Opponent's report

Opposition of master thesis "Generating Traveller Location Data from a Microsimulation Model" by Johannes Schlagheck

Daniel Erlandsson

LINKÖPINGS UNIVERSITET

Thesis report summary

Limited traffic data impose a challenge to traffic planners when it comes to monitoring traffic, both live and at time spans that has already past. Gathering traffic data can be both time consuming, imprecise and costly. The presentation of data may also have a long delay depending on the method of collection.

This thesis evaluates the possibility of analyzing motorized traffic, real time or in the past, by using Call detail records (CDR) generated by mobile devices when used in a cellular network, e.g. GSM. But since there are no CDR data publicly available a set of CDR must first be created. The author uses data provided by Trafikverket from fixed sensors along a stretch of highway in Stockholm. This data acted as demand data for the traffic simulation. The thesis aims at being reproducible, especially when it comes to generation of synthetic CDR and focuses on sharing this knowledge to ease traffic research.

CDRs' are created by combining knowledge of mobile cellular networks and traffic simulation using a traffic simulation software called Aimsun. This script translates traffic data (trajectories) to connected mobile devices for each cell (or mobile device) in the given geographical area of research. Using traffic data from five Tuesday mornings, the author states that it contains regular morning peak traffic for the given area. CDRs' are generated not only for the highway, but also for residential streets surrounding the given highway stretch. Information about positioning of cell towers and cell identification where gathered from an open source called OpenCellID.

By assigning a portion of the vehicles to a mobile telephone cell (e.g. only motorist in x portion of the cars carries a mobile device connected to telephone operator X), it can be examined how different measurements in the mobile network can be interpreted as the current traffic status. Three scenarios where put up; (a) *Original scenario*, (b) *Free flow scenario* and (c) *Congestion scenario* where (b) and (c) are decreased respectively decreased by 20 % compared to (a). Results are also examined in the type of road dimension, i.e. *Highway*, *Highway and ramps*, *Residential streets* and *Total network*.

Measurements for the mobile network presented in the report are (a) *Total system load*, (b) *Average cell* size and (c) *Cell dwell time*. It is concluded after several analyses that (a) *Total system load* only works for a part of the data set and is therefore regarded as unreliable. However, with measurement (b) *Average cell size* there was concluded that general status of traffic (especially on highway) could be determined. Faster moving traffic connects to bigger cells than slower moving traffic. Furthermore, measurement (c) *Cell dwell time* could also be used as a measurement when slow moving traffic spends more time in the same cell compared to fast moving traffic.

A general OD estimation based on CDR was also examined but concluded to be unsuccessful with this type of data.

The author states that algorithms used are in a simple state and that microscopic models are limited in their size. This implies that algorithms may be developed more but are limited to the size of the models. Growing models will quickly increase in complexity.

For further work, the following areas are identified; (a) Different shapes of cell networks, (b) Different call likelihood parameters, (c) including non motorized traffic, (d) including noise CDR and (e) adding the dimension of public transport (e.g. busses).

Assessment of the report

Overall well written report on an interesting subject. The technical advantages by using CDRs' compared to other means of collecting data are clearly apparent since it might cut costs and provide real time data. Among the drawbacks are the imprecise location data and challenges with privacy and access to real life data. However, I find it to be a relevant and meaningful field of research.

There might be a coding or conversion error. The received PDF file contained some unexpected characters throughout the report. E.g. chapter *2.11*, second sentence.

Title and Abstract

Having read the thesis, the title corresponds well to the rest of the report. The Swedish sub title may not be needed since the report to the full extent is written in English.
The Abstract is well balanced and gives the reader a good understanding of the work.

Introduction

In general a good introduction that gives the reader a good picture of the problem and how the author has solved it. Some sections are, however, quite dense and contains a lot of information. As an example, the descriptions (especially regarding cellular networks concerning databases, estimation of cell locations etc.) in *1.3 Methodology* might belong to the Theoretical background chapter instead. This part might also not be common knowledge and may be needed to use together with references.

A bit vague when it comes to *1.4 Limitations*; can the defined delimitation for network structure and radio resource management be specified more than "as advanced as possible"? Since selecting the correct cells for each trajectory is a crucial step in this work, it might be better off with a better delimitation.

Theoretical background

The theoretical background is for most parts easy to understand and is relevant for the experimental procedure presented later in the report. Stating where the reader can find more to read on a specific subject is helpful. It covers the basics and need to know before describing the experimental procedure.

Figure 2.2 is hard to understand and is in need of further explanation. It might be that me (the opponent) is not familiar with this graph, but I don't succeed to read any of those abbreviations (AS, AD, AR, AB, etc.).

Great explanation in chapter *2.2.2 Signalling data in cellular networks* (of CDR, CCU and NMR), but how many of the users are available in the CDR? 50 %? 20 %?

Experimental procedure

Overall this is a well written chapter that is easy to follow for me as a reader. There are some assumptions that are not fully described which can influence the end result, e.g. traffic generated by residential areas, highway capacity and simulation model parameter settings.

Great explanation of the scripts used. By referring to a specific line in the code it makes it easier to for the next person reproducing the experiment.

Since the validation based on a real data set is not performed it might be quite a substantial room for error, even though that might not affect the outcome of this particular experiment.

