MODELLING EMISSION OF POLLUTANTS FROM TRANSPORTATION USING MOBILE SENSING DATA

ADVANCING MODELLING OF STREET LEVEL POLLUTION AND CLIMATE FORCING GAS EMISSIONS

ANDERS LEHMANN

PH.D. DISSERTATION
August 2017
Advisor: Niels Olof Bouvin & Allan Gross
A developed country is not a place where the poor have cars. But where the rich use public transport.

- paraphrased from Enrique Penalosa, former Mayor of Bogotà, Colombia

I dedicate this dissertation to my family.
The advent and the proliferation of the smartphone has promised new possibilities for researchers to gain knowledge about the habits and behaviour of people, as the ubiquitous smartphone with an array of sensors is capable of deliver a wealth of information.

This dissertation addresses methods to use data acquired from smartphones to improve transportation related air quality models and models for climate gas emission from transportation. These models can be used for planning of transportation networks, monitoring of air quality, and automate transport related green accounting.

More accurate transportation models can be obtained by using observed travel routes, acquired from smartphone data, rather than indirectly computed routes, as input to a model of route choice in a transportation network. Smartphone data can also be used to gain detailed knowledge of the driving style of individual drivers, by deducing driving modes from accelerometer data.

The work presented in this thesis spans different and diverse research fields. Transportation models are a subfield of econometrics, air quality modelling is a subfield of atmospheric chemistry, and driving mode detection and efficient database implementations are a subfield of computer science. I have worked to bring these diverse research fields together to solve the challenge of improving modelling of transportation related air quality emissions as well as modelling of transportation related climate gas emissions.

The main scientific contributions of the dissertation are:

- Algorithm for origin destination demand matrix creation from smartphone data.
- The development of a novel map matching algorithm suitable for a database.
- Using user experienced routes as a seed for a transport model.
- Driving mode detection from smartphone accelerometer data.
- A performant database implementation of Restricted Stochastic User Equilibrium transport model.
- An investigation of the accuracy of Global Positioning System in a stationary smartphone.
RESUMÉ

Fremkomsten og udbredelsen af smartphonen har lovet ny muligheder til forskere for at få ny viden om folks vaner og adfærd, da den alledennæværende smartphone med de mange sensorer er i stand til levere et væld af informationer.

Denne afhandling gennemgår metoder til at anvende data modtaget fra smartphones til at forbedre transport relaterede luft kvalitets modeller og modeller for udledning af klima gasser fra transport. Disse modeller kan anvendes til planlæg af transport anlæg, overvågning af luft kvalitet samt automatisering af transport relaterede grønne regnskaber.

Præcise transport modeller kan opnås ved at benytte faktisk brugte rejse ruter, ved at udnytte data fra smartphones, fremfor indirekte beregnede ruter, som input til en rute valgs model i et transport netværk. Data fra smartphones kan også anvendes til få detaljeret viden om individuel kørsels stil, ved udlede kørsel mønstre fra accelerometer data.


De væsentligste videskablige bidrag er:

- Algoritme til generering af transport behovs matricer udfra data fra smartphones.
- Udvikling af en ny database egnet metode til tilpasse positioner til et vejkort.
- Brug af faktisk brugte ruter som start punkt til en trafik model.
- Detektering af kørsel mønstre fra smartphone data.
- En effektiv implementering af Restricted Stochastic User Equilibrium (RSUE) trafik model i et database system.
- En undersøgelse af nøjagtigheden af Global Positioning System (GPS) i en stationær smartphone.
PUBLICATIONS

Papers published during the Ph.D study:


Unpublished papers included in the dissertation:


Previously published papers unrelated to this study:


ACKNOWLEDGMENTS

I would like to thank, my supervisors:
Niels Olof Bouvin og Allan Gross, for their patience and advice.
Anders Møller: chairman of the PhD commission of the Grade School of Computer Science at Aarhus university.
The management of the EcoSense project Kaj Grønbæk, Henrik Blunck, and Michael Goodsite.
Mikael Bergholz Knudsen from Aarhus School of Engineering, Aarhus Universitet, as well as:
Bülent Çatay, Sabanci University, Istanbul for generously hosting a research visit at Sabanchi University.
## CONTENTS

1 **MOTIVATION AND STATE-OF-THE-ART**  
   1.1 Introduction 3  
      1.1.1 Contributions 4  
      1.1.2 Introduction to papers 5  
      1.1.3 Structure of the dissertation 6  

2 **RELATED WORK** 9  
   2.1 Introduction 9  
   2.2 Air-quality models and vehicle emission models 9  
   2.3 Transport models 11  
      2.3.1 Origin destination demand matrix generation 12  
   2.4 Positioning 13  
      2.4.1 Map-matching 14  

3 **METHODODOLOGY** 17  
   3.1 Introduction 17  
   3.2 Hypotheses 17  
   3.3 Experiments 17  
      3.3.1 GPS accuracy 18  
      3.3.2 Driving modes 18  
      3.3.3 Experiments with EcoSense data 18  

II **EMISSION MODELS** 21  
4 **THE POLITICS OF CLIMATE GAS EMISSION** 23  
   4.1 Climate change policy development 23  
      4.1.1 Climate change organisations 23  
   4.1.2 Climate change mitigation 24  
   4.1.3 Climate change adaptation 24  
   4.2 Conclusion 24  

5 **MODELLING OF EMISSION OF CLIMATE GASSES FROM TRANSPORTATION** 27  
   5.1 Introduction 27  
   5.2 CO₂ emission modelling 27  
      5.2.1 IPCC methodology 27  
      5.2.2 Simple emission model 32  
      5.2.3 Driving mode detection 34  
      5.2.4 Modelling emissions from electric vehicles 35
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Air pollution dispersion Modelling</td>
<td>36</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Eulerian method</td>
<td>37</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Lagrange</td>
<td>38</td>
</tr>
<tr>
<td>5.3.3</td>
<td>DASK-GAS</td>
<td>38</td>
</tr>
<tr>
<td>5.3.4</td>
<td>OSPM</td>
<td>39</td>
</tr>
<tr>
<td>5.4</td>
<td>conclusion</td>
<td>40</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>43</td>
</tr>
<tr>
<td>6.2</td>
<td>Route choice models</td>
<td>43</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Deterministic User Equilibrium</td>
<td>46</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Stochastic User Equilibrium</td>
<td>46</td>
</tr>
<tr>
<td>6.2.3</td>
<td>System optimum</td>
<td>47</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Restricted Stochastic User Equilibrium (RSUE)</td>
<td>48</td>
</tr>
<tr>
<td>6.3</td>
<td>Travel demand</td>
<td>51</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Traffic counts</td>
<td>52</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Surveys</td>
<td>52</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Smartphone data</td>
<td>53</td>
</tr>
<tr>
<td>6.4</td>
<td>Driving mode detection</td>
<td>53</td>
</tr>
<tr>
<td>6.5</td>
<td>Conclusion</td>
<td>55</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>59</td>
</tr>
<tr>
<td>7.2</td>
<td>Accuracy of Global Positioning System</td>
<td>59</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Stationary experiment</td>
<td>60</td>
</tr>
<tr>
<td>7.3</td>
<td>Improving positioning accuracy from GPS</td>
<td>63</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Using alternative positioning systems</td>
<td>63</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Delta GPS and Sensor fusion</td>
<td>64</td>
</tr>
<tr>
<td>7.3.3</td>
<td>Map-matching</td>
<td>64</td>
</tr>
<tr>
<td>7.4</td>
<td>Trip detection, segmentation and origin-destination generation</td>
<td>66</td>
</tr>
<tr>
<td>7.5</td>
<td>Conclusion</td>
<td>68</td>
</tr>
<tr>
<td>8.1</td>
<td>Climate gas emissions</td>
<td>69</td>
</tr>
<tr>
<td>8.2</td>
<td>Air quality modelling</td>
<td>70</td>
</tr>
<tr>
<td>8.3</td>
<td>Transportation modelling</td>
<td>70</td>
</tr>
<tr>
<td>8.4</td>
<td>Positioning</td>
<td>70</td>
</tr>
<tr>
<td>8.5</td>
<td>Real life experiments</td>
<td>71</td>
</tr>
<tr>
<td>8.6</td>
<td>Impact</td>
<td>71</td>
</tr>
<tr>
<td>8.7</td>
<td>Future research</td>
<td>72</td>
</tr>
</tbody>
</table>
8.7.1 Transportation modelling  72
8.7.2 Emission modelling  72
8.8 Summary  73

V APPENDIX  75
A CASPER 2016 PAPER  77
B GLOBAL WIRELESS SUMMIT 2016 PAPER  85
C MDM 2017 PAPER  93
D TRANSPORTATION RESEARCH PART D 2017 PAPER  105

BIBLIOGRAPHY  120
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example of GPS data collected from users of an EcoSense app. As this trace is from a user flying in an airplane, it should be removed before map-matching is performed. 19</td>
</tr>
<tr>
<td>2</td>
<td>Emission factor as a function of speed. 31</td>
</tr>
<tr>
<td>3</td>
<td>Dividing driving into urban, rural and highway. 34</td>
</tr>
<tr>
<td>4</td>
<td>A closer look on figure 3. 35</td>
</tr>
<tr>
<td>5</td>
<td>The CO₂ emission factor from the Danish electricity grid. 36</td>
</tr>
<tr>
<td>6</td>
<td>Different contributions to air pollution (figure from [29]). 37</td>
</tr>
<tr>
<td>7</td>
<td>Street canyon model. The turbulence drives pollution towards the leeward side of the street (figure from <a href="http://www.au.dk/thor">www.au.dk/thor</a>). 39</td>
</tr>
<tr>
<td>8</td>
<td>Travel time vs traffic volume (BPR formula). 45</td>
</tr>
<tr>
<td>9</td>
<td>Map of Istanbul. The light green roads denotes roads where traffic has been assigned by the RSUE. 51</td>
</tr>
<tr>
<td>10</td>
<td>Convergence of RSUE algorithm. The error grows when new routes are added to the choice set. Notice the error axis is logarithmic. 52</td>
</tr>
<tr>
<td>11</td>
<td>Test route for driving mode experiment 54</td>
</tr>
<tr>
<td>12</td>
<td>Accelerometer scatter plots. On top are XY axes. Bottom left is XZ. Bottom right is YZ. 54</td>
</tr>
<tr>
<td>13</td>
<td>XY low pass scatter plot. The red lines show four of the clusters. 55</td>
</tr>
<tr>
<td>14</td>
<td>Example of grave GPS positioning errors found by using a running app. The blue trace is the smartphone reported running track. The runner supposedly crossed the river six times without using any of the bridges. The ground truth is shown in red. (My own study) 60</td>
</tr>
<tr>
<td>15</td>
<td>Location reported from stationary smartphone. X-axis is latitude and Y-axis is longitude. 61</td>
</tr>
<tr>
<td>16</td>
<td>Location reported with Wi-fi enabled (North is up). 61</td>
</tr>
<tr>
<td>17</td>
<td>Location reported with Wi-fi enabled, superimposed an aerial photo of the site of the experiment. 62</td>
</tr>
<tr>
<td>18</td>
<td>Daily variation of GPS accuracy 63</td>
</tr>
<tr>
<td>19</td>
<td>Maximum speed in 1 km by 1 km grid and observed locations coloured by client. 67</td>
</tr>
</tbody>
</table>
Figure 20  GPS trip data map-matched to a road map of Herning. Each trip has a different colour, but links shared by more trips only show the top colour.

LIST OF TABLES

Table 1  IPCC considered species  28
Table 2  Data collected in Herning  66

LISTINGS

Listing 1  The code for the Dijkstra shortest path search  50

ACRONYMS

BEV  battery powered electric vehicle
BPR  Bureau of Public Roads
DUE  Deterministic User Equilibrium
GIS  Geographic Information System
GNSS  global navigation satellite system
GPS  Global Positioning System
ICE  Internal Combustion Engine
IPCC  Intergovernmental Panel on Climate Change
ACRONYMS

ITS  Intelligent Transport System
OBD$_2$  On-Board Diagnostics 2
OD  origin destination demand
OSM  Open Street Map
PSL  Path Size Logit
RSUE  Restricted Stochastic User Equilibrium
SUE  Stochastic User Equilibrium
SQL  Structured Query Language
VoT  Value of Time
In the following I will argue for the importance of the research presented in the dissertation, and I will present the State-of-the-art of the research fields that I have built upon and extended.
INTRODUCTION

1.1 INTRODUCTION

Emissions from transportation play a large part in both climate gas emissions and health problems induced by aerial pollutants. 26% of climate forcing gasses in Denmark (2015)[48] are contributed by emissions from vehicles, of which passenger cars accounted for 57%. As the transportation demand continues to grow, the emissions from traffic will also grow, even though government regulation forces the emission of individual combustion engine vehicles to lower levels. The conversion of traffic from Internal Combustion Engine (ICE) into battery powered electric vehicles (BEVs) will change the emission pattern from the local concentrated emission we experience in our cities today, to emission from electricity generation, which will be generated either from sustainable energy sources with no emission or from highly regulated electricity plants. Electric vehicles will also provide a temporal shift in the emission producing power demand, as the power demand is created while charging the battery, not while driving the vehicle. Unfortunately, the transformation towards electric vehicles is too slow to significantly change the emission patterns in the coming decade. Hybrid electric vehicles will only marginally change the emissions patterns as the main power generator in these vehicles is still a combustion engine, although if a large part of the vehicle fleet is converted to battery powered or hybrid electric vehicles the pollution from congestion can be somewhat alleviated.

The problems from vehicle emissions might also be exacerbated by the increasing urbanisation. Forecasts predict that 66% of the world’s population will live in cities by 2050 [46]. This will, with current transportation demand patterns, lead to increased congestion and thus increased concentrations of pollutants in heavily congested areas. For this reason, it will continue to be important to be able to monitor and predict the emissions from vehicle traffic, and thus important to improve the methods used for modelling the emissions from transportation.

Smartphones have seen a dramatic penetration in the mobile phone markets. The introduction of smartphones with multiple sensors created opportunities for large scale crowd sourcing of data. One of the main ideas of the EcoSense1 project, which has funded this study, was to explore the possibility of studying users "in the wild" by

1 http://ecosense.au.dk
getting access to data from the sensors in their phones. One of the original goals of the EcoSense project was to be able to model the emission of climate forcing gasses from transportation, to be used for instance in green accounting, from smartphone data. Another aim of the EcoSense project was to be able to model air quality in congested cities, by using smartphone data.

During the study I did a number of investigations. First, I performed a literature study on how emissions from vehicles are modelled, in order to determine the needed smartphone data. In [41] I extend the traditional emission models by using machine learning on accelerometer data to detect different driving modes.

Because of the versatility of GPS (Global Positioning System) for acquisition of positioning and speed information, an experiment to estimate the accuracy of GPS in smartphones was executed [39].

To alleviate the problem of not being able to collect data from all travellers in an area, I implemented a performant traffic model [40] in a database system, capable of modelling very large cities.

An important input to a traffic model, apart from data describing the road network, is the traffic demand, defined by how many people need to go from one place to another at a specific time. To create the origin-destination matrix for the travel demand, I propose to extend existing methods, such as survey and traffic counts, with actual travel data collected from smartphones [38].

1.1.1 Contributions

The contributions of the work performed in the Ph.D. study are listed below, grouped into the relevant research fields.

1.1.1.1 Transport modelling

- **Generating origin destination demand (OD) matrix from smartphone data.** An algorithm for using data collected from smartphone, to supplement existing methods for generation of OD matrices (see 7.4).

- **Route choice set seeded by GPS traces.** Using GPS traces from smartphone as an initial choice set of routes ensures that real experienced routes are represented in the choice set.
• **Database implementation of restricted stochastic user equilibrium transport model.** I created an efficient database implementation of a transport model, which is scalable to very large urban transportation network.

### 1.1.1.2 Positioning

• **Map-matching.** A topological map-matching algorithm tailored for use in a relational database was developed.

• **Position accuracy in smartphones.** I investigated the accuracy of a smartphone’s positioning system, to show that in a stationary setting over an extended period significant positioning errors occur. The investigation also revealed that using the smartphone’s builtin combination of Wifi positioning and GPS positioning will not always perform better than solely relying on GPS.

### 1.1.1.3 Air quality

• **Air quality model from smartphone data.** I show that it is possible to enhance existing air quality models by using data from smartphones. By using data from smartphones to catch actually travelled routes more realistic travel models can be created. Better traffic models will create better estimates of the local emissions, and thus lead to better estimates of local concentrations of pollutants and air quality.

• **Driving mode detection.** Improve emission factors by detecting the different driving modes (idle, acceleration, cruise, brake, left turn, right turn).

### 1.1.2 Introduction to papers

In the following, a short introduction to the four papers, produced in the Ph.D. project.

#### 1.1.2.1 CASPer 2016

In "Using crowd sensed data as input to congestion model,"[40] a traffic model for large cities is presented. The paper was presented at the 3rd International Workshop on Crowd Assisted Sensing, Pervasive Systems and Communications colocated with PERCOM 2016 in Sydney, Australia. The main contribution of the paper is showing that a database implementation of the RSUE [68] performs well on large cities. Istanbul, with more than 14 million inhabitants, is used as a showcase to exemplify the running
time of the implementation of the model. The database implementation of the transport model completes in approximately half an hour on modest computing hardware (a 2013 laptop).

1.1.2.2 GWS 2016

The paper "Effect of GPS errors on Emission model." was presented at Global Wireless Summit 2016 in Aarhus, Denmark [39]. The paper explains experiments regarding the position accuracy of GPS receivers in smartphones. The paper shows that location accuracy fluctuates over time even for a stationary measurement setup. Furthermore, the paper gives examples of position accuracy which is not always improved by using GPS in conjunction with other location technologies, such as base station triangulation and WIFI triangulation.

1.1.2.3 MDM 2017

The paper titled "Towards vehicle emission estimation from smartphone sensors" was presented at the 18th conference on Mobile Data Management in Daejeon, South Korea in May 2017 (acceptance rate 27%)[41]. The paper explores how to model vehicle emissions using smartphone data. Location data from GPS is used to gain speed information, which is important due to the emission characteristics of combustion engines. Furthermore, vibration data from accelerometers is used to gain information on driving modes. A simple clustering method is employed to discern between the six driving modes: idle, accelerate, cruise, braking, right turn, and left turn.

1.1.2.4 Transportation Research Part D 2017

To report how data from smartphones can be used to create traffic demand matrices, the paper [38] investigates ways of improving the Origin-Destination matrices generated by user surveys and traffic counts. The paper is not yet published.

1.1.3 Structure of the dissertation

The dissertation consists of four parts. The first part contains the motivation, state-of-the-art of the research fields, and a methodology chapter.

The second part contains a summary of the research into modelling emissions from passenger transport. There are two main goals of emission models, the first is to model
emission of climate forcing gases in order to fulfil the contractural obligations stipulated in the Kyoto protocol with oversight by the United Nations. The other goal is to model local air quality by modelling the concentrations of pollutants.

The third part of the dissertation gives a summary of transport models and how the contributions of this dissertation go beyond the state of the art. The data collected from smartphones can only show a small part of the transport demand in an area. To be able to gain a complete picture of emissions from a transport system, a transport model is needed.

Accuracy in positioning systems in smartphones is the topic of the last part of the dissertation. To be able to reliably attribute emissions from a traveller to a specific location, there is a need to know the accuracy of the position of the traveller and how to improve insufficient accuracies.
RELATED WORK

2.1 INTRODUCTION

The work presented in this dissertation is a combination of different research fields. In this section, an overview of the state of the art of the research fields is presented.

The goals of the research presented in this dissertation are:

1. Modelling of emission of climate forcing gases from traffic.
2. Modelling of street level air quality.
3. Using crowd sourced data from smartphones.

The relevant research fields, related to obtaining the goals are: Air quality modelling, Transport modelling, and Positioning.

2.2 AIR-QUALITY MODELS AND VEHICLE EMISSION MODELS

The state-of the art in urban air quality modelling is the Open Street Pollution model OSPM [3, 33]. In OSPM Long Range pollution transport models are used (i.e. Danish Eulerian Hemisphere model [11]) for regional pollution sources, and a urban background model is used for urban pollution sources and lastly a street canyon model is used for local pollution sources. In the street canyon model, turbulent winds produce difference in pollution concentrations across the street as the turbulent winds force the pollution to the leeward side of the street. The model is used to monitor the adherence to the limit values set up by national and international regulating bodies [12], and have been extended to completely cover Denmark [30].

The OSPM model incorporates data from traffic counts, traffic surveys, traffic models, fleet information, and meteorological data.

Carbon dioxide emissions are related to fuel consumption, as the primary source of CO$_2$ is the combustion of hydrocarbons.

luftenpaadinvej.au.dk
Hilpert et al. [27] presents a system based on readout of the On Board Diagnostics system (On-Board Diagnostics 2 (OBD2) 2). The data collected gives information about the motor system. Emissions from the vehicle are calculated, by the relation between engine airflow, the fuel to air ratio, and the CO\textsubscript{2} emission factor for the fuel. No actual data are reported.

To detect and analyze driving styles to prevent car accidents, the Johnson et al. [32] use smartphone sensor fusion, and data obtained from the internal CAN\textsuperscript{3} bus in vehicles. The paper shows detection of driving events is possible and also how to classify the driver aggressiveness.

