Reports*, **2**, December 2012, ISSN 2045-2322.

Widhalm, P., Y. Yang, M. Ulm, S. Athavale and M. C. González (2015) Discovering urban activity patterns in cell phone data, *Transportation*, **42** (4) 597–623.

Xie, K., K. Deng and X. Zhou (2009) From trajectories to activities: a spatio-temporal join approach, paper presented at the *Proceedings of the 2009 International Workshop on Location Based Social Networks*, 25–32.

Yan, Z., D. Chakraborty, C. Parent, S. Spaccapietra and K. Aberer (2011) SeMiTri: a framework for semantic annotation of heterogeneous trajectories, paper presented at the *Proceedings of the 14th international conference on extending database technology*, 259–270.

Yang, Y., P. Widhalm, S. Athavale and M. C. González (2015) Mobility Sequence Extraction and Labeling Using Sparse Cell Phone Data.

Yuan, N. J., Y. Wang, F. Zhang, X. Xie and G. Sun (2013) Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach, paper presented at the *2013 IEEE 13th International Conference on Data Mining (ICDM)*, 877–886, December 2013.

Zhang, Y., X. Qin, S. Dong and B. Ran (2010) Daily OD matrix estimation using cellular probe data, paper presented at the *89th Annual Meeting Transportation Research Board*.

Zhao, J., A. Rahbee and N. H. M. Wilson (2007) Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems, *Computer-Aided Civil and Infrastructure Engineering*, **22** (5) 376–387, July 2007, ISSN 1467-8667.

Zheng, Y. (2015) Methodologies for cross-domain data fusion: an overview, *Big Data, IEEE Transactions on*, **1** (1) 16–34.

Zilske, M. and K. Nagel (2014) Studying the Accuracy of Demand Generation from Mobile Phone Trajectories with Synthetic Data, *Procedia Computer Science*, **32**, 802–807, ISSN 1877-0509.

Zilske, M. and K. Nagel (2015) A Simulation-based Approach for Constructing All-day Travel Chains from Mobile Phone Data, *Procedia Computer Science*, **52**, 468–475, ISSN 1877-0509.