Figure 3.2; it would be helpful to mark out and explain the centeroid in the figure description since this can look different in other applications or as an explanation to a reader who has not modelled traffic with Aimsun before.

In chapter *3.1.3 Demand data input*, you need to clarify which dates are used as an input, (wrongly stated 26th of April to the 14th of May).

In chapter *3.3.4 Implementation of the mobile connectivity model* it is stated that a realistic Cell dwell time is 3 to 7 minutes. Why? Driving time?

Results

The Results chapter is also well written. I would however like to see some more structure concerning the description of scenarios. Instead of including all scenarios in body text, try to break them out one by one to make it clearer that (1) there are different scenarios and (2) the differences between the scenarios.

The different measurements are presented well and connect to the relevant sections in the theoretical background. Both positive results (e.g. Average cell size and Cell dwell time) and negative results (e.g. OD estimation) are presented. This approach makes it easier for the reader to understand the opportunities as well as the challenges using CDRs' for this application.

There are some concerns regarding numbers that doesn't add up or is not explained well enough. E.g. *Table 4.4* which I am making a comment on in the questions section later in this document. It had also been interesting to see if and how these different measurements can be combined and thereby create a small model that might identify different scenarios using CDRs' as input.

Conclusion

The author points out some very good points in this section, proving that he knows many of the drawbacks using synthetic CDRs'. After reading the chapters up until the conclusion, most of the conclusions comes natural in the same way for me as a reader that the conclusions that the author has made.

The chapter *5.2 Recapitulation* connects most other parts of the report and shows a red line through most of the report.

The future outlook has the potential to be more clearly stated in terms of better defined areas of suggested upcoming research. Most of them involves increased complexity by adding noise or randomness into the model.

Fulfilling of aims

In the report aim, five aims where stated. These aims have been more or less fulfilled as described below;

How can CDR data, collected in an urban region, give information about the current traffic state?

It is concluded that CDR data can be used, or at least some of the measures, to give information about current traffic state. The lack of noise and exact positioning of mobile devices in the experimental phase means that it might be just a theoretical finding, but the author shows that some measures gave credible results.

Is it possible to filter out the travelling users from the data and what are the characteristics of their records?

This is a partly fulfilled aim. One can draw the conclusion that if a user is for example not changing cells or using a cell with a small size, that this user is stationary. However, introducing real life noise could have implications that lead to issues when trying to separate stationary users (changing cells) from moving motorists at residential streets. The experimental procedure also lack synthetic CDRs' from stationary users.

To what grade is it possible to distinguish a fast from a slow traveller?

This aim is to be considered as fulfilled. By using measurements like cell size and cell dwell time, it is possible to separate a slow moving user from a fast moving user.

How is a changing demand of a traffic system indicated in a relating CDR?

By describing different measurements and the result of the simulation of CDRs', it is shown that for different measurements different indications can be used. In some cases, a specific measurement couldn't deliver any promising results, but in other cases and with other measurement real time status of traffic could be examined.

In how far is it possible to distinguish travellers' origins and destination or to identify specific route choices from CDR collected in a suburban scale?

The answer, regarding the data and type of geographical area, is in this case none. It has been concluded that OD estimation cannot be done using the same prerequisites as this experiment has done.

Questions for the oral opposition

- Can you describe the overall aim for your thesis work? Is that aim fulfilled?

- Since CDRs' are created using data from highway traffic sensors there is no noise from e.g. pedestrians/cyclists and stationary users. Furthermore, cells might not be selected as optimal as described in this report. Fast moving mobile users may jump from one cell to another, not in a straight line, but just happening to pick up a strong signal from somewhere else. If an approach like this is to be used in real life, what are your thoughts on elimination such noise?

- Can you see a risk in that you have created CDR records that are too far away from reality? Would this have an effect on your results?

- Are your findings applicable in other fields than traffic planning?

- In chapter *2.1.2 Models for microscopic simulation* you describe the **gap acceptance model** and the **route choice model**. Are those models used in the Experimental procedure? In that case, how?

- Would the model or algorithms have benefitted from collecting highway sensor data in real life free flow instead of using the modified sensor data given for the Tuesday morning peak in order to simulate free flow?

- Do the model take into consideration that the way out from this geographical area (represented by Aimsun) might be a restrictor of traffic flow? I.e. the southbound highway decreases from 4 lanes to 2 lanes just some kilometres south of this specific area. Does that have any implications regarding the morning peak traffic for your model?

- In chapter *2.3.1 State of the art in traffic data collection* you mention Bluetooth and Wi-Fi devices. At a speed of 100 km/h this technique has an average detection rate of about 80 %. How does that compare to CDR used in the experimental face? Can you compare the main advantages and drawbacks collecting data whit these sensors compared to CDR?

- In the Result chapter, a couple of individual measurements on the (GSM) network are described. It appears that none of these measurements individually can give a good estimation of the traffic state; is that correct? Have you ideas of a methodology or approach that might combine these measures to a better estimation of the current traffic situation?

- In real life, all CDRs' will be accumulated. Would you be able to see congestion on the Residential streets when the highway has a free flow?

- In *Table 4.4*, why is there such a big difference between *Everything pattern* and the other three patterns combined? E.g. first row; Everything (2 688) and the other three (77 + 1 084 + 484 = 1 645).