Another use of OBD\textsubscript{2} is presented in [62]. The data collected from the OBD\textsubscript{2} system is compared to gas sensors in the tail pipe. Three different numerical models are presented and related to the tail pipe measurements. Tail pipe sensors and OBD\textsubscript{2} were used in [16] too. The paper presents estimated emission factors related to different driving modes. The study contributes the fact that emission factors for cold start driving is significantly different from steady state driving. The cold start emission factor has been incorporated into the emission calculation program COPERT [43].

Boriboonsomsin and Barth [6] presents how the effects of the road grade (road inclination) influence fuel consumption, with measurements of OBD\textsubscript{2} data. Two different routes between a start point and an end point are compared over multiple trips. The difference between the two routes is that the first route is flat and the other route has an inclination as it goes through a mountain pass. The fuel consumption is measured by sampling OBD\textsubscript{2} data. The model CMEM (Comprehensive Modal Emission Model) is used to model the fuel consumption.

Hemminki et al. [23] uses accelerometers to detect different transportation modes. The paper presents an overview of methods, applications and problems.

Using Hidden Markov Models classifiers to discern between different transportation modes are given in Mun et al. [45]. The paper argues for creating a personalised environmental impact report generated from mobile sensed data.

To detect idle BEV from idle ICE vehicles Wüstenberg et al. [70] uses accelerometer data to measure the revolutions of an idle combustion engine. The paper uses signal analysis and classifiers to detect the vehicle differences from engine vibrations. The paper also shows that noise generated by a moving car, makes it difficult to discern between a BEV and an ICE vehicle.

Ghose et al. [19] proposes to use a smartphone accelerometer to measure road quality and shows how to detect the positions of potholes.

\begin{footnotesize}
\begin{enumerate}
\item OBD\textsubscript{2} is a common standard in modern vehicles, for monitoring and controlling vehicle status
\item "Controller Area Network", an intra vehicle network standard ISO 11898
\end{enumerate}
\end{footnotesize}
Another use of OBD2 is presented in Beusen et al. [4] where GPS and OBD2 are used to get engine rotational speed (RPM, rotation per minute) information and other engine parameters, to test the long-term impact of Eco-driver training. The paper presents a model to detect different driving modes, such as accelerating, braking and at which engine speeds, gear shifts are performed.

2.3 TRANSPORT MODELS

Transport models are concerned with modelling how properties in a transport network develop under different constraints. A popular underlying concept for transport models is the idea of actors (travellers) who act to increase their economic gain (utility) (or decrease their cost). This leads to models that look for the equilibrium in the utility of the actors, i.e. when no-one can increase their utility by changing the choice of route through the transport network, without sacrificing their utility (increase their cost) (see chapter 6).

In [17] trips and routes are detected from GPS data. The paper presents how to use the previous derived routes to predict current trips. The similarity of an ongoing trip with previously recorded routes, determines the most likely route for the current trip.

Herrera et al. [25] finds that congestion can be accurately predicted with a 2-3% of all travellers reporting their position with GPS devices.

Route Choice Modelling was founded in the 1960’s. A textbook is written by Sheffi in 1985 [61]. The book is still a preferred textbook for teaching the introduction to transport modelling.

The transport model presented in the two papers [57, 68] is a combination of the Stochastic User Equilibrium (SUE) (SUE see chapter 6) and the Deterministic User Equilibrium (DUE), called the RSUE. By restricting the SUE model to only consider a subset of all possible routes, the runtime of the solving of the model can be significantly reduced. The paper by Watling et al. [68] is a presentation of the theoretical background for the development of the algorithm.

The paper by Rasmussen et al. [57] presents a detailed explanation of an algorithm for solving the RSUE. The paper also includes experiments using different error components to model reuse of links in different routes. The paper shows the versatility of the model, and the performance impact of the different link reuse strategies. Two implementations are presented and the running times for running the ArcGIS 4 on a model of the Copenhagen region is given. The ArcGIS implementation completes the model in approximately one hour.

4 http://www.esri.com/arcgis/about-arcgis
The review papers [54] and [53], show different available choices for implementing the random variable for the link cost in the Logit method (see chapter 6) to model the influence of routes with overlapping links, with comments on runtime costs and model performance.

2.3.1 Origin destination demand matrix generation

Yang et. [71] derives OD matrices from traffic counts by solving a Generalised Least Square problem.

In the paper by Zhang et al. [75] information from mobile phone cell towers is used to create an OD matrix. The paper estimates the OD matrix by applying a Horvitz-Thompson estimator. It is assumed cell phone traces generated from cell-phone towers are due to vehicle movements. Triangulating between the connected cell towers is used to calculate position of the cell phone. The travel demand of non-participating part of the population is estimated, but as a the authors report a penetration of cell phones of 87%, only a small fraction is not covered by the data set.

Triangulation between cell phone towers is presented in [65], where cell tower ids are used position the participating cell phone. The purpose of the presented project is to generate OD matrices from the revealed travel routes as a complement/ replacement of household surveys and traffic counts.

In [31] presents an alternative approach to generating OD matrices. The paper uses position metadata from social media updates. As an example the paper uses Foursquare data from the Austin, Texas region. In the presented experiment good correlation with survey based estimation of OD matrices is obtained.

An overview of data collecting methods for OD matrix estimation is given in [9]. The data collecting methods can be divided into self-reporting methods (surveys, personal GPS receivers with self-reported trip classification) and automatic methods (traffic counts, cell phone data, Bluetooth readers, license plate recognition). The paper does not establish methodologies for converting the collected data into OD matrices.

Since the advent and ubiquity of smartphones, many projects have utilised the new opportunities for data collection in the wild. The big problem, from an academic point of view, is that it can be difficult to persuade users to submit their data to a research community. Some research projects have used the smartphone position data indirectly, by using position data included in status updates for social media [34, 67]. Other projects have used cell tower data from network operators, to coarsely trace movements of individual mobile phone users [75]. Others, like in this project, have succeeded in using data directly from smartphones to perform their research studies [20, 24, 45].
2.4 POSITIONING

The GPS system is well documented and very useful [13] for primarily outdoor positioning. The GPS system consists of three parts. A network of ground based control stations, send correctional information to the space located part of the GPS system.

The second part of the GPS is comprised of 27 satellites (of which three are spare), which orbit the earth in approximately 20,000 km altitude. The orbits are configured in 6 planes in semi-geosynchronous orbits, so that each satellite passes the same location on earth twice daily. The satellites broadcast data signal containing information of the location of the satellite and the time according to a high precision onboard atomic clock.

The third part of a global navigation satellite system (GNSS), such as the American GPS, the Russian GLONASS, the Chinese BeiDou or the future European Galileo, is a passive receiver of the signal sent from the satellites. By comparing the travel time of signals from at least three satellites a fix of the position on earth can be calculated. The height above earth can be determined, if data from more than three satellites is available. The position and speed accuracy of non-military commercial receivers have been investigated in different settings.

In Witte and Wilson [69] GPS accuracy is measured by a bike driving on an athletics running track. The GPS receiver used had an external antenna placed on the biker’s helmet, which is different from a smartphone with a built-in antenna placed in a pocket or bag. The paper uses the placement of the antenna to explain the drop-in location accuracy during curves, since the biker’s head will be at another place than the bike, due to the leaning necessary for keeping the balance on the bicycle while turning.

In Modsching et al. [44] the GPS location accuracy from four different GPS receiving devices is investigated in an urban environment. The authors report high error rates due to buildings shadowing the radio signals from the GPS satellites, and provide map-matching as a tool to improve the location accuracy. Furthermore the paper finds that there is no correlation between actual location error and device reported error level.

A thorough investigation of error statistics of GPS location for Geographic Information System (GIS) applications, is given in Zandbergen et al. [74]. The error statistics are measured by a mapping grade GPS receiver in a fixed outdoor position over eight hours. The mean error is reported to be 2.4 meters.

To improve the accuracy of GPS location and speed estimation from smartphone data, the paper by Nitsche et al. [49] employs a Kalman filter on the raw GPS data to reduce the location uncertainty. The presented experiments show how the Kalman filter removes outliers, but also show residual inaccuracy when the filtered GPS data is drawn on a map.
2.4.1 Map-matching

Map-matching is a technique for finding the most probable road from a noisy location data. The noise in the data can be from different sources, either the inherent inaccuracy of the location measurement as in the GPS, or due to low sampling of the location measurement, i.e how to select roads travelled between location measurements.

In [5] smartphones are used to track transit vehicles. The authors apply a sort of map-matching to the tracked GPS data in order to geolocate the transit vehicles to provide arrival time information to transit passengers. The GPS traces of the transit vehicles are mapped to previously recorded traces from test vehicles and not to a real road map.

Quddus et al. [56] divides the different types of map-matching algorithms into four categories. The geometrical map-matching algorithms look at single location measurements and find the nearest road segment as the solution for the real position of the user. The geometrical algorithms often have problems with measurements at intersections or when roads run in parallel to each other (i.e. a rural road along a highway).

The second category is the topological map-matching algorithms. These algorithms exploit the topology of the road network to find the route which fits the location measurements the best. The third category is the probabilistic map-matching algorithms, which attach a region (elliptical or rectangular) around each location measurement with the size of the region proportional to the expected error in the position measurement. The road segment inside the region is then considered the road the device is moving along. These algorithms have the same problems as the geometrical algorithms, so extra consideration has to be taken when more road segments are inside the error region.

The fourth category is called the advanced map-matching algorithms. The advanced methods of map-matching often use data from other sensors, like dead reckoning sensors, to complement the position sensors. Many of the mentioned advanced map-matching algorithms use a Kalman filter to fuse the data from localisation, speed, and dead reckoning into a position on the map. The advanced map-matching category comprises a wide range of techniques, such recursive Bayesian inference or Multi Hypothesis Technique.

Quddus et al. [56] also provides an overview of the accuracy of the mentioned algorithms, as well as a look to how the future Galileo satellite based positioning system will impact the need for map-matching. Quddus et al. concludes that map-matching will still be needed in intelligent transport systems, especially in places where satellite signals are deteriorated by multi path propagation.

Brakatsoulas et al. [8] presents two different approaches to topological map-matching, an incremental map-matching method, which considers local measurement points
around every GPS point, and a global method to find a route in the road network which is most similar to the GPS trace. To compare the different approaches a quality measure based on the Fréchet distance is used.
3.1 INTRODUCTION

In this section, an overview of the research methods used in this project is given. The overall scientific process is an abductive reasoning approach, with inspiration from design thinking [14]. This has been an extension of the way the author has worked as an engineer, working in product research, design, and development. One of the important lessons of the Ph.D. study has been a training in the scientific method, and always to question my assumptions, leading to reflection on the results achieved and the method of achieving the results.

3.2 HYPOTHESES

After literature studies and small tests, the following hypotheses were formulated:

1. The positioning accuracy of smartphones varies with time, location, weather, and movement of the smartphone.

2. Air quality models for street level pollution concentrations can be refined by smartphone travel data, by taking driving mode of vehicles into account. The driving mode is divided into six different classes: idle, acceleration, deceleration, cruise, left turn, and right turn.

3. Emission models for climate change gasses from transportation can be improved by using data from smartphones.

4. Smartphone trip data can be used to create OD travel demand matrices as an extension in conjunction with traditional methods.

3.3 EXPERIMENTS

A number of experiments have been performed to test the hypotheses. A brief overview of the experiments and their goals are given below.
3.3.1 GPS accuracy

To test the time and weather related positioning accuracy of hypothesis 1 a stationary experiment with a smartphone was performed. The experiment could not correlate the positioning accuracy to certain times of day or certain weather events. The experiment did find differences in the positioning accuracy, whether the positioning relied only on GPS or both GPS and Wifi (see more in section 7.2.1).

3.3.2 Driving modes

The hypothesis 2 was tested by driving different cars with a smartphone on board, to capture positioning and accelerometer data. The experiment showed that driving modes could be extracted from the accelerometer data, but a model for the fuel consumption could not be validated because of inaccurate fuel measurements (see 6.4).}

3.3.3 Experiments with EcoSense data

Data from the EcoSense project was used to perform various experiments. The database for the "Herning cykler til månen" (Herning bikes to the moon) contains data from 909 trips from 17 users and 750000 data points, after removing non road transport related data. As the data collection was done by people with no relation to the Ph.D. study and not known by the author, the smartphone application has been used in many different use cases. Some users have used it for the official purpose of measuring the cumulated biking distance of the citizens of Herning municipality, but as can be seen from the collected data, the application has also been used while users have been walking, running, driving a vehicle, travelled by train and even by airplanes (see figure 1) and boats. To focus on road transport the data has to be filtered to remove traces not related to passenger transport in cars. The following sections describe the experiments performed with the data.

3.3.3.1 Transport model

To solve the problem of having transportation data for only a small part of the travellers in the transport system, a transport model was implemented. To make the transport applicable to different cities, the implementation was made as generic as possible. Changing the model to work in a different city would only entail changing some input parameters. As the transport model has to know about roads and street in the city under investigation, and since a global, free, and open source of map data containing
Figure 1: Example of GPS data collected from users of an EcoSense app. As this trace is from a user flying in an airplane, it should be removed before map-matching is performed.

The demand matrix is necessary for creating a transportation model for a road network. The hypothesis 4 was tested by filtering the EcoSense data to a data set containing only trips by a motor vehicle. Walking, running, and biking was determined by the maximum speed of a trip as measured with the GPS receiver in the smartphones. Low speed trips were excluded from the data set. To remove flying from the data set both height and speeds were necessary to discern flying from highway driving[38]. (see section 7.4.)

Origin destination pairs is determined as start and end points of the remaining map-matched trips in the data set. The demand for each origin destination pair is estimated by population density of the origin area and the distance to the destination.
Part II

EMISSION MODELS

In this part of the dissertation I will present my work on creating models for vehicle emissions using smartphone data, discuss the United Nations framework for climate change related activities, and present an overview of the existing models.
4.1 CLIMATE CHANGE POLICY DEVELOPMENT

4.1.1 Climate change organisations

It has been decided in the United Nations to create two bodies for devising policies for managing the anthropogenic global warming. The main political body is the United Nation Framework Convention for Climate Change (UNFCCC) \[59\]. The name refers to both a convention created in 1994, and a secretariat that services the operation of the Convention. This body is tasked with creating policy initiatives and negotiating treaties with the members of the UN. A visible part of these negotiations are the COP (Conference of Parties) meetings held annually. Since 154 nations have ratified the convention, the negotiations are very difficult, and progress in agreeing on firm goals and plans to reach the goals is slow \[51\].

To support the UNFCCC with scientific knowledge the Intergovernmental Panel on Climate Change (IPCC) was created. The panel gathers scientific evidence into reports on different aspects of climate change, the root causes, the consequences and ways to mitigate and adapt.

Even though the panel is a scientific panel, it is also a part of a political process. This means that the reports made by Intergovernmental Panel on Climate Change (IPCC), especially the "summary for policy makers" are also heavily negotiated by the participating government officials.

The latest reports from IPCC are the AR-5 (fifth assessment report) series.

The IPCC reports are divided into sections, done by different workgroups. Workgroup 1 delivers a scientific background section containing the latest scientific knowledge about emissions, measured consequences, melting of Icecaps etc. As a new feature in the assessment report a number of scenarios are envisioned, to estimate the consequences of different policies. The scenarios are called "Representative Concentration Pathways", and are named by the expected rise in global mean temperature. Four scenarios are envisioned in the report: RCP2.5, RCP4.5, RCP6, RCP8.5 \[64\], where the numbers refer to the maximal temperature increase due to anthropogenic climate
change. These scenarios are then used as a guiding principle for the rest of the sections in the AR-5.

The workgroup 1 report is generally viewed as a high quality and reliable source of knowledge.

Workgroup 2 delivers a report on "Impacts, Adaptation and Vulnerability".

Workgroup 3 is about mitigation of climate change.

4.1.2 Climate change mitigation

From the start of the debate on how to combat Anthropogenic Global Warming, there have been two competing approaches. The mitigation approach, where an effort is made for reducing the emissions of climate gases, is the approach which has received the most attention [52]. The EcoSense project is part of the mind-set behind this approach. Whereas the mitigation approach previously has focused on creating more efficient machines, to either produce energy with less emissions or produce machines that produce more useful work per energy unit [10], EcoSense focuses on how the machines are used in combination. Instead of focusing on a single energy consuming item, we are trying to analyse how the energy consumption changes as the individual items work together. In other words, developing tools for studying network effects in the transportation sector, as well as in other sectors, is a recent idea.

4.1.3 Climate change adaptation

The second approach is called the Adaptation approach. Since it is probable that there will be significant changes in the climate, even if we succeed in keeping the CO₂ emissions within the agreed limits, we have to invent ways to adapt to these changes. The Adaptation approach has not received much attention or research funding. It seems that IPCC in the fifth report from Working Group 2 is beginning to give more attention to the adaptation approach, as a consequence of realising that some level of Climate Change is inevitable [36, 63].

4.2 Conclusion

In this chapter I have presented the political framework for the United Nations monitoring of emissions of climate gases and the two different approaches to combat the effects of climate change; climate change mitigation and climate change adaptation. In
the past, these two views has been perceived as opposing strategies, but recently it has become apparent that both strategies has to be considered to develop climate change policy. The emission model methodology developed by IPCC is used as a foundation for the rest of the dissertation.
5.1 INTRODUCTION

In this section emission models for vehicles are presented. The starting point for the emission models is the methodology developed by IPCC. The three tiers of IPCC is presented as well as the proposed extension by adding driving mode detection to the CO₂ emission model.

5.2 CO₂ EMISSION MODELLING

In this section, the IPCC [66] methodology for calculating emissions of climate forcing gases from transportation is described. The focus of the IPCC approach is to offer methods for building national inventories for climate gas emissions. The following is an attempt to summarise the key methodologies, and I will use this as a basis for developing models for single trip emissions.

5.2.1 IPCC methodology

The IPCC has made a number of reports on how to calculate emissions of climate forcing gases and pollutants. The gases that IPCC is describing methods for are divided into four groups (see table 1):

Group 1 is a group of pollutants for which there exists detailed methodology for estimating the emission. The emissions are calculated from activity data, such as driving conditions and engine conditions.

Group 2 are pollutants which can be estimated from fuel consumption, when there is a direct connection between the burning of fuel and the emission. The estimates of emissions of the Group 2 pollutants are regarded to be as precise as the estimations for the Group 1 pollutants, even if the methodology differs. Group 3 and Group 4 are pollutants, where no detailed methodology exists for estimating the emission, so a simple method is used to calculate the emission.
## Group 1

<table>
<thead>
<tr>
<th>Compound</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon monoxide</td>
<td>CO</td>
</tr>
<tr>
<td>Nitrogen oxides</td>
<td>(NO\textsubscript{x}: NO and NO\textsubscript{2})</td>
</tr>
<tr>
<td>Volatile organic compounds</td>
<td>(VOCs)</td>
</tr>
<tr>
<td>Methane</td>
<td>(CH\textsubscript{4})</td>
</tr>
<tr>
<td>Non-methane VOCs</td>
<td>(NMVOCs)</td>
</tr>
<tr>
<td>Nitrous oxide</td>
<td>(N\textsubscript{2}O)</td>
</tr>
<tr>
<td>Ammonia</td>
<td>(NH\textsubscript{3})</td>
</tr>
<tr>
<td>Particulate matter</td>
<td>(PM)</td>
</tr>
</tbody>
</table>

## Group 2

<table>
<thead>
<tr>
<th>Compound</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon dioxide</td>
<td>(CO\textsubscript{2})</td>
</tr>
<tr>
<td>Sulphur dioxide</td>
<td>(SO\textsubscript{2})</td>
</tr>
<tr>
<td>Lead</td>
<td>(Pb)</td>
</tr>
<tr>
<td>Arsenic</td>
<td>(As)</td>
</tr>
<tr>
<td>Cadmium</td>
<td>(Cd)</td>
</tr>
<tr>
<td>Chromium</td>
<td>(Cr)</td>
</tr>
<tr>
<td>Copper</td>
<td>(Cu)</td>
</tr>
<tr>
<td>Mercury</td>
<td>(Hg)</td>
</tr>
<tr>
<td>Nickel</td>
<td>(Ni)</td>
</tr>
<tr>
<td>Selenium</td>
<td>(Se)</td>
</tr>
<tr>
<td>Zinc</td>
<td>(Zn)</td>
</tr>
</tbody>
</table>

## Group 3

<table>
<thead>
<tr>
<th>Compound</th>
<th>(PAHs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent organic pollutants</td>
<td>(POPs)</td>
</tr>
<tr>
<td>Polychlorinated dibenzo dioxins</td>
<td>(PCCDs)</td>
</tr>
<tr>
<td>Polychlorinated dibenzo furans</td>
<td>(PCDFs)</td>
</tr>
</tbody>
</table>

## Group 4

<table>
<thead>
<tr>
<th>Compound</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alkanes</td>
<td>(C\textsubscript{n}H\textsubscript{2n+2})</td>
</tr>
<tr>
<td>Alkenes</td>
<td>(C\textsubscript{n}H\textsubscript{2n})</td>
</tr>
<tr>
<td>Alkynes</td>
<td>(C\textsubscript{n}H\textsubscript{2n-2})</td>
</tr>
<tr>
<td>Aldehydes</td>
<td>(C\textsubscript{n}H\textsubscript{2nO})</td>
</tr>
<tr>
<td>Ketones</td>
<td>(C\textsubscript{n}H\textsubscript{2nO})</td>
</tr>
<tr>
<td>Cycloalkanes</td>
<td>(C\textsubscript{n}H\textsubscript{2n})</td>
</tr>
<tr>
<td>Aromatic compounds</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: IPCC considered species
The fourth group of pollutants are species, where the emission is calculated as a fraction of the Non-Methane Volatile Organic Compound (NMVOC).

The most important gases when considering Climate forcing gases are in Group 1 and 2, as they are the gasses with the largest impact on climate change and human health. Thus the focus of this Ph.D. study has been on CO\textsubscript{2}, CH\textsubscript{4} and NO\textsubscript{2}. The amount of CO\textsubscript{2} emissions is by far the largest, but since the radiative forcing of CH\textsubscript{4} is 21 times higher than CO\textsubscript{2} and the radiative forcing of NO\textsubscript{2} is 206 times higher (per molecule), they contribute significantly to the global warming.

The way emission inventories are created is described in the Guidelines from IPCC and extended in EMEP/EEA Guidelines \[50\]. In the guidelines three different methods are described, each method more accurate than the previous.

5.2.1.1 IPCC Tier 1 model

The Tier 1 method is based on numbers for national sales of hydrocarbons (Gasoline, Diesel, Natural gas etc.). These numbers are readily available for most countries and are converted into emission inventories by multiplying emissions factors (grams of the specie pr. kilogram of fuel) for each type of fuel. The Tier 1 method is the simplest, but also the most crude way of estimating the national emissions. The Tier 1 method does neither account for import or export of fuel, nor that not all emissions are directly related to fuel consumption. The only data points needed are: Volume of sales of the different fuel types and the emission factors for the different fuel types as shown in equation 1.

\[
E_{p}^{\text{CALC}} = \sum_{i} V_{\text{sales},i} \times e_{i,p}
\]  

(1)

Here the \(E_{p}^{\text{CALC}}\) is the calculated emission for the pollutant \(p\), where \(i\) is the fuel type, \(V_{\text{sales},i}\) is the sales for fuel type \(i\) and \(e_{i,p}\) is the emission factor of pollutant \(p\) for the fuel type \(i\).

This method is inadequate to estimate single trip emissions.

5.2.1.2 IPCC Tier 2 model

In the Tier 2 method the emission inventories are estimated by estimating the traffic volumes for different categories of vehicles, and multiplying emission factors (gram pr. kilometre) for each category. The vehicles are divided into six main categories: Passenger Car, Light Duty vehicles, Heavy Duty vehicles, Buses, Mopeds, and Motorcycles.
For each of the main categories, a subdivision is made, to accommodate for different emission characteristics stemming from pollution regulation, fuel type and engine size. For instance, in Europe, passenger gasoline cars are subdivided into 13 different types, according to the legislation governing allowed emissions. These regulations have been changed and tightened 13 times since the first emission control legislation was ratified in the early nineties. For each vehicle category, vehicle type and legislation class, activity data has to be obtained. The activity data consist of the number of vehicles, and the number of kilometres they drive pr. year, for each class. The IPCC has generated tables of emission factors (as g/km) for each class of vehicles. By multiplying these emission factors with the estimated kilometres and number of vehicles in the class, the total emission of a pollutant can be estimated for the vehicle class. The total annual emission from transport can then be calculated as the sum of the total emission for each vehicle class.

\[
E_{p}^{\text{CALC}} = \sum_{v} \sum_{i} D_{v,i} \cdot e_{i,v,p}
\]  \hspace{1cm} (2)

In equation 2 the \( E_{p}^{\text{CALC}} \) is the calculated emission, where \( i \) is the fuel type, \( D_{v,i} \) is the estimated distance travelled vehicle type \( v \) and fuel type \( i \) and \( e_{i,v,p} \) is the emission factor for the fuel type. \( p \) is the pollutant, for which we are finding the emission.

The Tier 2 model contains data that can be used for a simple emission model for single trip emissions.

5.2.1.3 IPCC Tier 3 model

The Tier 3 method is based on the Tier 2 method and improves on the estimated emission, by also considering the velocity distribution of the travelled distances, by considering the effects of cold-starts and vehicle age on the total emissions.

There are two ways proposed to calculate the effects of speed on exhaust emissions. Either by dividing the travelled distance into road types with different speed characteristics, i.e. urban, rural and highway. In this case, the total emission for a vehicle class will be calculated as the sum of the product of travelled distance on a road type and the emission factor for that road type and vehicle type.

The other method uses a measured speed to emission curve and a speed distribution function to estimate the emission. For some pollutants, an emission factor (pollutant emission pr. kilometer) is given, but for pollutants that are directly linked to fuel consumption (i.e. CO\(_2\)) the fuel consumption is calculated as function of speed (for instance equation 4). The emission calculated from the fuel consumption is then given
by a formula derived from the chemical combustion reaction. For CO$_2$ the relationship between emission and fuel consumption is given by the formula 3.

$$E_{CO_2}^{CALC}, km = 44.011 \times \frac{FC_{CALC}}{12.011 + 1.008 \times r_{H:C} + 16.000 \times r_{O:C}}$$  \hspace{1cm} (3)

This formula is based on the assumption that all fuel is consumed, i.e. that each carbon atom in the fuel is oxidised into CO$_2$, which has the molar weight of 44.001g/mol. The numerator $FC_{CALC}$ is the calculated fuel consumption factor, and this is divided with the molar weight of carbon plus the molar weight of hydrogen multiplied with the ratio of hydrogen to carbon, $r_{H:C}$, plus the molar weight of oxygen multiplied with the ratio of oxygen to carbon, $r_{O:C}$, in the fuel. There are tables for different ratios $r_{H:C}$ and $r_{O:C}$ for different fuel types, given in the COPERT IV methodology guidebook [50].

$$FC(V) = \frac{a + c \times V + e \times V^2}{1 + b \times V + d \times V^2}$$  \hspace{1cm} (4)

As an example, for passenger vehicles that are regulated by EURO 1 the fuel consumption factor curve is given by equation 4. The coefficients $a$ to $e$ are tabulated for the different engine sizes.
An emission factor (in g/km) versus speed curve can be seen in figure 2 for a vehicle with engine size less than 1.4l [43], as the blue curve. It can be seen that the emission factor is larger for low speeds and for high speeds and lowest at moderate speeds (app. 60 km/h). For low speeds, the higher emission factor is due to the engine being underused. For higher speeds, the opposing forces from wind and friction in bearings, increase the work from the engine, and thus a higher emission factor. By considering the emission pr. minute, obtained by multiplying FC(V) with the speed (km/min), the green curve can be calculated. From the green curve, it can be seen that the emission pr. time unit is a monotonic growing function, that resembles a second order polynomial, which fits well with the physical observation that the aerodynamic drag on an object is proportional to the square of the speed. The idle fuel consumption can not be calculated by equation 5, because the fuel consumption equation 4 is only valid for speeds larger than 10 km/h.

$$FC_t(V) = FC * V = \frac{a * V + c * V^2 + e * V^3}{1 + b * V + d * V^2}$$ (5)

The emission is calculated as the integral of the speed distribution multiplied by the emission function as seen in equation 6.

$$e_{i,k,r} = \int e(V) * f_{k,r}(V) dV$$ (6)

i is the pollutant for which the emission is calculated, k is the vehicle type, r is the fuel type, e(V) is the emission function of speed, and f_{k,r} is the speed distribution function for the vehicle type k and fuel type r.

The data needed to use the Tier 3 method is quite extensive. For each of the 13 vehicle classes (divided by fuel type, vehicle size) activity data for mileage for urban, rural and highway travel, hot start/cold start, as well as data for the number of vehicles in each emission regulation is needed. Fortunately, the program COPERT IV contains the emission factors, and country specific activity data can be downloaded to use for evaluating national inventories.

This model will form the basis of a complex model for emissions of single trips.

5.2.2 Simple emission model

In this section I will condense the contributions I have made to the EcoSense project. The emission factors for combustion engines, as used in the national emission inventories, are based on measured emissions from standardised test runs [2], and are used
for a bottom up estimate of the total national emissions, by estimating the number of kilometres the total national fleet has driven [47]. This estimate is compared to an estimated emission calculated from the amount of gasoline and diesel sold nationally, and the difference between these two estimates leads to a correction factor for the total national emission.

I have developed and implemented a simple model for estimating the emissions from single trips. The trips are divided into sections by the transportation mode detection algorithm. For each transportation mode, an emission factor is calculated and multiplied with the length of the section. The total emission for a trip is calculated by accumulating the emissions from the sections. The emission factors for cars, trucks, buses and trains are derived from the reporting of the national emission inventories to UNFCC under the Kyoto Protocol.

### 5.2.2.1 Mobile sensing

When using smartphones as mobile sensing devices, a number of different sensors can be used. In the following the built-in accelerometer and the GPS sensor are considered. The accelerometer sensor delivers a timestamped accelerometer measurement, consisting of a reading of the 3-axis accelerometer. The accelerometer is a very low power device and can thus be sampled more frequently (approximately 200 Hz) [37].

The GPS measurements consist of longitude, latitude, altitude, speed, bearing, number of satellites visible, accuracy and a timestamp of the measurement. The GPS sensor contains a radio receiver, and complex logic to decode the radio signals from multiple satellites and calculate a position from the data, thus it uses more electric power. To reduce the power drain on the smartphone battery, the GPS can be turned off between readings of the position.

When calculating the distance travelled between two GPS measurements a Haversine function is used. This function takes the curvature of the earth into account, and therefore gives a more precise value for the distance. The resolution of the test data, does not warrant the use of the Haversine approach, since it is sampled every second, but was chosen in anticipation on more crudely sampled GPS data, due to power conservation considerations.

### 5.2.2 Tier 2 derived model

The emissions are estimated for three different driving patterns: Urban, with many stops and low speed. Rural road driving, with few stops and moderate speeds, and lastly highway, with no stops and high speed. These three modes of road traffic can be distinguished by analysing GPS data. Figure 3 illustrates an example of a speed curve
for a series of trips. The figure shows how the trips are divided into the different driving types. The distance travelled for each road type can be found by adding the distance between sampling points for points in each driving type. These distances are multiplied with the emission factors from the Tier 2 methodology.

In figure 4 a closer look on figure 3 is shown. From the figure, it can be seen that the division with fixed speed limits delivers a quite crude division and gaps where the speed changes across a boundary between driving types.

To explore the data, and develop the implementation IPython Notebooks 1 with the numerical and scientific packages NumPy 2, Scipy 3 and Pandas 4 was used.

5.2.3  Driving mode detection

Driving mode detection is a technique for classifying vehicle movements into different modes. In this dissertation the driving modes of interest are: idle, acceleration, deceleration, cruise, left turn, and right turn. The aim is to associate different emission factors to each of the driving modes to obtain a more detailed emission model. A K-means

---

1 http://ipython.org/notebook.html
2 http://www.numpy.org
3 http://scipy.org
4 http://pandas.pydata.org
classifier is trained on accelerometer data from an experiment, where cars were driven at the same test route. The result of the classification for the idle case, was that the driving mode was misclassified in up to 25% of the classified idle measurement for ICE vehicles. For electric vehicles the misclassification of idle mode was below 1%.

5.2.4 Modelling emissions from electric vehicles

It is possible to calculate the emissions for electric cars by using near real time emission data from supervisor of the public Danish electricity grid, Energinet.dk, under the assumption that electric cars will be charged with electricity from the public grid. There is for now no way to detect if the charging of electric vehicles is done through the public electric grid, or by private solar electricity producers or other producers. These charging situations would have different CO\(_2\) emission factors.

The dataset from Energinet.dk contains production data for the large power plants, the distributed power plants, wind farms and solar. For each kind of producer, the production is given for east and west Denmark, and the import/export of power to the neighbouring countries is also given. There is data for the temperature and wind speed (for a single location), and an estimated emission factor in g\(\text{CO}_2\)/kWh. The resolution of the data is 5 minutes. The CO\(_2\) factor from the Danish public grid for a period from 2011-2016 can be seen in figure 5.
Figure 5: The CO₂ emission factor from the Danish electricity grid.

5.3 AIR POLLUTION DISPERSION MODELLING

This section gives an overview of different ways for calculating the change, over time and space, in concentration of pollutants. Traditionally, a main focus for investigation has been how pollutants move from a pollution source towards surrounding areas. The concentrations of the species (each kind of pollutant is regarded as a specie) change partly by dilution (the plume widens), partly by deposition pollutants on the ground or in water, and partly by chemical reactions in the atmosphere. The chemical reactions are dependent on different factors, such as light and temperature. In the section on OSPM (section 5.3.4) the effect of the landscape (here a street canyon) is discussed.

To model the concentration of pollutants at a specific place, the researcher has to consider the different sources of the pollution, as pollution is transported over wide ranges. To model the long-range transport of pollutants coarse models are used. The long-range models will contribute to a regional background pollution concentration. For a city, more fine grained models (Urban Background Models) are used to calculate the contribution from the city, which leads to the calculation of a city increment of emissions added to the regional background concentration. When street level pollution is the target of the calculation, a model for the pollution contribution and distribution for the street is needed, which considers local emissions and local conditions to calcu-
late the street level increment added to the regional background concentrations and urban background concentrations (see figure 6 from [29]).

5.3.1 Eulerian method

The Eulerian method is also called a grid model. For the area under investigation, the volume of air above the area is divided into boxes. In at least one box there should be a source of pollution. The concentration of the pollutants in the box can be calculated as the amount of the species coming into the box (either through the boundaries of the box, from chemical reactions or from sources located inside the box) minus the amount of the species lost (through the boundaries, by deposition or by chemical reactions).

For each time step the concentration of each specie is calculated by solving an ordinary differential equation (ODE) and the calculated concentrations are used to determine the start conditions for the next time step, as well as the boundary conditions for the surrounding boxes [78].

Examples of grid models are Danish Eulerian Hemisphere Model (DEHM) [11], which calculates concentration of 63 different species and 120 chemical reactions in different scales (size of a single grid) (150 km x 150 km, 50 km x 50 km and 17 km x 17 km) [18].
5.3.2 Lagrange

In the Lagrange method, a parcel of air is considered. The movement of the parcel is simulated by considering the local meteorological conditions, and the statistical mechanics of the particles in the parcel. By simulating many traces of parcels the advection of the pollutants from a source can be visualised. In this method for calculating the advection of pollution the observer follows the advection from the perspective of the pollution, instead of from a fixed position [60]. Inside the parcel the reactions of the gases are modelled while the parcel is moving.

The modelling can also be performed backwards to ascertain where the air pollution at a given position is originating from.

5.3.3 DASK-GAS

DAnish Solver for chemical Kinetic - Gear, Analysis, Solver (DASK-GAS) \(^5\) is a chemical compiler and solver. The idea of DASK is to automate the code generation for handling chemical reactions. In DASK you give the chemical reactions and their reaction rates as input, and the program transforms the input into Fortran code.

The reactions are given as the names of chemical compounds in a reaction schema, with input reagents on the left side and resulting compounds on the right side. The temperature dependency of the reaction rate is given as a Fortran function, and if the reaction is driven by photolysis it is marked in the data file. For each of the reactions Fortran code is generated to calculate loss and gain of the species. The generated Fortran code is called inside the ODE solver, to make it easy to add new chemistry [26].

The setup of the chemical reactions used in the simulation is called a mechanism, a number of standardised mechanisms exist, such as European Monitoring and Evaluation Programme, EMEP, Regional Atmospheric Chemistry Mechanism, RACM and Regional Acid Deposition Model Mechanism, RADM2 [21].

Beside the chemical compiler, there is also a simple way to set up scenarios for the solver to model.

The specification of photolysis takes into account the amount of sunlight available in the period and place which is simulated. This is done by listing values for the sunlight for each time interval within the simulation period in a separate data file.

---

\(^5\) unpublished solver from Prof. Allan Gross
The Operational Street Pollution Model is a model for calculation of pollution in urban environments. Due to turbulent wind conditions in urban street canyons, the pollutants do not mix well with the surrounding air. The heavier pollutants will concentrate on the leeward side of the street canyon (see figure 7). The OSPM model considers the effects of street geometry, wind speed, emission factors and atmospheric chemistry (i.e. the NO – NO$_2$ – O$_3$ cycle), as well as meteorological data, background pollution originating from city contributions and from regional background due to long range air transport. The COPERT IV model for activity based emission models is built-in to OSPM, as an example of the extensive data used to model the street level pollution concentrations.

To use the program, detailed information on the geometry of the street and surroundings must be provided. The height of the surrounding buildings, the angle of the street with respect to the wind and the distance to intersections of street corners has to be input to the program.

http://envs.au.dk/videnudveksling/luft/model/ospm/
The program uses emission data for background concentrations from European Monitoring and Evaluation Programme (EMEP), as input but actual traffic volume for the considered street has to be provided [3].

5.4 CONCLUSION

In this chapter I have presented the methods for modelling air quality in urban environments. The current models for air quality combines research into pollution dispersion over long distances, chemical reactions in the atmosphere, and transportation models based on traffic surveys and traffic counts. Furthermore a method for analysing detailed emissions from vehicles through driving mode detection using data from smartphones, was presented.
Part III

TRANSPORTATION MODELS

Previously in the dissertation I have presented models for emissions from individual vehicles from smartphone data. As it is unfeasible to get complete knowledge of all vehicle movements in a transport system, there is a need for a traffic model to predict the speed and traffic volume at each street, in order to be able to model vehicle emissions for a large transport network.
Transport of people, goods etc. has a crucial place in our society. Optimisation of transport can therefore save the society money, impact the air quality and improve the quality of life in cities. Transport models can be used as a tool for optimisation as the model can predict the outcome of changes in the transport network such as congestion. In the following I will present an overview of the transportation models I have considered for creating a model, capable of describing the traffic volumes and speed at street level.

6.2 ROUTE CHOICE MODELS

The basic idea in the field of route choice models comes from econometrics [61]. The idea is to find an equilibrium that satisfies the needs of all travellers. A traveller is considered to have a travel need (travel demand), for travelling from an origin to a destination. The travel demand can be satisfied by many different routes through the transport network. The traveller chooses a route by measuring the utility (value) gained by travelling that route. All travellers are expected to choose a route that maximises their utility (or minimises the perceived cost), and the equilibrium is obtained when all travellers have chosen routes, and changing route would decrease the utility (increase the cost) for the traveller.

For most travellers, fulfilling the transportation need is associated with cost, and this cost is, when considering different routes to take, mainly the travel time associated with the route. In the following the equilibrium condition is met when it is not possible for any traveller to change their route and achieve a smaller travel time.

Modelling actors in a travel network as selfish travellers, leads to congestion of the most used roads. Congestion should lead travellers to choose other less congested routes, and thus the travel time of each road is modelled as dependent of the number of travellers using the road. The more travellers on a road, the longer travel time for that road.
To better understand the details of the route choice models, some of the key concepts need to be defined:

- **Link.** A link connects two nodes. Traffic flows into and out of the link at either end and cannot be created or disappear on the link.

- **Node.** Nodes connect links. Nodes are typically placed at road intersections.

- **Route.** Routes consist of links.

- **Origin and destination nodes.** Travellers enter the travel network at origin nodes, and exit the travel network at destination nodes. The travel demand for the travel network can be described as an OD matrix, where the element at row $i$ and column $j$ describes the number of travellers going from origin node $i$ to destination node $j$.

- **Value of Time.** To convert the travel time into a cost (negative utility), the concept of Value of Time (VoT) is introduced. The VoT is a subjective figure, measuring the perceived cost of the time spent in a specific transportation mode. An example is that the VoT of driving in a queue is larger than driving in the Copenhagen Metro [1, 15], which means that the time spent in the metro feels more efficient than spending time in a queue.

- **Congestion.** The travel time is a function of the number of cars on each link in order to model congestion. The Bureau of Public Roads (BPR) defined the classic travel time function as equation 7.

\[
\text{Traveltime}(x_a) = t_0^a \left(1 + \alpha \left(\frac{x_a}{C_a}\right)^\beta\right)
\]  

(7)

The travel time for link $a$ with the traffic volume $x_a$ is given in 7, where $\alpha$ and $\beta$ are constants depending mainly on driving culture - values of $\alpha = 0.15$ and $\beta = 4$ has typically been used[61]. The value $C_a$ denotes the capacity of road, and $t_0^a$ is the free flow travel time for link $a$, i.e. the time it will take a traveller to traverse the link if there is no congestion while observing speed restrictions.

A graph of the BPR function can be seen in figure 8, for a link with a capacity of 1000 vehicles pr hour and a free flow travel time $t_0$ of 1 minute. As the traffic volume is low the travel time is equal to the free flow time, but when the traffic volume becomes larger than the capacity the travel time increases rapidly. The reason for having a travel time function, that operates on travel volumes above the capacity road, is to avoid deadlock situations in iterative model solving algorithms. The links that are assigned high traffic volumes will get high travel times which in a subsequent step
of the algorithm will lead to traffic moving to other links to avoid the congested link, thus lowering the traffic on the congested link.

![BPR travel time function](image)

Figure 8: Travel time vs traffic volume (BPR formula).

The stated problem of transport modelling is how users choose a route from an origin to a destination to satisfy their transport need with the lowest cost. If we consider the transport system as a graph, where the links are vertices, and the endpoints of the links are the nodes, the task of the model is to assign traffic volumes to each link in such a way that the travel demand of all users are satisfied and the equilibrium condition is met. To find the best (fastest) route a naive algorithm would find all the routes in the transport network that can satisfy the transportation need and choose the fastest. Unfortunately the number of routes in a transportation network grows rapidly with the size of the network. The problem of counting the number of routes is a #P-complete problem\[58\]. On one hand it is possible to find the shortest path through a network with Dijkstra or A* algorithms, on the other hand it is hard to enumerate all possible routes, which would be needed to consider alternative routes when congestion rises.

There are different ways to state the equilibrium condition, with different algorithmic properties. Here I present the two classic ways of stating the equilibrium condition of the Route choice problem, the **DUE** and the **SUE**.
6.2.1  *Deterministic User Equilibrium*

In the *DUE* every traveller has perfect knowledge about the transport network. All travellers know all possible routes, the travel time on all links, and what route all other travellers have chosen. The equilibrium condition can then be expressed as no traveller can change their chosen route and get a shorter travel time.

The *DUE* can be solved iteratively by the Frank-Wolfe [61] method, a convex combination minimisation method.

6.2.2  *Stochastic User Equilibrium*

In the *SUE* we introduce an element of uncertainty, so that the equilibrium condition can be expressed as: all travellers think they know the fastest route to meet their travel demand. The uncertainty is introduced in the link travel time equation by adding a random variable to the BPR formula:

\[
\text{Traveltime} = t_a + \epsilon. \tag{8}
\]

(The \( t_a \) is the travel time given by BPR 7, and \( \epsilon \) is a Gumbel distributed random variable).

The random variable captures the uncertainty of the travel time for a given route.

There are two general ways to model the statistical properties of the *SUE*, logistic regression (Logit) and probabilistic simulation (Monte Carlo methods). Monte Carlo methods have not been considered in this work.

6.2.2.1  *Logit stochastic assignment*

The general Logit structure for the probability of choosing a route \( k \) is given by the Logit equation 9:

\[
P_k = \frac{e^{V_k}}{\sum_{i=1}^{K} e^{V_i}}, \tag{9}
\]

where \( V_k \) is the utility of route \( k \) and \( K \) is the number of routes available for the traveller.
The general Logit formulation does not model correlation between different routes, so different correction terms have been developed to overcome this weakness (see [54]). In this work only the Path Size Logit (PSL) [54] has been considered, as the PSL shows good performance without being overly complicated. The PSL formula is given by equation 10:

\[
P_k = \frac{e^{V_k + \beta \cdot \ln(PS_k)}}{\sum_{j=1}^{K} e^{V_j + \beta \cdot \ln(PS_j)}}
\]

where a correction factor is added to the utility. This correction factor is called the Path Size correction factor for route k and is given in equation 11:

\[
PS_k = \sum_{i \in \Gamma_k} \frac{1}{\sum_{l \in C} \delta_{il}}
\]

In equation 11 \( \Gamma_k \) is the set of links in route k, \( \delta_{il} \) is the link-path incidence function (equal to 1 if link i is in route l, otherwise 0) and C is the choice set of routes.

The Path Size correction factor decreases the utility of a route if links in the route are used in other routes. The maximum of the correction factor is one, if the route is unique, thus the logarithm of the correction factor is less than or equal to zero. The factor \( \beta_{PS} \) has to be estimated for the given transportation network.

### 6.2.3 System optimum

The previous mentioned models for finding an equilibrium in the transport flows have the selfish behaviour of the drivers as the core principle. If the transport system is designed for the most efficient fulfilment of the travel demand, the transport system would be aiming towards minimising the mean travel time among all travellers. This is the definition of the system optimum.

A transport system operating at system optimum, where traffic flows throughout the transport network, might involve travellers having longer travel time than other travellers with the same origin and destination. The longer travel time of the few would mean shorter travel time for the rest, thus resulting in shorter overall mean travel time. Another way of looking at the system optimum is to avoid congestion by having a small number of travellers taking another route, or increasing the capacity of the most congested links in the transport network.
The system optimum can be calculated from the OD matrix, in a similar way as the DUE.

6.2.4 Restricted Stochastic User Equilibrium (RSUE)

The RSUE [57, 68] is a way of combining DUE and SUE. The problem with SUE is that every route in the transport network is assigned a probability, and throughout the solving algorithm all these routes have to be considered. By restricting the set of routes to a smaller set, the problem with the large number of routes with very low probability can be reduced. The number of routes in the choice set increases as unused routes with a low travel time is found.

As shown in [57] the cost function has to be transformed to allow for the mathematical equilibrium conditions of DUE and SUE to match.

6.2.4.1 Description of algorithm

In the following I will give an overview of the algorithm to solve the RSUE (in the MIN formulation). The algorithm for solving the RSUE consists of a preparation stage and an iterative assignment loop. In the preparation stage the algorithm finds an initial set of routes with a shortest path algorithm with the free flow travel times as weights. These routes form the start of the restricted set of routes, to be considered further in the algorithm. The chosen routes are referred to as the "choice set". The links are assigned traffic volumes by using the OD matrix and the found routes. Then new travel times are calculated for each link, based on the resulting traffic volumes (links can be shared by different routes, thus the traffic volume for a shared link is the sum of the traffic for each route sharing the link).

The loop consists of four steps. First the shortest travel time path from each origin to each destination is found, using a shortest path algorithm with the travel time of each link as weights on the vertices in the graph (Dijkstra or A* are both usable). If any of the shortest paths are different from the routes in the choice set and have a travel cost lower than lowest route in the choice set, the new routes are added to the choice set with traffic volume equal to zero.

The second step consists of balancing the traffic volumes between the different routes in the choice set, so that for each OD pair, the travel times for each route in the choice set covering that pair are equalised. Different ways of performing the balancing of traffic flow between the routes in the choice set exist (see [57]). In the implementation the pair wise swapping method was chosen, where the routes in a choice set is ordered...
by cost\(^1\). The routes in the ordered choice set are paired, so part of the traffic from the highest cost route is moved to the lowest cost route. Then traffic from the next highest cost route is moved to the next lowest cost route. The amount of traffic moved is determined by a swap factor calculated from the normalised difference of the costs. The swapping reduces the traffic of the routes which has the highest cost (which are the most congested routes) and adds to the routes with the lowest cost (least congested), leading to more equal cost between the routes. The balancing has to be performed for the choice sets for all OD pairs.

Step three consists of calculating updated link travel cost based on the new traffic volumes generated in step two.

The last step in the algorithm is a convergence measurement, for determining when the algorithm is close to convergence. The convergence condition to stop the algorithm is defined to be the minimal difference in costs between the different routes. In the RSUE the convergence condition is split in two. The first convergence condition measures the relative difference in cost between the equilibrated routes in the choice sets. The other convergence condition considers the cost of possible routes outside the choice sets (the unused routes). If there are unused routes with a cost close to the minimal cost in the equilibrated choice set, these routes might be chosen in a later iteration.

6.2.4.2 Case study: Istanbul

The implementation of the RSUE algorithm was created during a research stay at Sabanci university in Istanbul, which led to using Istanbul as a case study. Istanbul is a large city with more than 14 million inhabitants living on both sides of the strait of Bosphorus. The road network data, acquired from OSM, contains over 300 thousand roads, which are converted into 500 thousand one-way roads. The strait of Bosphorus has only three bridges (the latest opened in August of 2016), which creates natural choke points. The bridges are not the only places where congestion is occurring, as visits to the city will show.

The OD matrix was synthesised as traffic counts were not available. A number of places (46) was selected as origins (and destinations) and a constant demand was allocated to each origin destination pair as a test for the implementation.

The RSUE algorithm is implemented in Python, which connects to the database system to execute Structured Query Language (SQL) queries to perform the main calculations. The results of queries are often saved in new database tables to keep the data in the database domain.

---

\(^1\) The cost mentioned in this paragraph is not the real cost but a transformed cost, which is necessary to combine the DUE and SUE.
Listing 1: The code for the Dijkstra shortest path search

```python
def get_many2many(db, table, source, dest, reverse='false'):
    pg_kdijkstra = {
        "pgr_dijkstra('select id, source,
            target, dyncost as cost, reverse_cost
            from {table}', array{source}, array{dest}, {reverse}) ""
            .format(
                table=table, source=source,
                dest=dest, reverse=reverse))
    
    route = """DROP TABLE IF EXISTS temp;
    CREATE TABLE temp as SELECT start_vid as source,\
        end_vid as id1,node as id2, \
        edge as id3, cost \
        FROM {}""".format(pg_kdijkstra)

    path_cost_dijk = """drop table if exists new_paths_costs;
    create table new_paths_costs as select source,\
    id1 as target, sum(cost) as cost
    from temp group by source,target;"""

    cur = db.cursor()
    cur.execute(route)
    cur.execute(path_cost_dijk)
    db.commit()
```

In listing 1 an example of the Python code is shown. The code is a function to perform a shortest path search for all origins to all destinations. The Python function is called `get_many2many` and takes as parameters an open database connection `db`, a table name in the variable `table`, which contains the road network with updated link costs, lists of origin and destination node ids `source` and `destination`, and a reverse parameter to signal if the reverse cost should be considered in the search. The Dijkstra shortest path search is a function installed in the database by the `pgrouting` extension to PostgreSQL. The function is called from a SQL query, and the query is assembled in Python by formatting strings. The base query is in the `route` string variable, which is formatted with the Dijkstra function call in the `pg_kdijkstra` string variable. The result of the route query is a recreation of a table called `temp`, and no data is moved from the database to the Python interpreter. This step will be performed many times during the algorithm.

---

2 [http://pgrouting.org/](http://pgrouting.org/)
The map of Istanbul is shown in figure 9. An overlay in light green shows the roads that the algorithm has found to be used to accommodate the travel demand.

In figure 10 the error of one of the convergence measures is shown. The convergence is measured by the imbalance in costs of different routes in the choice set for an origin destination pair. When a new route with a low cost is discovered, the route is added to the choice set with zero flow. This causes an imbalance in costs for the routes and thus the convergence measure increases. As the traffic is balanced between the routes in the choice set, the convergence measure decreases.

6.3 TRAVEL DEMAND

As have been mentioned multiple times previously, the travel demand for a transport system is given by a matrix of origins and destinations. The traditional approaches to synthesise this matrix have been through traffic counts, traffic surveys and data from mobile phone cell towers [28]
52 TRANSPORT MODELLING

Figure 10: Convergence of RSUE algorithm. The error grows when new routes are added to the choice set. Notice the error axis is logarithmic.

6.3.1 Traffic counts

Traffic counts are performed at street level, typically over a period of weeks. The traffic counts measure the number of vehicles passing a particular point, but cannot capture information of where the vehicle originates from or where the destination of the vehicle is.

The OD matrix can be synthesised by iteratively generating OD matrices and using the generated matrices for assigning traffic in the transportation network. The differences between the measured traffic counts and the assigned traffic are then used to adjust the OD matrix.

6.3.2 Surveys

In traffic related surveys participants are interviewed, among other things, about their travel related needs and practices. From a representative sample of travellers in the transport system a transportation demand matrix can be synthesised, by extrapolating the answers obtained from the survey with respect to community size [77].
The survey can give information of both the origin, destination, route chosen, possible alternate routes known to the informant, travel frequency, preferred transportation, Value of Time and other very detailed information about transportation habits.

As the answers collected by the surveys are subjective responses to detailed questions about ingrained habits or questions informants are not used to reflect upon, there can be some unwanted side effects of using surveys. People might not remember the precise routes they are using, they might forget to mention important details, or the traveller has extraordinary travel needs. These problems can cause biases and a skewed traffic demand matrix.

6.3.3 Smartphone data

Smartphone data have the potential to give very accurate data on the traffic needs and traffic patterns of travellers in a transport system. The smartphones are carried by the users, and through the sensors of the smartphone a very detailed picture of the actual travelled routes can be obtained. The data used in this study are three-dimensional acceleration data from the accelerometer and positioning data from the GPS receiver, sometimes extended by positioning data from Wifi radio strength measurements.

6.4 Driving mode detection

To investigate if the smartphone data can be used to detect the driving style of a driver, an experiment was conducted. A test route was chosen and driven a number of times in different vehicles with a smartphone placed in the vehicle. The purpose of the tests was to determine if the accelerometer data can be used to detect the driving modes of the driver, i.e. if the vehicle was idle, accelerating, cruising, braking, turning left, or turning right.

The test route can be seen in figure 11. It is a 13 km long route comprising both urban and rural driving.

In figure 12 scatter plots of the accelerometer data for a trip is shown. From the figure of the acceleration in the XY plane it can be seen that, apart for some outliers, most of the points are located close to the centre where there is no acceleration in either direction. Around the centre cluster, four smaller clusters are visible. From the XZ and YZ plots it can be seen that when there is acceleration in the Z direction, there is no acceleration in the X or Y direction, and likewise when there is acceleration in X or Y direction there is no acceleration in Z direction.
Figure 11: Test route for driving mode experiment

Figure 12: Accelerometer scatter plots. On top are XY axes. Bottom left is XZ. Bottom right is YZ.

The previous analysis indicates that K-means clustering can be used to detect the different driving modes. In the experiment idle mode could be detected correctly in
more than 70% of the data points for ICEs and more than 99% for electric vehicles. The experiment lack ground truth detailed enough to adequately calculate the accuracy of the categorisation of the other driving modes. To convert the analysis into a working tool for driving mode detection, an app would need to perform a training step every time the smartphone is reoriented relative to the vehicle, to calculate the right cluster coefficients. When the application has been trained each data point can be categorised immediately.

6.5 CONCLUSION

In this chapter the basics of the field of route choice and how to create a transport model from the concept of user equilibrium, how to detect different driving modes, and how these driving modes impact fuel consumption was presented. The driving mode detection algorithm uses the accelerometer sensor to determine the driving mode, by using a simple K-means clustering algorithm.

A database implementation of the RSUE was also presented and a case study of the performance of the model on a large city (Istanbul with more than 14 million inhabitants) was shown.
Accurate positioning is necessary to be able to localise vehicle emissions to a street level emission model. In this part I discuss problems in positioning with smartphones and methods to overcome some of the difficulties.
7

POSITIONING

7.1 INTRODUCTION

Accurate positioning of vehicles is very important for modelling emissions from vehicles, especially for air quality modelling purposes. Smartphones offer an easy way to obtain real-time positioning, but also poses some challenges. The smartphone positioning subsystem uses a number of methods to calculate the position of the smartphone. The main part of the positioning system is GPS, the American satellite based system. The positioning subsystem also uses measurements of radio signal strength from radio transmitters in range, be it mobile phone towers or Wi-fi networks. The sensor fusion in the positioning subsystem of a smartphone combines the error patterns of the different positioning algorithms.

In this chapter I will discuss some of the error patterns seen in smartphone positioning, and explore methods to overcome the positioning errors of smartphones in road based transportation.

7.2 ACCURACY OF GLOBAL POSITIONING SYSTEM

When using a smartphone for recreational purposes it is not unusual to experience errors of the positioning subsystem. The positioning data can change rapidly for no obvious reason as witnessed by figure 14.

In figure 14, a running route in Florence, Italy is mapped with a running app on an iPhone 4S. Apparently, I should have crossed the river several times, without using one of the bridges and at one occasion went into the river for several meters. The ground truth for the track (shown in red) is that I followed the roads in Florence along the river and crossed the river twice using two of the bridges. The trace shows how large inaccuracies suddenly appear in the recorded trace. The first two kilometres of the trace are not very accurate but recognisable (the trace follows mostly the right roads, although the trace misses the used bridge). After two kilometre there is an abrupt change in the trace, as if the smartphone has lost track of the necessary satellites to achieve accurate positioning. The possibility for such an interruption of the positioning
service is part of the reason for the investigation presented in this chapter, the other reason is to investigate the more general expected accuracy of smartphone positioning.

7.2.1 Stationary experiment

The GPS is a space based positioning system that works by triangulation from a point on earth to three or more satellites, in medium earth orbits approximately 20200 km above the earth surface. In this orbit each satellite passes the same earth position twice a day. To investigate if the handover from vanishing satellites to appearing satellites has an effect on GPS accuracy, an experiment was performed.

The purpose of conducting the experiment was also to investigate the influence of weather, specifically if rain would have an influence on the GPS accuracy, and thus explain the sudden GPS accuracy deterioration. The hypothesis was that water molecules in the atmosphere could dampen the radio signals from the GPS satellites. Meteorological data was acquired from the Danish Meteorological Institute¹ to test this hypothesis.

¹ http://dmi.dk
The experiment consisted of a smartphone placed at the same place for three weeks and the reported location coordinates and self-reported accuracy was recorded. The experiment was split into two periods - one period the positioning data was recorded with the Wi-fi radio activated, and a second period where the Wi-fi was deactivated to investigate the influence of the combined Wi-fi and GPS positioning versus the GPS positioning alone.

In figure 15 the self-reported location of the iPhone 4S that was used in the experiment, is shown. In the left-hand figure (a) is the location with Wi-fi enabled and in the right-hand figure (b) the location reported with Wi-fi disabled.

From the figure 15 it can be seen that the maximum location error does not differ significantly between the two plots. The plot with Wi-fi enabled seems to be more spread out in specific directions, whereas the GPS only plot is more centred with a number of outliers.
In figure 16 the Wi-fi enabled locations are shown with longitude as X-axis, so north is up. In figure 17 the error directions are shown on a aerial photo of the location of the experiment with red lines. The red lines mostly match the directions to neighbours, who have Wi-fi networks which can be received from the smartphone at the site of the experiment. The neighbouring Wi-fi networks have been identified by their SSIDs.

The conclusion of the experiment with respect to the location accuracy when Wi-fi is combined with GPS seems to be, that for stationary location it might be more precise to only rely on GPS. It is surprising that Wi-fi positioning can add errors in directions toward the Wi-fi transmitter. The reason could be that the Wi-fi positioning is determined by triangulation of the Wi-fi radio signals from known Wi-fi SSIDs. The strength of the received Wi-fi radio signal is dependent of, amongst other things, the humidity [73]. The varying humidity in the spring season could cause variations in the received strength of the Wi-fi radio signals and could thus explain the variations in locations in directions towards the neighbouring Wi-fi radios.

To investigate the weather influence on GPS positioning local rain measurements for the site of the experiment, was requested from the Danish Meteorological Institute. The rain data was compared with the self-reported location accuracy of the smartphone location subsystem, but no correlation was found.

To investigate the hypothesis of accuracy degradation from handover between satellites vanishing below the horizon and satellites appearing from below the horizon, the
self-reported accuracy was visualised in a three dimensional plot with the hour of day at Y-axis, the date at the X-axis, and the accuracy (in meters) at the Z-axis. The reported accuracy is the radius of a circle within which the true position is located. The higher accuracy number the larger uncertainty of the true location. The plot can be seen in figure 18. Data with an accuracy of less than 20 m has been removed, as the focus of the experiment is the large errors that can explain a sudden disruption of the location service.

From the plot it is not possible to see any correlation between time of day and large values of the reported accuracy, either with a 12 hour view or a 24 hour view. There is no indication of sudden loss of positioning accuracy can be attributed to the handover from satellites disappearing below the horizon to satellites appearing.

### 7.3 Improving Positioning Accuracy from GPS

#### 7.3.1 Using alternative positioning systems

The American GPS system is not the only space based positioning system. The Russian GLONASS system, the Chinese COMPASS system, and the future European GALILEO system offer similar positioning services as the GPS system, and in theory the combination of these positioning systems should offer better accuracy. The iPhone 4S used in the previously described experiment was the first iPhone model using the GLONASS positioning service as well as the GPS system, testifying that improvements in position accuracy are still needed. It is not known how Apple has implemented the GLONASS support. It might not be used as a vehicle for improved accuracy but merely exists as a fall-back solution if the GPS positioning is failing.
7.3.2 Delta GPS and Sensor fusion

To improve the accuracy of the GPS positioning an often-proposed method is differential GPS. Differential GPS utilises an extra radio transmitter located at a precisely known location. This local signal is adequate to improve the positioning accuracy to centimetres.

The accuracy of the GPS positioning can be improved by considering the data from the accelerometer through Kalman filtering. The Kalman filter combines the determinism of differential equations governing movement with noisy position measurements, and determines the most likely true location.

7.3.3 Map-matching

For applications where the location data is acquired from transport, map-matching can often be used to remove the positioning error. The idea is to move GPS located measurement points which are not placed at a road according to a map to a point on an existing road.

To use map-matching, the data set is filtered to remove data from other activities than road transport in motor vehicles. In figure 1 an example of a GPS trace, which is received from a user using an EcoSense app in a sports airplane. Such traces must be removed before map-matching is performed.

A number of different methods for map-matching exist, with different processing demands and different failure modes. In this section, I will briefly present the ideas behind the two methods I have considered.

7.3.3.1 Geometric map-matching

The simplest method to match a GPS point to a road segment is to choose the nearest point on the nearest road. This method is called the geometric map-matching approach. This works well as long as the roads are widely separated, and less well when roads are close and parallel or at intersections.

7.3.3.2 Topological map-matching

To reduce the problems of choosing which road to move the GPS point to, previous (and in offline scenarios subsequent) points are considered to inform which road is the right one to choose. In the paper Brakasoulas et al reports two different approaches to
7.3 Improving Positioning Accuracy from GPS

Topological map-matching, an incremental map-matching which only considers a few measurement points around every GPS point, and a global method to find the route in the road network which is most similar to the GPS trace. To compare the different approaches a quality measure based on the Fréchet distance is developed.

To match the location of the data acquired from the EcoSense data set, the data set is first segmented into trips (see 7.4). The trip data is then filtered to remove trips which are not related to travel by car, by removing very fast trips (airplane flights) and very slow trips (walking, running and bike riding).

The remaining trips are map-matched by a topological method. The map data is acquired from the OSM project, and converted into a topological PostGIS 2 database table by the OSM2PO 3. The OSM2PO converts OSM roads (called "ways" in OSM parlance) into a proper graph, so each row in the database table corresponds to a link in the road network. This is needed as OSM stores streets spanning multiple nodes as a single row, which does not lend itself to tasks such as routing, which is an important part of topological map matching.

The topological street data is converted into one-way streets in order to use the street direction in the map-matching algorithm. In OSM (and thus in the converted dataset) all streets have a cost value to describe the time it takes to traverse the link from the source node to the target node, and a reverse_cost value for traversing in the opposite direction. One-way streets are denoted by either a very large cost or a very large reverse_cost. To convert a two-way street into two one-way streets it is necessary to locate the two-way streets by searching for street segments where the reverse_cost is less than double the cost value of the segment 4.

A new row is added to the table with the source and target nodes interchanged, the cost value of the new row is set to the original reverse_cost of the two-way segment and the reverse_cost field is set to a very large value. Finally the original road segment’s value for the reverse_cost is changed to a very large value.

The location data of the trips are matched to the road network by first locating the nearest street segment for each location. This is a quite costly procedure, since the database implementation goes through all street segments for each location point to find the nearest map segment. When all road segments of a trip is located, there will be a number of falsely detected segments, that is segments which are not part of the travelled route. This happens due to accuracy problems from the smartphone positioning system. The problems arise at intersections, where the nearest road segment to a point can be one of the not taken roads. The misdetection can also be at parallel

---

2 http://postgis.refractions.net
3 http://osm2po.de
4 The relation of forward cost to reverse cost is chosen arbitrarily to be less than two for two-way street and larger for one-way streets
Table 2: Data collected in Herning

<table>
<thead>
<tr>
<th>Number of data points</th>
<th>1.2 million</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>17</td>
</tr>
<tr>
<td>Number of trips</td>
<td>2179</td>
</tr>
<tr>
<td>Number of flights</td>
<td>11</td>
</tr>
<tr>
<td>Number of slow trips (less than 25 km/h)</td>
<td>1251</td>
</tr>
<tr>
<td>Number of map-matched trips</td>
<td>934</td>
</tr>
</tbody>
</table>

road segments, which are close enough for the inaccuracy to falsely indicate a location point closer to the not-travelled road.

A way to reduce the amount of work the database query has to do is to only consider location points corresponding to a moving vehicle. This can be achieved by excluding points where the vehicle is in idle mode, or where the smartphone user is not engaging in a travel related activity.

7.4 TRIP DETECTION, SEGMENTATION AND ORIGIN-DESTINATION GENERATION

The acquired data from smartphones is stored in a central document database (MongoDb) \(^5\). The data is sent from the smartphone in packets of observed data, and also stored in the database as multiple observations from one smartphone (client) in one document. To convert the data from the packet format convenient for communicating with the server into a trip based data format, the data is transformed from the document database to a SQL database (PostgreSQL) with one timestamped observation per row.

The data in the SQL can be visualised, as seen in figure 19, by plotting the observed GPS points on a map. In figure 19 is also shown the maximum speed aggregated over points inside a 1 km by 1km grid. The observed points are coloured according to the smartphone client reporting the location.

Trip detection is performed by searching for discontinuities in the timestamps of the observations. If the difference between two timestamps for the same client is larger than five minutes, a new trip is flagged. The trips are then filtered to remove trips that are not related to vehicle road transport. The filtering is performed with regards

\(^5\) http://mongodb.com
to the maximum speed of trip, and the maximum altitude reported for the trip. Low maximum speed trips (below 25 km/h) are probably related to walking, running, or biking, and high speed and high altitude trips (speed above 150 km/h and altitude above 300 m above sea level, Herning is located at approximately 100 m above sea level) are related to flying.

The remaining trips are map matched to the road network obtained from the OSM project. The map-matching algorithm used, is a topological algorithm, where knowledge of the road network is used to choose which roads to match the GPS observations to. First the nearest road from each observation is found with a nearest neighbour query. The roads found by the nearest neighbour function are filtered to exclude road segments which are not part of a route, by searching for segments that are connected in both ends of the segment. The remaining segments are connected and form the route for the trip except for the first and last segment, but for OD matrix generation only the general part of the city is needed, not a particular address. The exclusion of the first and last segment of the route also ensures the privacy of the participants in the experiment. This approach works well on the acquired data because of the high sampling rate of the data (approximately 1 Hz).

In figure 20 the result of the filtered and map-matched GPS traces is shown.

---

6 The query function is adapted from https://gis.stackexchange.com/questions/14456/finding-the-closest-geometry-in-postgis#14457
Figure 20: GPS trip data map-matched to a road map of Herning. Each trip has a different colour, but links shared by more trips only show the top colour.

The OD matrix is derived from the filtered and map-matched trips, by assigning transport demand to an OD pair from the distance between the start and end points of the route and the population size the part of the city the route starts. A simple power law can be adopted to estimate the demand from the distance, but for larger cities an exponential law might be more appropriate [42].

7.5 CONCLUSION

In this chapter accuracy of the positioning subsystem of smartphones was discussed. The position accuracy has been investigated for stationary use of smartphone indoor positioning. In [39] a discussion of the self reported accuracy of moving smartphones, which concludes that the self reported accuracy varies very much.

Using position data from smartphone to reliably generate traces of routes taken by travellers was discussed.

An algorithm for generating an OD matrix, was presented. The idea of using the routes found in GPS traces as the initial restricted route choice set in the restricted stochastic user equilibrium was detailed.
The main scientific contributions of this dissertation are:

- Algorithm for origin destination demand matrix creation from smartphone data [38].
- The development of a novel map matching algorithm suitable for a database [38].
- Using user experienced routes as a seed for a transport model [38].
- Driving mode detection from smartphone accelerometer data [41].
- A performant database implementation of Restricted Stochastic User Equilibrium transport model [40].
- An investigation of the accuracy of Global Positioning System in a stationary smartphone [39].
- Improve air quality models by using smartphone data [41].

8.1 CLIMATE GAS EMISSIONS

Even though there has been discussion about anthropogenic climate change for the last 50 years, the threat of massive changes in climate has not been mitigated. There have been active and focused efforts to reduce the emission of climate forcing gasses from the countries which have the highest emission per capita, but still we see the CO₂ levels rise unabatedly [72]. One of the areas that has seen the least effect of the mitigation efforts is the transportation sector. The vehicle engines have become more efficient and (especially in Europe) the engines of passenger cars have become smaller. These changes has been offset by an increase in the number of cars and the amount of travel each car make each year, thus the amount of emissions from transportation has continued to increase. The only sign of a future decline in the emission of climate gasses from the transportation sector, is if the motor technology changes to a more effective type, such as electric motors (battery or hydrogen fuel-cell powered) or combustion engines driven by sustainable fuels. The change of motor technology is going
to be slow and is not expected to have a large impact before late in the next decade \(^1\). Because of this, work on modelling the emission of climate gasses from transportation will still be relevant in the next ten years.

### 8.2 Air Quality Modelling

Heavy traffic of ICE vehicles is responsible for much of the air pollution in urban environments. As stated above this is unlikely to change drastically in the next decade, due to increasing demand for transportation [22]. The pollution from traffic in expanding third world cities is going to be especially challenging, as the compounding factors of influx of people to the city and growing income, means growing transport demand over longer distances coupled with growth of a more wealthy middle class demanding private transportation in passenger cars. Even drastic initiatives such as congestion charges in city centres have only had marginal effect on the air quality [7, 35].

The above makes the work in this thesis with regards to modelling of street level pollution in potentially large cities relevant, and will remain relevant for the coming years.

### 8.3 Transportation Modelling

The transportation models discussed and implemented in this work, were necessary to reach the goals of the two above mentioned sections, emissions of climate change gasses and air quality modelling. In this thesis an implementation of the RSUE is shown to be applicable to a large city. As presented in this thesis, the use of smartphone data is a novel way to create the travel demand matrix, that can supplement the traditional methods, traffic counts and traffic surveys.

Driving mode detecting from smartphone positioning and accelerometer data, was performed with the use of a simple K-means clustering approach.

### 8.4 Positioning

The accuracy of the positioning system of smartphone has been investigated through a couple of experiments. Prompted by observations of sudden deterioration of GPS accuracy in recreational use of smartphone location tracker, an experiment of how

---

\(^1\) see http://ec.europa.eu/eurostat/statistics-explained/index.php/Passenger_cars_in_the_EU for statistics of registrations alternative fuel vehicle in EU (2015 data)
stationary smartphone position accuracy developed over time, was performed. The experiment did show variations in accuracy, but the variation in accuracy could not be explained by weather data or satellite orbit movements\(^2\). The experiment did show that using Wifi as part of the positioning system changed the variations in accuracy, in such a way that the error was more expressed in directions toward neighbouring Wifi radios, whereas the error did not have preferred directions when position only relied on GPS.

8.5 Real life experiments

When you as a researcher ventures outside the laboratory, and make experiments that are close to real life, unexpected data and behaviour will occur. While obtaining the acquired data, the expectation was to have to filter the data set into different terrestrial transportation modes. It took a while to recognise that the EcoSense data set also includes travels by amateur aviators.

8.6 Impact

The research completed during the Ph.D study can be of use for a variety of problems. In my opinion, the practicality of the implemented transport model for use in very large cities could have the biggest impact. Apart from using a transportation model to estimate emissions from vehicles for climate change gas emission reporting and air quality monitoring, the model can be used in urban planning to predict and localise congestion in connection with urban growth, large road works, and development of new residential or industrial areas. Furthermore, the model can be used to evaluate different congestion minimising proposal, i.e. new highways, adding lanes to existing highways, buildout of public transport and route planning for public transport.

There are possibilities for adapting the results of this dissertation to automate parts of the transport related green accounting. If owners of vehicle fleets, installs a GPS sensor in the vehicles the total emission of the fleet could calculated from the received positioning data.

\(^2\) The orbital movement of GPS satellites is synchronised to earth rotation in the semi-synchronous orbit. Any accuracy variation from the movement of satellites should have a recurring property, which was not found in the data.


8.7 Future research

8.7.1 Transportation modelling

Traditional transport modelling use traffic counts as the foundation for estimating the traffic demand. Transport modelling then assumes that travellers will choose the shortest route in terms of travel time to satisfy their transport demand. Previously, knowledge about travellers observed route choice was based on user surveys [55], but by using acquired GPS traces from real travelled trips, the actual travelled routes are revealed. Instead of starting the transport model by using shortest path algorithms to create the initial route choice set, the proposal of this dissertation is to use routes found in smartphone GPS traces as seed for the initial choice set.

The database implementation of the RSUE is not well optimised. When the model was implemented and performed in less than expected time, no further work was performed in order to further optimise the runtime. Using recently added functionality to PostgreSQL and the pgrouting library can reduce the runtime considerably. Intelligent application of indexes in the created tables might also be beneficial, but this needs careful investigation as the indexes would have to be recreated often, and the time saved for looking up data, could be offset by the time it takes to build the indexes.

Application of trajectory mining [76] to find recurring use of routes and further explore behaviour of travellers in the transportation network, might lead to new insight into how habits govern route choice, and give idea for Intelligent Transport System (ITS) on how to decrease congestion.

8.7.2 Emission modelling

The advent of a significant part of the road transport demand to be satisfied by BEVs, will need some careful adjustments to the existing emission models. The emissions related to BEVs is not local as it is associated with charging of the battery. The emissions related to charging of a BEV depend on the source of the electricity used in the charging process, which can be from sustainable sources with only no or small emissions (solar, hydro, wind, or batteries powered by solar or wind) or from conventional sources with large emissions 3. In the end it might be that emission from BEV transport will have to be moved from the transportation part of the climate gas inventory reporting to the energy part.

---

3 I have excluded nuclear and other sources with small emissions due to their diminishing or small influence.
Driving mode detection might possibly be improved by applying more advanced machine learning techniques to the problem. The K-means algorithm I have used in this study is quite simple and other methods as Kalman filter, support vector machine, or deep learning networks can provide better detection of driving mode.

8.8 summary

I have combined research from different fields to find solutions for modelling air quality and emissions of climate gasses from transportation. The fields of research I have drawn upon are:

- Air quality research.
- Climate gas emissions from transportation.
- Transport modelling.
- Computer Science.

The research contributions of my work are:

- Algorithm for origin destination demand matrix creation from smartphone data [38].
- The development of a novel map matching algorithm suitable for a database [38].
- Using user experienced routes as a seed for a transport model[38].
- Driving mode detection from smartphone accelerometer data [41]
- A performant database implementation of Restricted Stochastic User Equilibrium transport model[40].
- An investigation of the accuracy of Global Positioning System in a stationary smartphone [39].
- Improve air quality models by using smartphone data [41].
Part V

APPENDIX
Using Crowd Sensed Data as Input to Congestion Model

Anders Lehmann*, Allan Gross†
*Department for Computer Science
University of Aarhus
alehmann@cs.au.dk
†Department of Business Development and Technology
University of Aarhus
agr@auhe.au.dk

Abstract—Emission of airborne pollutants and climate gasses from the transport sector is a growing problem, both in industrialised and developing countries. Planning of urban transport system is essential to minimise the environmental, health and economic impact of congestion in the transport system. To get accurate and timely information on traffic congestion, and by extension information on air pollution, near real time traffic models are needed. We present in this paper an implementation of the Restricted Stochastic User equilibrium model, that is capable to model congestions for very large Urban traffic systems, in less than an hour. The model is implemented in an open source database system, for easy interface with GIS resources and crowd sensed transportation data.

I. INTRODUCTION

Transport contributes to the emission of climate gasses, i.e. in Denmark transport related climate gas emissions are 24% of the total emissions in 2014 [11]. To enable better tools for urban traffic planning, both in terms of environmental, health, and economic impact of congestion, we have investigated how to model congestion in road transport systems. To improve the accuracy of the existing models, we propose to use crowd sourced data, collected from smartphones of users travelling in the transport system. The data collected will inform our model of travel demand and preferred routes.

In this paper we present a congestion model for Urban traffic systems implemented in a relational database system. The input to the model is Geographical data of an urban road system, and travel demand data gathered from Smartphones and traffic counts. The implementation of the model in a relational database system allows for integration to Geographical Information System (GIS) data, and shows reasonable performance for even large Urban transport system. An example of the performance shows results for modelling congestion in Istanbul, a city of more than 14 million inhabitants. We propose to extend existing Route Choice Models in two ways. Firstly since the collected data gives detailed information on chosen routes, we propose to use the observed routes as a start choice set for the Route Choice model. Secondly we propose to create the origin destination demand data from users of the transport system, whose smartphone app reports their travel behaviour [18]. Since we cannot get a complete data coverage of a certain transport system, we propose to combine this approach with existing for creating the origin destination matrix, such as traffic counts and surveys.

The map data for constructing the road network is provided by Open Street Map, a Crowd sourced GIS data provider.

The database system is chosen to be the open source database PostgreSQL, with the extensions PostGIS, for handling geographic data from map providers, and the extension pgrouting, for routing.

In figure 1 are screenshots of a pair of the smartphone apps, which have been used for the data collecting experiments. The one pictured left is "Elbil parat" (are you Electric Vehicle ready), the one on the right is from "Herning cykler til Månen" (Herning bikes to the moon).

In section two an overview of related scientific work is given. The congestion model and the implementation is discussed in section three. Section four presents the results of using the model for the megacity of Istanbul. Section five describes lines of further research.

Fig. 1. Screenshots of data collecting smartphone apps
II. RELATED WORK

The research presented in this paper combines two different scientific fields, the field of crowd sensing for creating demand data, and the field of route choice modelling from transport research.

In [5] the authors describe how to convert GPS data into trips and routes. This work use the calculated routes to predict current trips. By measuring the similarity of an ongoing trip with prerecorded routes, the most probable route for the current trip can be estimated.

For low coverage of traffic by GPS trackers [6] conclude that congestion can be predicted even with a small number (2-3% of total number of travellers) of speed reporting devices.

An overview of data collecting methods for origin destination matrix estimation is given in [4]. The data collecting methods can be divided into self reporting methods (surveys, personal GPS receivers with self reported trip classification) and automatic methods (traffic counts, cell phone data, bluetooth readers, license plate recognition). The paper do not establish methodologies for converting the collected data into origin destination demand matrices.

Origin destination demand matrices can be derived from traffic counts as is shown in [19]. The derivation involves solving a Generalised Least Square problem.

To create an origin destination demand matrix from information from mobile phone cell towers is used in [20]. The paper used a Horvitz-Thompson estimator to estimate the OD matrix. This work assumes that each cell phone trace generated from cell phone towers data is equal to a vehicle trace. The trace is made by triangulating between the connected cell towers. The authors take the cell phone market penetration (87% of the population in the researched area has a cell phone) into account and use the estimator to estimate the missing population.

A similar approach is presented in [16], where cell tower id’s are used as a location identifier, in a project to verify the use of mobile phone data as a complement/ replacement of household surveys and traffic counts for generating origin destination matrices in a highway planning scenario.

In [7] an indirect approach to creating origin destination matrices is presented. The paper analyses social media messages, that incorporate location and time information. The paper uses Foursquare data from the Austin,TX region. The data show good correlation with survey based estimation of origin destination matrices.

An example of using smartphones to track transit vehicles can be found in [2]. The goal of the paper is to use the tracking to provide arrival time information to waiting transit passengers. The authors applies map matching to the tracked GPS data in order to correctly geolocate the transit vehicles.

In our project we plan to get information on traffic demands from travellers via apps installed on their smart phones. The data from the smartphones allows us to know the origin and destination of the route, but also the actual route which the traveller chose. The challenge from the collected data is to estimate the total traffic demand from the observed travelled routes.

Modelling of Route Choice was founded in the 1960’s. The basic textbook is written by Sheffi in 1985 [15], and is still a preferred textbook for teaching.

The model for congestion used in this paper is presented in detail in two papers [17] [14] by the inventors. This work presents two different implementations, one in MATLAB and one in ArcGIS (with extra extensions for route choice modelling). The authors report running times in the ArcGIS implementation for modelling the Copenhagen region, of about an hour on a normal sized desktop computer.

The paper [14] presents experiments using different error components to model reuse of links in different routes to show the versatility of the model, and performance impact of the different link reuse strategies.

From the review papers [13] and [12], we have chosen to use the Path-size Logit method to model the influence of routes with overlapping links, due to good runtime and model performance.

The previous traffic modelling research, as reported, has only been tried in small to medium sized cities [9][8][3]. Our implementation makes it feasible to model the largest cities in the world as we show in the results section for Istanbul (21st largest city 1).

III. CONGESTION MODELLING

The effect of congestion in road transport, is primarily that the travel time of a congested road segment increases as the traffic load approaches the traffic capacity of the road segment. The travel time increase can be modelled with the BPR (Bureau of Public Roads) [15] formula:

\[ t = t_0(1 + \alpha \left( \frac{f}{C} \right)^\beta ) \]

Where \( t_0 \) is the free flow travel time, \( f \) is the volume of traffic (traffic flow), \( C \) is the capacity of the road. The constants \( \alpha \) and \( \beta \) are country specific numbers capturing the way drivers react to congestion (the values of \( \alpha = 0.5 \) and \( \beta = 4 \) yield results that correlate well with observations). As can be seen from the formula the travel time will stay at the free flow travel time until the flow is very close to the capacity. The travel time then increases rapidly as the flow increases above the capacity.

To model where congestion occurs, we are using the Route Choice Modelling framework. The Route Choice Model is based on the assumption that all travellers in a transport system, are choosing the route in order to maximise the utility of the travel (or minimise the cost of their travel). The cost of the travel consists of actual costs for fuel, toll and wear of the vehicle, and perceived cost which is modelled as a Value of Time (VoT), associated with the travel time. The Route Choice Model seek the equilibrium state, where all travellers travelling from the same origin to the same destination has the same low travel cost. In this state of the traffic flow any change in choice of Route will increase the travel cost.

1https://en.wikipedia.org/wiki/Megacity
For the Deterministic User Equilibrium, the assumption is that all travellers have perfect knowledge of all routes and their associated cost. Each traveller will then choose the route with the lowest cost.

In the Stochastic User Equilibrium model, each traveller thinks that they choose the route with the lowest cost, but does not have perfect knowledge, and therefore there is a probability, that the chosen route is not the one with the lowest cost. This method leads to algorithms that consider a large number of routes with very low probability of being chosen, since all paths have to be considered.

The Route Choice in the Stochastic User Equilibrium is modelled by considering the travel times on the different routes and adding a error term, which models the uncertainty of the travel time for the route. A problem in the formulation of the SUE is that links can be shared between routes and this reuse of links must be taken into account. There has been proposed a large number of ways to model the reuse of links in the Route Choice literature [13]. The probability of choosing a the $k^{\text{th}}$ route can be expressed by the Path Size Logit formula (2).

$$P_k = \frac{\exp(V_k + \beta_{PS} \ln(PS_k))}{\sum_{i \in C} \exp(V_i + \beta_{PS} \cdot PS_i)}$$

In equation2 $V_k$ is the utility of the $k^{\text{th}}$ route, $PS_k$ is a path size correction factor measuring the amount of reuse of links in route $k$. The $\beta_{PS}$ is a number controlling how much the Path Size similarity of the routes should affect the utility. The $PS_k$ is given by :

$$PS_k = \sum_{a \in \Gamma_k} \frac{L_a}{L_k} \frac{1}{\sum_{i \in C} \delta_{ai}}$$

In 3 $L_a$ is the length of the link $a$, $L_k$ is the length of route $k$, $\Gamma_k$ is the set of links making of route $k$, $C$ is the set of all links to chose from and $\delta_{ai}$ is one if $l$ equals $a$ or zero otherwise. The number $PS_k$ is one, if none of the links in route $k$ is shared with other routes, and less than one for routes with shared links.

The Restricted Stochastic User Equilibrium [14] combines the two mentioned equilibrium formulations by restricting the SUE to only consider a restricted number of possible routes.

The necessary data to model congestion with Route Choice is a digital road network, where each link (edge) is specified with free flow travel time and capacity. Further more an origin destination demand matrix, which specifies the travel demand in the network is needed.

In this paper the road network is created from the Open Street Map data, by converting the GIS date into a topology, to ensure that the routing functions can be used.

The demand data can be obtained in a number of ways. The use of surveys to create the demand data has been used [9], as well as interviews combined with traffic counts [8]. We propose to use smartphone apps to facilitate creating the demand data. By having users participating in the generation of the demand data, the hope is to get more accurate and up to date information.

The users have to install a smartphone app, which will record GPS and accelerometer data and send the data to be stored on our servers.

Since it is not possible to get a complete coverage of travel data, for a specific area, due to not all travellers participating in the data gathering process, methods for generating the Origin Destination matrix from a small number of respondents, needs to be created.

For Istanbul, where we were not able to deploy any smartphone apps, we used synthetic data. By using demographic data on the population densities of the municipalities of the Istanbul megacity, origins where chosen as single points near major roads, in the different municipalities. The destinations were chosen at the same points, and the origin destination matrix was created as a symmetric matrix with zeroes in the diagonal. The demand level were chosen to be the same for each origin destination pair.

For the danish cities of Aarhus and Herning we have collected data from participating users. In Herning users would use an app to measure their bike travels to participate in a municipality driven contest of "Biking to the moon". The sum of all bike ride in the municipality should reach the distance to the moon. As an aside users would also give the researchers data from their other travel activities. The data from Herning is divided into single trips and the trips are truncated at the start and end to the closest traffic junction, to anonymise the user. The trips are then grouped by start area and end area, to create the origin-destination matrix. Within each origin-destination group the trips are sorted in time buckets to find rush hour patterns. To account for the incomplete coverage of data, the origin-destination data thus created has to be complemented with other data sources as census data, traffic counts and household surveys.

In Aarhus a project called "Are you electric vehicle ready" 2, participants first use our app to register the driving demand. After a month use of the app participants can then borrow an electric car for a month to see how it is to drive an electric vehicle. The data from the app is used both to grade the suitability of an electric vehicle, but also for our research purposes.

The data collected from our crowd sensing apps, can be used to examine qualitative properties of the transport system. Examples of what low coverage data can be used for are:

- The origin and destinations of the collected trips are realised travel routes, and can be used as a seed for generating the origin destination demand matrix.
- A speed much lower than the speed limit, will indicate a congested link in the road network, and we can use occurrences of this low speed pattern as a check, since the model also must predict that the link is congested.

The crowd sensed data also presents actually chosen routes, and thus these routes should be present in the chosen routes in the model. The gathered data is biased towards passenger traffic due to the way participants are recruited. To persuade users to use the apps there must be a benefit for the user. This perceived benefit will create a bias in the group of people choosing to use a data providing app. As an example the app "Herning cykler til månen" (Herning bikes to the moon), will have a bias towards people who likes to use their bicycles. The car driving data we receive from these participants, might not be representative for the population of Herning.

The low coverage is a problem, for estimating travel demand. On one hand the data is very accurate and detailed. The origin and destination of a trip is clearly discernible, as is the chosen route, but the low number of users, does not provide enough data to generate the origin destination demand matrix.

To create a origin destination demand matrix we propose to use traffic counts on links in the road network as described in [19], where we use the GPS route traces gathered from the smartphone apps, as constraints for solving the Generalised Least Square problem.

In figure 2 the traffic counts are visualised for the municipality of Aarhus. The width of the coloured roads relates to the traffic counts (thicker lines equals more traffic). The colours signifies whether the traffic count is measured or estimated. The green colour are the measured traffic counts and the red colour are for the estimated traffic counts.

IV. RESULTS

All model simulations were run on a MacBook Pro 15" with Intel 2 GHz Core i7 processor. The database system was chosen to be PostgreSQL version 9.4.1, with the extensions PostGIS and pgrouting. The road data for the model is from the Open Street Map project (OSM). OSM is a crowd sourced map database. Anyone can create a user and start create or update map data. When data is added to the database the committed changes will be made visible at the next update, and other OSM users and automatic rule checkers will review the changes. OSM provides several different ways for add and retrieve map data.

To convert the data from OSM to a topology importable by PostgreSQL, the OSM2PO was used. The Istanbul road network consists of over 300000 bidirectional road segments. These segments are expanded to unidirectional arcs to make sure that the routing algorithm does the routing adhering to normal traffic rules.

The result of the route choice model can be seen in figure 3. The map only shows the road network of Istanbul, but it is quite easy to recognise the topography of the city. The Bosphorus strait is visible in the center with the two bridges connecting Europe and Asia. The white spots in the map are mountainous areas, with only a few roads. The overlay with darker color is the routes chosen by the model to serve the origin destination demand matrix. The demand is simulated, by choosing 46 points in Istanbul as origins. The points were chosen at central locations in different parts of the city, where the traffic could easily be dispersed without creating congestion at the origin. The points represents the traffic demand from an area around the chosen point. Each origin creates traffic demand to all other origins, with a constant demand, to create an origin destination matrix with 2070 non zero entries.

1http://www.postgresql.org
2http://www.openstreetmap.org
3http://osm2po.de
The runtime of the algorithm is visualised in figure 4.

The runtime of an iteration increases as the number of iterations increases. The main contributors to the runtime are the shortest path searches (we are using the \textit{kDijkstra} function from pgRouting), and the test to see if a new found shortest route is already in the set of possible routes. This test compares the hash value of the new route to the hash values of already found routes. As the number of routes increases the number of tests for determining if the found shortest routes are already in the choice set also increases.

There is a possible performance improvement by adding indexes to some of the tables in the database.

In figure 5 the convergence of the algorithm is shown. The error is a measure of the differences in travel times for the different routes for each origin destination pair.

The x-axis is the number of iterations, the model has been run through, and the y-axis is the logarithmic view of the error.

The figure shows how the error decreases fast in the initial steps of the algorithm. The error further decreases as the algorithm progresses. The sudden increases in the error is when a new route is found, with a lower travel time. The route is added to the set of routes without any traffic assigned. The traffic distribution for that origin destination pair is thus unbalanced, and hence contribute with a large error. As the traffic volumes for the different routes are redistributed, the travel times of the routes move closer to each other, thus reducing the error.

V. Further Work

To further improve the congestion model, we plan to expand the model to more kinds of traffic, by integration commercial transport of goods, and public transport into the traffic mix.

To create the origin destination demand matrix from the collected data, we want to create a stochastic model for generating the traffic demand, guided by knowledge gained from local observers, census data, transit data, and crowd sensed data.

Including turn delays and intersection modelling [10] would further increase the accuracy of the model.

There is a possibility for improving the runtime of the model, by utilising a new feature of pgRouting. With this feature the shortest path for the origin destination matrix can be found in a single query to the PostgreSQL database system. As implemented in this work the shortest paths are found for one origin to all destinations at a time. This forces the pgRouting function to rebuild the graph of the road network for each origin in each iteration.

VI. Conclusion

In this paper we have presented the efficient database implementation of a congestion model based on the Restricted Stochastic User Equilibrium algorithm. We propose to extend this model by adding routes derived from crowd sensed data to the starting choice set, instead of relying solely on shortest path algorithms.

Using a database has been a fortuitous choice, since the performance of the database has resulted in acceptable run times of the algorithm, without having to look for advanced performance improving techniques. This also means that it is highly likely that the performance can be drastically improved, by a careful examination of the queries used.

The congestion modelling can be used for a number of purposes: for transport and urban planning, the model can be used to show effects of new roads, planned road works, or increased traffic. The model result can also be used as input to Environmental and Air Pollution models [1], to further increase the accuracy of such models.

ACKNOWLEDGMENT

We want to thank professor Bülent Çatay, Sabancı University, Istanbul, Henrik Blunck, and Niels Olof Bouvin for their generous support.

This work has been supported by The Danish Council for Strategic Research as part of the EcoSense project (11-115331), and the BIOGAS2020 project supported by the European Regional Development Fund for Öresund-Kattegat-Skagerrak, EU Interreg (NYP 20200437).
REFERENCES


Effect of GPS errors on Emission model

Anders Lehmann*, Allan Gross†

*Aarhus School of Engineering University of Aarhus
anders@ase.au.dk

†Department of Business Development and Technology
University of Aarhus
agr@btech.au.dk

Abstract—In this paper we will show how Global Positioning Services (GPS) data obtained from smartphones can be used to model air quality in urban settings. The paper examines the uncertainty of smartphone location utilising GPS, and ties this location uncertainty to air quality models. The results presented in this paper indicates that the location error from using smartphones is within the accuracy needed to use the location data in air quality modelling. The nature of smartphone location data enables more accurate and near real time air quality modelling and monitoring. The location data is harvested from user data gathered in the wild.

I. INTRODUCTION

Air quality can have serious health effects, both immediate and longterm. In many countries limit values for different pollutants are imposed, to keep the health consequences to a minimum. To monitor that the air pollution is well below these limit values, and to be able to alert the affected communities when concentrations are above the limits, extensive air quality models have been developed. These models take different emission sources like long range transport of pollution from regional or transnational sources, urban pollution sources, and emission from traffic into account. All these pollution sources are considered to estimate the influence on the street level air quality. The effects of the weather is also factored in as wind patterns in street canyons will increase the pollution on the lee side of the street and temperature and precipitation will change the rate of chemical reactions and deposition of pollutants.

Air quality in cities is dependent on a number of factors. Pollution can travel for long distances, and thus affect the air quality in a distant city, but often more important are sources local to the city. The local sources of pollution affecting the city air quality are for instance local residential and district heating, local processing industry and traffic. We will in this paper focus on how we can improve the modelling of traffic induced emissions, and thus the traffic related effects on air quality, through data obtained from smartphones used by users of the urban traffic system. Figure 1 shows the different levels of the pollution model.

To model the impact of traffic on air quality the existing models uses traffic counts to estimate the traffic. The traffic counts are typically points measures done rather seldom and for intervals of a few weeks or months. Ordinarily traffic counts only gives information on the number of vehicles passing through the measuring gates, and thus the speeds of vehicles are not measured. Only a few streets in a city is exposed to traffic counts, and the traffic in the remaining streets are estimated from traffic models.

In this paper we consider the quality of data collected from smartphones of people moving in an urban transport system, and if this data can be used as a dynamic supplement to the more static traffic counts. We identify two issues to investigate. First the uncertainty of location measurements based on Global Positioning Services (GPS), which will influence how to count vehicles on the streets. The second issue is the uncertainty on the speed measurement from smartphones, which will influence the outcome of the level of emissions estimated from the model.

The data used in this paper is collected in deployments of the project EcoSense, which have provided smartphone apps for different projects promoting sustainable energy, energy savings, sustainable transport and climate change mitigation efforts.

The paper is first considering the available literature on the subjects of air quality and location accuracy in section II. The experiments are presented in section III. Section IV presents the results of the experiments. The paper ends with a discussion of the contributions and results, as well as an outlook towards future work.
II. RELATED WORK

Air quality modelling in urban areas has been a scientific field since the privatisation of urban transport became abundant. The large concentration of cars in city centres, leading to serious health issues, has spurred the evolution of methods to monitor the level of pollution. The active measurements of pollution led to an understanding of the sources of the pollution, but was too costly to be deployed in more than a few places in each large city. Modelling approaches were considered to be put in place of active measurements for the rest of the urban areas. One of the more thorough models for urban air quality is the Operational Street Pollution Model [1], [2]. OSPM combines results from Long Range Transport models (i.e. Danish Eulerian Hemisphere model [3]) for regional pollution sources, with Urban background models for urban pollution sources and a street canyon model for local sources and variations due to turbulent winds, to calculate the pollution concentration at the street level. The model is used to monitor the adherence to the limit values set up by national and international regulating bodies [4].

The OSPM model uses an extensive amount of data from diverse sources in order to estimate the emission of pollutants, the mixture and reactions of pollutants, and the transport of pollutants. The data sources stems from traffic counts, traffic surveys, traffic models, fleet information, and meteorological data.

Since the advent and ubiquity of smartphones, many projects have utilised the new opportunities for data collection in the wild. The big problem, from an academic point of view, is that it can be hard to persuade users to submit their data to a research community. Some research projects have used the smartphone location data indirectly, by using location data included in status updates for social media [5], [6]. Other projects have used cell tower data from network operators, to coarsely trace movements of individual mobile phone users [7]. Others, like in this project, have succeeded in using data directly from smartphones to perform their research studies [8]–[10].

The GPS system is well documented and very useful [11]. The location and speed accuracy of non-military commercial receivers has been investigated in different settings. In Witte and Wilson [12] GPS accuracy is measured by a bike driving on an athletics running track. The GPS receiver used had an external antenna placed on the bikers helmet, which is different from a smartphone with a built-in antenna placed in a pocket or bag. The paper uses the placement of the antenna to explain the drop in location accuracy during curves, since the bikers head will be at another place than the bike, due to the leaning necessary for keeping the balance on the bicycle while turning.

In Modsching et al. [13] the GPS location accuracy from four different GPS receiving devices is investigated in an urban environment. The authors report high error rates due to buildings shadowing the radio signals from the GPS satellites, and provide map-matching as a tool to improve the location accuracy. Furthermore the paper finds that there is no corre-

lation between actual location error and device reported error level.

A thorough investigation of error statistics of GPS location for GIS applications, is given in Zandbergen et al. [14]. The error statistics are measured by a mapping grade GPS receiver in a fixed outdoor position over eight hours. The mean error is reported to be 2.4 meters.

To improve the accuracy of GPS location and speed estimation from smartphone data, the paper by Nitsche et al. [15] employ a Kalman filter on the raw GPS data to reduce the location uncertainty. The presented experiments show how the Kalman filter removes outliers, but also shows residual inaccuracy when the filtered GPS data is drawn on a map.

III. EXPERIMENTS

A series of experiments was conducted, to investigate the error behaviour of GPS in smartphones.

The first experiment concerns stationary indoor accuracy to test experienced system accuracy, without introducing possible distortions from mobility.

The second experiment is about experienced accuracy of moving GPS receivers in smartphones. Data has been collected from smartphones of different users and the reported uncertainty of the GPS location were evaluated.

Lastly an experiment was made to investigate the sensitivity of the Air Quality model to the average speed of the modelled traffic.

A. Stationary GPS accuracy experiment

The GPS system involves 27 satellites (24 is the designed number of satellites, the rest are in reserve but operational) in six different medium earth orbits at 20,000 km above earth. The orbit is chosen in such a way that each satellite will trace the same area of the earth two times each day. The recurring events of satellites appearing and disappearing over the horizon, could lead to recurring errors in the localisation of a GPS receiver.

To test this hypothesis we recorded the location and error estimate of a smartphone placed in a fixed indoor location, over a period of 5 weeks. The fixed location was chosen to minimise the effects of changing environments, which could cause location errors from multi-path radio propagation, shadowing or other degradation of the received radio signals. The indoor location was chosen to ensure the integrity of the smartphone and to keep it powered during the experiment. The indoor location may also cause some extra damping of the radio signals from the satellites, which can influence the accuracy of the location estimation. The location was chosen to be a single story home in a residential area. The location and horizontal and vertical error were sampled every second.

The experiment was made in two variants - one with Wifi connected to the local Wifi network, and one with the wifi radio turned off. As the smartphone reported location accuracy is determined from the combined localisation the GSM network, Wifi networks and GPS localisation, the two variants were made to test if the smartphone connected to a
TABLE I
STATISTICS FOR THE "HERNING CYKLER" DATA SET

<table>
<thead>
<tr>
<th></th>
<th>Number of trips</th>
<th>Number of location points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herning Cykler data</td>
<td>900 thousand</td>
<td>250 million</td>
</tr>
<tr>
<td>In Herning proper</td>
<td>150 thousand</td>
<td>3 million</td>
</tr>
</tbody>
</table>

local wifi network would influence the location accuracy. The GSM network was disabled in both variants, by removing the SIM card.

B. Mobile GPS accuracy experiment

To assess the accuracy of mobile smartphone localisation data from EcoSense deployments were used. Data on reported GPS accuracy and localisation was extracted from the "Herning cykler" data set. As the data set is collected from users who have downloaded an app from an appstore, the app has been used in many different places, and for different purposes (including boating). As we wish to focus on accuracy in urban environments, points located outside a bounding box containing the city of Herning were removed. The resulting data contains 250 million location points from 900 thousand different trips.

C. Emission modelling speed sensitivity

The OSPM [1] modelling program was used to investigate the potential influence of speed inaccuracy on the results for the modelled air quality. OSPM models the air quality at the street level by combining long range pollution transport models, with urban background pollution model and traffic based street level pollution contributions. In our experiment we choose a street with heavy traffic in a built up area of a city, where the surrounding buildings create a street canyon. In the street canyon, turbulence from cross winds will concentrate the pollution on the lee-side of the street, and thus exacerbate the effects of the traffic emissions. The traffic load was chosen from information on traffic counts performed by the municipality. The builtin meteorological data for the city and fleet data for Denmark was used. A number of simulations was performed where the average speed of the traffic was varied. Numbers for the air quality measures (the concentration of the different pollutants at the height of an virtual sensor) was retrieved from the resulting reports to show the impact of different average speeds.

IV. RESULTS

We present in this section a number of results from our investigation. As the experiments span from single phone stationary experiments, over multiple mobile phones to computer simulations, there are quite diverse results.

A. Stationary GPS accuracy experiment results

The experiment concerning indoor stationary GPS accuracy, was conducted by having an iPhone 4s, without a SIM card, running the app "SensorLog" for five weeks. In the first three weeks the wifi was enabled and connected to a residential wifi network. In the last two weeks the experiment was repeated with the wifi radio disabled.

In Figure 2 a scatter plot for 1.2 million position measurements. To be able to see if there are structures apparent only points with a reported location accuracy below 200 meters has been shown. The figure shows that most points are grouped close to each other. The largest distance between points are 67 m. The central grouping shows some signs of preferred directions, and do not resemble a gaussian distribution of points from a fixed centre.

In Figure 3 locations from the first three weeks, where the smartphone was connected to the wifi network, and where the wifi radio could be used to aid the location system, by triangulation between neighbouring wifi networks. The single outlier from the previous figure is gone, but the preferred directions are more clearly discernible. When looking at a map of the residential area of the experiment, it is a quite suggestive to match the preferred directions to directions from the smartphone toward the neighbouring wifi access points. There app. 800,000 location points in this data set. The largest difference between the locations is 54 meters.

Figure 4 shows the location data from the second experiment...
where the wifi has been disabled. The figure shows the single outlier from figure 2 and a very tight center with a few strays around the center. The preferred directions are less pronounced in this figure and the largest difference between points is 68 meters, and if the outlier is removed the largest distance between points is decreased to 33 meters. There are almost 430000 location points in this experiment.

From the examination of these two experiments we can conclude that using wifi networks as part of a localisation solution, does have an effect, but maybe it is better to solely rely on GPS for stationary use.

To see if there are any recurring accuracy events, which could relate to the recurring constellations of GPS satellites, we show in figure 5 the reported accuracy and the date (x-axis) and time of day (y-axis). The measurements which shows low location uncertainty (less than ten meters) has been removed to not clutter the diagram. Figure 5 shows the 30 location measurements for the 3 weeks where both GPS and wifi was used for location estimation, where the reported accuracy was above 10 meters. There are no apparent recurring daily events so we cannot conclude that satellite constellations seriously affect the accuracy of stationary location measurements. There seems to a small tendency towards larger uncertainty at five in the morning and at ten in the evening.

In figure 6 we show the same diagram as in figure 5, but for the two weeks with only GPS as localisation device. In these two weeks there were 330 incidents of reported errors larger than ten meters, but there are no apparent recurring events of large localisation uncertainty, thus we cannot find support for the hypothesis that satellites constellations influence GPS accuracy. In this GPS-only experiment it seem that not relying on wifi for localisation introduces more incidents of large accuracy uncertainty.

From the above analysis a few observations can be made. First it can be seen from the figures 2, 3 and 4 that the accuracy of GPS location without wifi based location is better than GPS with wifi based location, except for a single point. From the figures 5 and 6 it seems that the variability of the reported errors is higher when the wifi is switched off, even though that the scatter plots suggested otherwise. It seems that the self reported error is overly relying on the access to known wifi access points.

B. Mobile GPS accuracy experiment results

In this section we report the results of our investigation into reported accuracy from smartphones as used in the wild. The data used is collected through the "Herning cykler" deployment. Even though the app was only promoted in Herning, Denmark, the deployment has generated data from other places in Denmark, and outside Denmark, the last part probably because of some of the informants going on holiday, and bringing their smartphones with them. The dataset contains data from 900 thousand trips, of varying lengths, with 250 million location measurements.

In figure 7 a histogram for the reported uncertainties. The histogram is abbreviated to uncertainties below 200 meters, and the bins are ten meters wide. It is clear to see than the most of the reported uncertainties are ten meters and below. But there is still almost half of the measurements with larger uncertainties.

The figure 8 shows some of the long tail of the uncertainty distribution. The small uncertainties (below 20 meters), has
ben removed to enable the view of the smaller counts of the larger uncertainties. In the figure 8 one can see that even though most of the location measurements have uncertainties below 100 meters, there is still a considerable amount of measurements with quite large uncertainties above 500 meters.

C. Emission modelling speed sensitivity results

The results of the experiment to quantise the sensitivity of the air quality model towards the speed of the traffic.

The figure 9 shows the concentration of $NO_2$ at a simulated measurement point in a street canyon. The figure shows that the concentration of $NO_2$ drops as the average speed increases. The drop in concentration is congruent with the speed dependency of emission factors [16], for low speeds the emission per meter is high due to idling or low utilisation of the motor. At higher speeds the motor output is used more effectively, thus the emission factor is lower. As the speed increases towards highway speeds the emission factor grows as the motor has to overcome increasing wind forces. We have only made the simulations for speeds less and equal to 120 km/t, as higher speeds are not deemed relevant for traffic in an urban street canyon.

From the figure 9 it can be seen that the speed sensitivity of the emission model is largest for low vehicle speeds. To accurately model the street level pollution it is important that the traffic data is accurate for low speeds.

V. APPLICATION OF GPS LOCATION DATA IN STREET LEVEL AIR QUALITY MODEL

To improve street air quality models like OSPM, with data obtained from smartphones of urban travellers, two essential pieces of information can be supplied. The location of travellers can be used to account for emissions from a trip to the locations traveled through. The speed at different locations is the second important information that the model needs in order to estimate the emission amount.

The smartphones could also supply information on the type and motor size of the vehicle, which could further improve the model estimate, but currently this information has to be gained by asking the user of the smartphone. Research has been done to automatically detect transportation mode [17], but currently it seems impossible to detect motor size and fuel type. One exception is to detect electric vehicles versus internal combustion engine vehicles [18].

VI. DISCUSSION

From the experiment with the OSPM model, it can be seen that the speed of vehicles are very important to the air quality at the street level. But it is also important to be able to attribute the emissions from vehicles surveilled by applications like ours to the right streets. That means that the accuracy of the location of vehicles monitored by smartphone apps must be high enough to assign the location to a single street.

To evaluate the influence of GPS accuracy an air quality modelling we have to determine the connection between location accuracy and speed accuracy and possible methods to overcome inaccuracy problems.

From the GPS design documents the best case error estimate is below six meters, which should be achievable in 95% of the time [11]. The speed is (falsey) reported to be larger than zero in 29 cases in the stationary experiment. As the reported
locations as seen in figure 2, 3, 4 should result in more reported nonzero speed events, if only the raw GPS locations were used to calculate the reported speed. The largest reported speed in the 29 cases of non zero speed is 2 km/h.

Map matching algorithms can reduce the inaccuracy from the GPS data, as seen in [10]. A median filter might also provide a stable location trace.

The errors as seen in the real life data set from “Herning Cycler”, points toward the conclusion that we can use smartphone data to supplement our other data sources, for air quality modelling at the street level. This possibility enables more dynamic and up to date air quality monitoring, and possibly also new kinds of health alerts and urban transport planning overview.

Since the projects that have enabled the data collection has been targeted at special audiences, it is likely that there is a bias towards people who cares about sustainability, environment and man made climate change. The validity of the analysis of GPS accuracy does not suffer from this bias.

VII. FUTURE WORK

In this paper we have only considered the collected data to be from person transport in cars. To further the quality of the collected data, methods to discern between different modes of transportation, to correctly adjust for the emissions from the vehicles, is needed.

VIII. CONCLUSION

We have measured the accuracy of GPS data from smartphones, and investigated if the data is As traffic models become more ubiquitous, and are able to model the largest cities [19], urban transport planning get more tools for simulation of urban policy initiatives and new infrastructure initiatives, we hope that the work in this paper will lead to urban planning to take air quality into account.

ACKNOWLEDGMENT

We want to thank Henrik Blunck, and Niels Olof Bouvin for their generous support.

This work has been supported by The Danish Council for Strategic Research as part of the EcoSense project (11-115331), the TUD COST Action TU1302 Satellite Positioning Performance Assessment for Road Transport (SaPPART), and EU Interreg project Biogas2020.

REFERENCES


Towards vehicle emission estimation from smartphone sensors

Anders Lehmann
Aarhus School of Engineering
University of Aarhus
Email: anders@ase.au.dk

Abstract—CO₂ emissions from transport constitutes a large, and growing, part of the total carbon emissions. We present a model to estimate CO₂ emissions from passenger cars on basis of GPS and accelerometer data gathered from the driver’s or passengers’ smartphone apps. As part of an experiment to establish ground truth, a method for measuring fuel consumption without instrumenting vehicles is presented. As part of this estimation model, a method for discerning between different driving modes (idle, accelerating, cruising and turning) is presented, using the K-means clustering method. This method will increase the accuracy of the emission model by estimating the fuel consumption while accelerating and braking, as well as estimate idle consumption. The model will enable a more detailed emission inventory, both in terms of location and time. Apart from emission of CO₂, the method can also be used for estimation of other transport related emissions. The detailed localised emissions can be used to monitor air quality in an area in near real-time as well as help in creating precise emission data for green accounting.

I. INTRODUCTION

Smartphones bring new opportunities to researchers who are studying the behaviour of modern human beings. For research in mobility, transport and atmospheric chemistry modelling [1], [4] the proliferation of smartphones yields the potential to lower the cost for gathering relevant data considerably [18].

In modern societies, carbon emissions from transport are a large and growing part of the total carbon emissions from human activities. For instance, the emission of CO₂ from transport in Denmark was 24% of the total carbon emissions in 2012 [21], and road transport is responsible for 67% of the carbon emissions from transport. In order to create viable measures to lower the carbon emissions from passenger transport, there is a need for more detailed models quantifying these emissions. By using data from smartphones, we aim to obtain more accurate information on where and when emissions from transport are created. An obvious problem in using smartphone data for calculation regional or national emission inventories is the incomplete coverage of the traffic system that the method offers. Not all drivers have smartphones and not all smartphone owning drivers are willing to share their data. By combining the gathered data with a traffic congestion model, as reported in [17], we can use the incomplete data to estimate the traffic in the transportation system. These models can help policy makers and planners make informed decisions on future changes to the road transport system and related infrastructure. Another important application of these models is found in green accounting [8] for businesses. Here the models offer a more affordable and precise methodology to create the foundational data for reporting climate impact of a business.

As carbon emissions are closely related to the fuel consumption, experiments were designed and carried out which gathered driving-related data from a smartphone and data on the amount of fuel consumed, from measurements.

For the experiment three cars was used: a Citroen Xantia 1.8 from 1999, a Citroen Xantia TD 1.9 from 1999 and a Tesla model S 90 D from 2015. To accurate measure the fuel consumption during the experiments, a measuring method, which did not involve instrumenting the vehicles was developed.

The remainder of the paper is organised as follows: In Section II we review related work, including methods for measuring the fuel consumption and for creating related models of fuel consumption from sensor data. Additionally, we briefly review methods to determine different driving styles and driving modes. In Section IV the models used to convert the experimental data into emission data are presented. The results of the experiments which partly relate to fuel consumption measurement and modelling and partly to reliable detection of driving modes, are presented in Section V and the paper is concluded in Section VII.

II. RELATED WORK

Carbon emissions from passenger cars are closely linked to fuel consumption, as the primary source of CO₂ is the combustion of hydrocarbons. To model the fuel consumption from smartphone data, we need to establish a correlation between the observed data, and the consumed fuel.

Hilpert et al. [12] propose a real-time data gathering system and model based on On Board Diagnostics system (OBD2). From the data gathered emissions from the vehicle are calculated, by a simple relation between airflow into the engine, the stoichiometric fuel to air ratio, and the CO₂ emission factor for gasoline/diesel. No actual data are reported. The OBD2 option was not available in cars used in the experiment. The Citroen Xantias were too old and the Tesla does not allow access to consumption data via the external OBD2 connector.
For detection and analysing of driving styles for prevention of car crashes, the authors in [14] make use of sensor fusion of sensors from smartphones, and data obtained from the internal CAN\textsuperscript{1} bus in cars. The authors show that it is possible to accurately detect driving events and to classify the aggressiveness of the driver from the used sensor. In this paper, we want, to be able to determine turns as well as the aggressiveness of accelerating and braking events.

Models for the emission from gasoline cars are evaluated in [24] by combining simulation with measurements done in a combination of data from the OBD2, and sensors placed in the tail pipe. Three different numerical models were compared to the measurements. The same instrument setup (OBD2 together with tailpipe sensors) were used in [9], where the goal was to estimate emission factors for different driving modes. One conclusion from this study was that the emission pattern for cold start driving is significantly different from steady state driving. As our experiment is designed to be able to model the most used driving patterns, we have focused on warm engine mode of operation, thus we perform a warm up drive before the actual test run.

Lee and Gerla [15] provide an overview of different projects using mobile sensing platforms. The paper is mostly concerned with network topologies for mobile sensor network and back-end support for data storage and retrieval, but also has some input on different sensors used.

In Boriboonsomsin and Barth [3] the effects of the road grade (road inclination) on fuel consumption is investigated, through the use of OBD2 measurements and models. Two different routes between an Origin-Destination pair (one route through a flat area, and one through a mountain pass) are compared over multiple trips. The fuel consumption is determined from a binary fuel cut signal from the OBD2 data, and compared to fuel consumption modelled by CMEM (Comprehensive Modal Emission Model).

Hemminki et al. [11] presents an overview of methods, applications and problems, for using accelerometers to determine transportation modes. Examples of how to use Hidden Markov Models classifiers to discern between different transportation modes are given. Mun et al. [20] adds a personalised environmental impact report generated from mobile sensed data. Where these papers focus on transportation modes, i.e. to discern between different modes of transportation, in our paper the focus is on different driving modes i.e. how the driving is performed.

Wüstenberg et al. [25] gives an example of discerning idle electric vehicles from idle internal combustion engine vehicles, by using accelerometer data to measure the revolutions of combustion engine in idle mode. In idle mode, electric vehicles do not have a turning engine. The paper shows how signal analysis combined with classifiers can detect the vehicle differences from engine vibrations. The paper also shows that noise when cars are driving makes engine detection difficult.

To monitor road condition Ghose et al. [10] proposes to use the accelerometer to measure road quality and detect the positions of potholes.

Development and yearly reporting of national emission inventories was agreed on in the terms of the Kyoto protocol. This has led to the development of rigorous models and protocols for reporting of the total national emission of climate forcing gases. To be able to make detailed reports of the transportation related emissions the COPERT\textsuperscript{2} [19] program was developed as a European transport emission model.

Beusen et al. [2] used GPS and OBD2 to get information on engine rotational speed (RPM, rotation per minute) and other engine parameters, to test the long-term impact of Eco-driver training. The information gathered was used to model different driving modes, such as accelerating, braking and at which engine speeds, gear shifts were performed.

In the papers mentioned in this overview a cross section of methods and tools for estimation of fuel consumption and emission from light duty vehicles and passenger cars were presented. The tools used in these papers are OBD2, GPS, accelerometers and tail pipe emission sensors. The methods used are controlled driving cycles, simulation models, classifiers and sensor fusion.

There is also a host of commercial and free apps offering driving style analysis directly on the smartphone.

The work we present in this paper extends and complements the work in the above-mentioned papers by using measurements from smartphone accelerometers and controlled measurement of fuel consumption, and by exploring a different classifier than presented in the above papers. This work can be used for vehicles that do not have an OBD2 connector and cannot be instrumented with mass flow meters for measuring fuel consumption as in [6], [13].

We have chosen to focus on a flat test course, in order to avoid the road grade as a consideration. We have also chosen to use a single route with the same point as start and finish, as our test course.

### III. Experiment

An experiment was performed to get a baseline for our fuel consumption model. Data from a smartphone was combined with accurate measurements of fuel consumption.

Researchers have reported fuel consumption measurements by instrumenting the vehicles with flow gauges in the fuel line [6], [13]. These measurements allow for real-time measurements of the fuel consumption. In this paper, a manual measurement procedure was developed to obtain the fuel consumption measurements (see Section III-B) since it was not possible to instrument the cars with flow meters.

#### A. Route selection

A number of criteria were considered to select a route for the experiment. The length of the route should be long enough to have a measurable consumption but short enough

---

\textsuperscript{1}"Controller Area Network", an intra vehicle network standard ISO 11898

\textsuperscript{2}http://emisia.com/copert
TABLE I
DESCRIPTION OF INPUT DATA

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of trips</td>
<td>21</td>
</tr>
<tr>
<td>number of data points</td>
<td>1.5 million</td>
</tr>
<tr>
<td>trip length</td>
<td>13 km</td>
</tr>
<tr>
<td>number of traffic lights</td>
<td>19</td>
</tr>
<tr>
<td>number of roundabouts</td>
<td>4</td>
</tr>
<tr>
<td>elevation difference</td>
<td>6 m</td>
</tr>
</tbody>
</table>

to be repeatable and less than the range of a tankful of fuel. Furthermore, the route should comprise different driving patterns (urban, rural, high speed/low speed) to simulate real life driving patterns.

The route was chosen to mainly be in an urban traffic setting. The urban setting was chosen in order to get as many accelerations and idle periods as possible. The route did contain a short distance on a rural highway in order to have higher speed measurements as well. The route is shown in Figure 1.

The length of the chosen route was 13 km, which was short enough to be repeated, and long enough to make measurement of fuel consumption possible. The fuel consumption for the route was approximately one litre of fuel, depending on driving style.

To prevent cold start behaviour to influence the experiment, the vehicles were driven until the engine had reached the operating temperature before starting the test. The warmup drive was also performed for the electric vehicle to prevent battery conditioning effects to influence the experiment. The experiment was performed 21 times and the fuel consumption was determined after each run. The experimental parameters are summarised in Table I.

B. Measurement of fuel consumption

The following procedure was developed to determine the fuel consumption. At the start of the course the gas tank was filled and the fuel level in the gas tank funnel was measured. After the course was completed fuel was refilled to the same level of fuel in the funnel from a fuel canister. The weight of the fuel canister was measured before and after the gas filling process to determine the amount of fuel refilled, and thus the consumption of fuel for the trip. To get a reproducible reading of the weight of the canister the weight has to be placed at the same place, with the same orientation, for each measurement. A scale with 1g accuracy was used in the experiment. In order to reduce the effect of wind on the measurement the scale and canister was placed in a cardboard box during the measurement.

This manual method was needed since the use of the fuel dispenser at the fuel station proved inaccurate. The main problem of using the fuel dispenser to measure the amount of fuel consumed by driving the test route, is that it is hard to get the same filling level at each filling. Using the automatic filling stop mechanism is unreliable, as the filling level becomes dependant on how much foam the filling process has created in the tank and the funnel. This creates large deviations of the amount of fuel used to refilling the fuel after the test runs.

By accurately measure the filling level and refill the fuel from a canister, it was possible to have very small differences between subsequent test runs.

The accuracy of the measurement can be estimated by inspecting the uncertainty of the individual steps of the measurement. The tank filling level can be measured with an accuracy of less than one cm, and since the diameter of the funnel is five cm the maximum error of the fuel measurement will be less than 20 cm³ or 20 ml, which is 2% if the total fuel refill is one litre.

The use of fuel (energy) in the electric vehicle was measured from the internal energy consumption monitor. The dashboard energy consumption was read before and after the completion of the test course.

IV. MODELLING EMISSIONS

An accurate model for carbon emission from transport can be derived from an accurate model of fuel consumption of vehicles, since the carbon emissions are related to the combustion of fuel. The relation between fuel consumption and CO₂ is based on the assumption of perfect combustion, where all available carbon in the fuel is converted to CO₂.

The Equation 1, from [22], gives the emission $E$ in grams for the fuel consumption $FC$ also in grams. The coefficients $r_{HC}$ and $r_{OC}$ are the ratio between hydrogen and carbon, and oxygen and carbon, respectively in the fuel. The numbers in the equation are the molar weights of the CO₂ (44.011), oxygen (16.0), carbon (12.011), and hydrogen (1.008).

$$E = \frac{44.011}{12.011 + 1.008r_{HC} + 16.000r_{OC}} \cdot FC$$

(1)

The energy consumption of the electric vehicle was measured by the on-board trip measurement. Two examples of the dashboard (before and after the test route drive) can be seen in Figure 2. The highlighted areas show the energy consumed since last recharge (upper red circle), and the time of day (lower red circle).

The data received from smartphones comes from a variety of sensors. For this experiment, we used the accelerometer and GPS sensors. From the GPS sensor information on speed, heading and location is obtained with a sample rate $f \approx 1 Hz$, and from the 3d accelerometer the size and direction of acceleration is gathered with a sampler ate of $f \approx 100 Hz$.

A. Method from National Emission model

Emission from warm combustion engines is modelled from the speed of the vehicle in the COPERT IV [5] emission modelling program. As an example, the emission CO₂ is modelled by modelling the fuel consumption of the vehicle, since the CO₂ emission is directly linked to the amount of fuel burned in the engine. Equation 2 is an example of how COPERT IV models the fuel consumption of a petrol driven passenger car as a function of the speed of the vehicle [22].
There is a similar equation covering the fuel consumption of diesel cars.

\[ FC_{warm} = \frac{a + c * v + e * v^2}{1 + b * v + d * v^2} \]  

The constants \( a - e \) are found by driving cars on a test circuit while measuring speed and fuel consumption. The constants differ by car type, engine size, year of manufacturing, and regulation. The COPERT IV program uses mean values (best fit) for all tested vehicles within a particular class of vehicles.

Figure 3 shows an example of the fuel consumption of passenger cars from 1999 versus speed as calculated by Equation 2. The values for the parameters \( a - e \) are looked up in "Denmarks National Inventory Report" [21]. The consumption is measured in \( l/km \), and is seen to have a minimum at approx. 80 km/h for petrol cars and approx. 60 km/h for diesel driven cars. The fuel consumption is high for low speeds as that the engine is underused, and the fuel consumption is high at high speeds due to increased wind forces on the vehicle. The Equation 2 is only defined for the speed range 10 – 130 km/h.

From fuel consumption, the CO\(_2\) emission can be modelled under the assumption of perfect combustion (Equation 1). If all carbon atoms in the fuel is oxidised into CO\(_2\) we only need to know how many carbon atoms there are in a litre of fuel.

**B. Single trip emission model**

The model developed in this paper builds upon the work from the national emission inventories. By using the instantaneous speed data obtained from smartphones, and the fuel consumption versus speed curve from the COPERT program, instantaneous emissions can be calculated.
Fig. 3. Fuel consumption versus speed for a 1999 passenger car.

By assuming a constant speed between two measurements of the speed of the vehicle and by measuring the distance between the two speed measurements, we calculate the fuel consumption for the distance by multiplying the distance with the fuel consumption given by Equation 2. The fuel consumption and thus the CO\(_2\) emission for a single trip can then be calculated as the sum of the instantaneous consumptions.

One problem with the above method is that the fuel consumption model, defined in Equation 2 is only valid for speeds between 10 and 130 km/h. This means that fuel consumption when the vehicle is idle (for instance waiting for green light at an intersection), is probably modelled too high.

To model the emission from accelerating vehicles, we first need to establish the different forms of driving modes. For each driving mode, an emission model has to be employed.

In order to model the effect of vehicle acceleration, we divide the trip into six driving modes, idle, forward acceleration, deceleration, cruise at constant speed, left turn and right turn. Reversing is not considered in this project, partly because we have no data and partly because it is an infrequent driving pattern.

C. Emission model for accelerating vehicles

There are a number of issues that we need to consider to use the accelerometer to discern between the different driving modes. The position and orientation of the smartphone in the vehicle is a matter for consideration. We kept the smartphone orientation fixed, with the X-axis of the accelerometer pointing forward (usually the top of the smartphone), and the screen facing upwards (the Z-axis). We made this choice to fix the smartphone direction to eliminate noise from smartphone movement not originating from the vehicle movement.

1) Forward acceleration: The different driving modes have different acceleration signatures. The forward acceleration would have a positive acceleration in the direction of the X-axis, and only a small acceleration component in the Y-axis and Z-axis. Likewise, the braking driving mode would have a significant component in the negative X-axis and only small components in the Y and Z directions. The emission model for acceleration and braking would not be the same though, as energy is spent during acceleration, but energy is not retrieved from braking in most vehicles. For the forward acceleration mode, the emission model we use is a model derived from physics. To accelerate a body of mass \(m\) a force \(F\) given by Newton’s Second Law is needed. The work \(W\) this force does is given by integrating the force over the distance the force works. This leads to Equation 3.

\[
W = \int_0^L F \cdot ds = \int_0^L m \cdot a \cdot ds
\] (3)

Equation 3 ignores the efficiency of the internal combustion engine. The efficiency is very dependent on the load and rotational speed of the engine, but since we cannot determine the actual engine speed during acceleration an efficiency value of \(\eta = 0.4\) was chosen [7]. The fuel consumption for an acceleration \(a\) over a distance of \(L\) is thus estimated as in Equation 4.

\[
FC(a, L) = \frac{1}{\eta} \cdot \frac{W_f}{d_f} \cdot \int_0^L M \cdot a \cdot ds
\] (4)

where \(W_f\) is the energy content of the fuel in MJ/kg and \(d_f\) is the density of the fuel resulting in the fuel consumption \(FC\) is calculated in litres.

2) Turning: Turning is characterised by higher acceleration in the Y direction, and left and right turn can be discerned from the direction. The emission model for turning is not being covered in this paper.

3) Cruising: Cruise is characterised by no significant acceleration in either X or Y direction. For cruising the emission model from Equation 2 can be used since the speed is constant.

4) Idle mode: As in cruise mode idle mode is characterised by a lack of acceleration in X and Y but the speed is zero in idle. The emission model for idle mode is an emission per time unit instead of emission per distance unit. The model is created from Equation 2 by taking the fuel consumption at lowest valid speed for the equation which yields the fuel consumption \(g/km\) and multiplying with the lowest valid speed to get the fuel consumption in \(g/h\).

5) Braking: The emission model for the braking and idle driving mode can be the same, since regenerative braking is not used in many passenger cars with only a combustion engine.

The clustering algorithm K-means is used for classifying the accelerometer data into the six driving modes. It is an unsupervised clustering algorithm, that given the number of wanted cluster centres, minimises the mean squared error of the distance between the cluster centres and the cluster members for all data in the data set.

V. Results

The experiment was carried out in Holstebro, Denmark with three different cars (two Citroen Xantias and one Tesla S 90 D). The data collection was done with an iPhone 4s running an
app called SensorLog\(^3\). The data gathered consisted of GPS, accelerometer. The GPS is sampled at approximately 90 Hz.

A total of 28 test runs were driven, seven test was not completed due to driver errors, leaving usable data from 21 test runs. It became apparent during the processing of data that the original way of measuring fuel consumption by using the meter at the gas pump was unreliable, and thus were a more reliable measurement method developed. Two test runs were completed with the improved method.

The chosen test route can be seen in the map in Figure 1.

As seen in Figure 5 the results of fuel consumption model is reasonable stable with a mean value of $1.31 \pm 5\%$ litres.

The modelled fuel consumption uses the Equation 2 for each sample point. The speed is the speed estimate gained from the GPS data. Each consumption factor derived from Equation 2 is then multiplied with the distance from the previous sample point. The distance is calculated from the GPS location and the previous GPS location, using the Haversine formula [23] to take the spherical nature of the Earth into account.

\(^3\)http://sensorlog.berndthomas.net/

An example speed curve of a run is shown in Figure 6. The different driving patterns are apparent in the Figure, as the large number of speed changes suggests.

**A. Acceleration data**

The data from the accelerometer of the smartphone is used for determine the driving pattern of the driver. Part of the driving pattern is determined by road conditions, such as intersection, stoplights, curves, and turning. Other patterns are due to personal driving style ie. heavy acceleration, overtaking, and braking versus smoother driving and lastly some patterns are due to traffic conditions like congestion, accidents, and slow or fast moving vehicles.

In Figure 7 the magnitude of the acceleration is shown in a spectrogram. The x-axis is time (sample number) and the y-axis is the frequency spectrum of the magnitude of the acceleration signal. The idle periods are visible as vertical bands with energy only at a few frequencies. The visible single frequencies are related to the revolutions of the engine. It can
Fig. 8. A spectrogram for an electric vehicle clearly demonstrating the lack of an engine signal in idle periods. The arrow points to an idle period.

be seen from the figure that the idle running of the motor when the vehicle is standing still is very visible, while the motor frequency is not discernible while the vehicle is moving due to noise from the wheels moving on the road.

The spectrogram for an electric vehicle is shown in Figure 8. The absence of the engine signal during pause in the driving cycle is quite apparent. The spectrogram indicates that the accelerometer signal can be used for detection of an idle running engine in a stationary internal combustion vehicle. The absence of vibrations for electric vehicles can be used to detect a stationary vehicle.

Fig. 9. Clustering of accelerometer data. Top X, Y plane. Left bottom X, Z plane, right bottom Y, Z plane

A simple K-means clustering algorithm was used to discern different driving patterns. There are six driving modes that are of interest: idle, forward acceleration, braking, cruising, left turn, and right turn, thus the number of clusters were chosen to six. Figure 9 shows the results of the K-means clustering algorithm on the data from the 18 test runs with internal combustion engine vehicles. The clustering is performed separately on each test run, but since the position and orientation of the smartphone was kept the same in all tests (parallel to the vehicle, with X toward the front face Z up), we present the results in the same figure. The upper figure shows the X and Y part of the cluster centres, and the two lower figures shows the X, Z part and Y, Z part respectively. There are many cluster centres close to the origin of the graph, corresponding to either idle or cruising driving mode. Many of the test have two centres at or close to the origin in the X, Y chart, indicating that the Z axis differs in the two cases, which is expected from the difference between idle (no road vibration) and cruising (road vibration).

The X, Z and Y, Z charts of the cluster centres are aligned along X = 0 and Y = 0 axis, again indicating that the Z value of centres are connected to low acceleration in the X, Y directions.

To further improve the detection of the driving mode a low pass filter was applied to the three different acceleration signals before performing the K-means clustering. The result for the X, Y plane is shown in Figure 10.

Fig. 10. Clustering of low-pass filtered accelerometer data (X, Y plane)

With the added low pass filter a good separation between the driving modes is apparent. The added red lines indicate how to discern between the four accelerating modes: forward, brake, left turn, and right turn. To discern between idle and cruising one have to take the Z axis signal into account as well or consider the speed signal from the GPS.

An example of the clustering combined with the speed curve is shown in Figure 11. The figure shows part of the speed curve for a test drive done in 2015. The excerpt is approximately five minutes long, and shows a part of the test drive performed within a city. The speed is low (below 40 km/h), and there are frequent accelerations and decelerations.
A few observations can be made from figure 11. First, it can be seen that only four of the six clustering’s are represented in this figure. The colours signify the different clusters to which the accelerometer data has been characterised into. From figure 11, it can be seen that the yellow colour is mostly associated with acceleration and the cyan colour is correlated with cruising. The blue colour seems to be correlated to braking and deceleration. There is a small amount of red at low speed which could be representing idle mode. However, when looking at the full trip in Figure 13 the red colour is concentrated at the beginning of an acceleration thus it could be related to turning.

The section in Figure 11 was taken from the full speed curve of the trip as shown in Figure 13.

To further investigate if the clustering with the K-means algorithm relates to driving modes the clustering is visualised on a map. In Figure 1 the GPS positions recorded by the smartphone are shown as dots on the map. The test drive starts in the upper right corner (north-east) goes east before proceeding to the south westerly corner of the route. The route continues east before going north west and eventually turns east to finish at the starting point. The colour of each dot relates to the cluster the point has been classified to.

Figure 12 shows an enlarged part of the map of the trip. This shows a drive in a residential area with a speed limit of 50 km/h (the direction of travel is right to left). The colour of the dots are almost all white until the route reach the corner where a little braking is necessary. This suggests that the clustering algorithm can detect the cruising driving mode from the accelerometer data. The light blue colour just before the leftmost corner suggests that the clustering algorithm detects the breaking, and the orange colour after the corner suggests a light acceleration after the braking through the turn.

To show how traffic lights can be detected from the clustering algorithm, another enlarged part of the route is shown in Figure 14. In this part of the route two traffic lights are shown (both with left turns) in the two Y-forks the route passes through. At the junction before the left turn two orange coloured dots are seen. These dots are examples of the idle mode detection of the clustering algorithm. The deep red dots after the two left turns signifies the rapid acceleration from standstill.

The examples shown here suggest that it is possible to match clusters to individual driving modes, but the examples also show that the matching is not perfect. For instance, the cyan colour in Figure 13 is found both at high speed, which would suggest cruising driving mode and at low speed approaching idle mode. To have a value of the mis-classification of the cluster algorithm, we are using the speed information from the smartphone as ground truth. By assuming that the cyan colour is classified as cruising, and the low speed datapoints classified as cruising are misclassified. The assumption is that in the tests there are no cases of cruising with a speed lower
than 10 km/h. We can calculate an estimate of the rate of cruise miscalculations for this trip as:

\[ R_{\text{cruise}} = \frac{N_{\text{cruise points < 10 km/h}}}{N_{\text{cruise points}}} \] (5)

The results for cruise detection with the K-means clustering algorithm is summarised in Table II. The mis-classification of the cruise mode for internal combustion engine vehicles is between 5% and 24% but very close to zero for electric vehicles. It is interesting that the reduced noise from an electric vehicle seems to make the detection of the cruise mode more accurate, than the cruise classification of an internal combustion engine vehicle.

B. Emissions from Electric Vehicles

To model the emissions from battery powered electric vehicles another method has to be used, since the emission is decoupled from driving the vehicle. To model climate forcing gas emission from battery powered vehicles, we have to look at the emissions from the electricity production which power the charging of the battery at the time of vehicle charging instead of the time of vehicle driving. In Figure 15 the CO\(_2\) emission per kilowatt hour is shown for the Danish grid over a five-year period. The CO\(_2\) emission data is reported as an average over five minutes, and the data shown in figure 15 is further smoothed by a rolling median filter and rolling average filter, to show the seasonal changes. The figure shows how varied the CO\(_2\) emissions are, both in terms of daily, monthly variations and through the year. The large electricity production from sustainable energy sources in Denmark creates much of the variation in CO\(_2\) emission. In 2015 wind turbines produced 42% of the Danish electricity consumption.

![Fig. 15. The CO\(_2\) emission factor from the Danish electricity grid](image.png)
VI. DISCUSSION
The speed information is generated by the GPS chip in
the smartphone, and as the Figure 14 show, not completely
accurate. The accuracy of GPS and possible effects are re-
ported in [16]. Even though this lack of accuracy the GPS
data from the smartphone is useful as it is synchronised with
the accelerometer. To use other sources for ground truth, we
could have used video which could be synchronised within
a few frames (~50 ms) but would lack in position accuracy
(manually annotation of position). Professional GPS receivers
could provide better location accuracy but would present a
synchronisation problem. So, we use the GPS data not only
because it is already available, but also because it is already
synchronised.

VII. CONCLUSION
In this paper, we have presented a method for modelling
carbon emission from smartphone data and a method for
discerning six different driving modes. The carbon emission
model is quite stable with a standard deviation of 0.03.

The K-means based method for discerning between driving
modes is able to discern between four acceleration modes
(forward, brake, left turn, and right turn) and two non-
accelerating modes. One of the remaining problems of the
above described method is the ability to accurately map driving
modes to the clusters found with the K-means algorithm. The
method can be extended to encompass all the emissions from
a vehicle. This will be a valuable contribution to atmospheric
chemical transport modelling since traffic emissions is the
largest anthropogenic emission uncertainty in these models.

ACKNOWLEDGMENT
We want to thank Henrik Blunck, and Niels Olof Bouvin
for their generous support.

This work has been supported by The Danish Council for
Strategic Research as part of the EcoSense project (11-
115331), the TUD COST Action TU1302 Satellite Positioning
Performance Assessment for Road Transport (SaPPART), and
EU Interreg project Biogas 2020.

REFERENCES
logging devices to study the longer-term impact of an eco-driving course.
Transportation research part D: transport and environment, 14(7):514–
520, 2009.
[3] K. Boriboonsomsin and M. Barth. Impacts of Road Grade on Fuel
Consumption and Carbon Dioxide Emissions Evidenced by Use of Ad-
vanced Navigation Systems. Transportation Research Record: Journal
integrated model study for europe and north america using the danish
eulerian hemispheric model with focus on intercontinental transport of
[5] I. COPERT. Computer programme to calculate emissions from road
transport. user manual, version 5. European Environment Agency,
cation of instantaneous emission modeling. Atmospheric Environment,
for development of mobile source emission factors. Proceedings of US
Environmental Protection Agency Emission Inventory Conference, pages
condition monitoring and alert application: Using in-vehicle smartphone
as internet-connected sensor. In Pervasive Computing and Communi-
cations Workshops (PERCOM Workshops), 2012 IEEE International
portation mode detection on smartphones. Proceedings of the 11th ACM
Conference on Embedded Networked Sensor Systems - SenSys ‘13, pages
51–64, 2013.
product carbon footprints in transportation processes based on OBD2 and
smartphones. Proceedings of the Annual Hawaii International
Conference on System Sciences, pages 1–10, 2011.
Networked Suite of Mobile Atmospheric Real-time Sensors. pages
a smartphone as a sensor platform. IEEE Conference on Intelligent
congestion model. In 2016 IEEE International Conference on Pervasive
Computing and Communication Workshops (PerCom Workshops), pages
Toward Zero-Effort Carpooling on Individual Smartphone. International
E. Howard, R. West, and P. Boda. {PEIR}, the Personal Environmental
Impact Report, As a Platform for Participatory Sensing Systems Re-
search. In Proceedings of the 7th International Conference on Mobile
Systems, Applications, and Services, (MobiSys ’09, pages 55–68, New
York, (NY, USA), 2009. ACM.
M. Mikkelsen, R. Albrektsen, M. Thomsen, K. Højgaard, L. Hoffmann,
P. Fauset, H. Bruun, V. Johannsen, T. Nord-Larsen, L. Vesterdal,
I. Møller, O. Casperson, E. Rasmussen, S. Petersen, L. Baunbæk, and
man, K.-h. Zierock, M. Andre, M. Winther, N. Gorrisen, P. Boulter,
R. Jounard, R. Rijkeboer, G. Vennard, and S. Hausberger. EMPE/EEA
European Environment Agency (EEA), 2012.
[22] D. Rick. Deriving the haversine formula. In The Math Forum, April,
1999.
ical models for simulation of real-world hot-stabilized fuel consumption
and emissions of gasoline light-duty vehicles. Transportation Research
tinguishing Electric Vehicles from Fossil-Fueled Vehicles with Mobile
Sensing. 2014.
The paper is submitted, pending review.
Origin destination matrices synthesised from Smartphone data

Anders Lehmann

Aarhus School of Engineering, University of Aarhus, Campus Herning, Birk Center Park 15, 7400 Herning, Denmark

Abstract

Modelling of pollution in urban areas is used to monitor violation of limit values for pollutants which can impact the health of the exposed citizens. Pollution modelling lessen the need of having an expensive network of measuring stations. As transportation is a significant contributor to urban emissions, a transportation model should be part of the pollution model. As the origin destination matrix is an input parameter to the transport model, describing the transport demand in the urban transportation network, it is important that the demand is described realistically and accurately. The main contributions of the paper are: First, an algorithm to synthesise the origin destination matrix from smartphone data is proposed in the paper. The algorithm is using GPS traces recorded from travellers using the transportation network. The GPS traces are map-matched to a road map, and the start and end points of the map-matched routes present a set of origin destination pairs. The second contribution is the use of map-matched routes as a set of experienced routes to form a starting point for the transport model.

Keywords: OD matrix, smartphone, transportation model.

1. Introduction

Modelling of transport related pollution in urban areas is important to monitor and predict pollution concentrations accurately. The modelling can be used for planning of the transport network, to plan for pollution mitigating actions, or to predict and warn if pollution limit values are violated.

In this paper, an extension to an congestion model for Urban traffic systems implemented in a relational database system is presented. The implementation was first presented in Lehmann and Gross (2016b). The input to the model is an urban roadmap, travel demand data gathered from smartphones, and can be supplemented with traffic counts or travel survey data. The model is implemented in a relational database system, which allows for integration to Geographical Information System (GIS) data, and shows acceptable performance for even large urban transport system. The city of Istanbul is used as a case study of the performance of the implementation. We propose to extend existing route choice models in two ways. Firstly, since the collected data gives detailed information on experienced routes, we propose to use the observed routes as a start choice set for the route choice model. Secondly, we propose to create the origin destination demand data from users of the transport system, whose smartphone app reports their travel behaviour (see Wüstenberg et al. (2014)).

The map data for constructing the road network is provided by Open Street Map1, which provides maps of almost every part of the globe. The Open Street Map data is converted into a road topology, ready to be imported into a relational database by the program OSM2PO2. The database system is chosen to be the open source database PostgreSQL3, with the extensions PostGIS4, for handling geographic data from map providers. The PostgreSQL extension pgrouting5 is used for routing.

1http://openstreetmap.org
2http://osm2po.de
3http://postresql.org/
4postgis.net
5http://pgrouting.org/
In figure 1 are screenshots of a pair of the smartphone apps, which have been used for the data collecting experiments. The one pictured left is "Elbil parat" (are you ready for an Electric Vehicle), the one on the right is from "Herning cykler til Månen" (Herning bikes to the moon).

In section two an overview of related scientific work is given. The congestion model and the implementation is discussed in section three. Section three also presents implementation details of the map-matching algorithm. Section four presents the results of using the model for the megacity of Istanbul as well as the results for applying map-matching to find origins and destinations. Section five describes lines of further research. This paper extends the work presented in the workshop paper (see Lehmann and Gross (2016b)).

Figure 1: Screenshots of data collecting smartphone apps

2. Related work

The research presented in this paper combines two different scientific fields, the field of crowd sensing for creating demand data, and the field of route choice modelling from transport research.

In Froehlich and Krumm (2008) the authors describe how to convert GPS data in to trips and routes. This work uses calculated routes to predict current trips. By measuring the similarity of an ongoing trip with prerecorded routes, the most probable route for the current trip can be estimated.

For low coverage of traffic by GPS trackers Herrera et al. (2010) conclude that congestion can be predicted even with a small number (2-3% of total number of travellers) of speed reporting devices.

An overview of data collecting methods for origin destination demand matrix estimation is given in Bricka et al. (2014). The data collecting methods can be divided into self reporting methods (surveys, personal GPS receivers with self reported trip classification) and automatic methods (traffic counts, cell phone data, bluetooth readers, license plate recognition). The paper do not establish methodologies for converting the collected data into origin destination demand matrices.

Origin destination demand matrices can be derived from traffic counts as is shown in Yang et al. (1992). The derivation involves solving a Generalised Least Square problem.

Zhang et al. (2010) use information from mobile phone cell towers to create an origin destination demand matrix . The paper used a Horvitz-Thompson estimator to estimate the OD matrix. This work assumes that each cell phone trace generated from cell phone towers data is equal to a vehicle trace. The trace is made by triangulating between the connected cell towers. The authors take the cell phone market penetration (87% of the population in the researched area has a cell phone) into account and use the estimator to estimate the remaining population.

A similar approach is presented in Tolouei et al. (2015), where cell tower id's are used as a location identifier, in a project to verify the use of mobile phone data as a complement/ replacement of household surveys and traffic counts for generating origin destination matrices in a highway planning scenario.
In Jin et al. (2013) an indirect approach to creating origin destination matrices is presented. The paper analyses social media messages, that incorporate location and time information. The paper uses Foursquare data from the Austin, TX region. The data show good correlation with survey based estimation of origin destination matrices.

An example of using smartphones to track transit vehicles can be found in Biagioni and Gerlich (2011). The goal of the paper is to use the tracking to provide arrival time information to waiting transit passengers. The authors applies map matching to the tracked GPS data in order to correctly geolocate the transit vehicles.

In Lehmann and Gross (2016a) the accuracy of GPS positioning in an indoor stationary experiment is investigated. The main result shows that WiFi assisted GPS sometimes has large position errors than standalone GPS.

Driving modes are classified from smartphone accelerometer data in Lehmann and Gross (2017) for different passenger vehicles. The detection is performed by a K-means clustering algorithm.

In our project we plan to get information on traffic demands from travellers via apps installed on their smartphones. The data from the smartphones allows us to know the origin and destination of the route, but also the actual route which the traveller chose. The challenge from the collected data is to estimate the total traffic demand from the observed travelled routes.

Modelling of Route Choice was founded in the 1960’s. The basic textbook is written by Sheffi in 1985 Sheffi (1985).

The model for congestion used in this paper is presented in detail in two papers by Watling et al. (2015) and Rasmussen et al. (2015) by the inventors. This work presents two different implementations, one in MATLAB and one in ArcGIS (with extra extensions for route choice modelling). The authors report running times in the ArcGIS implementation for modelling the Copenhagen region, of about an hour on a normal sized desktop computer.

The paper Rasmussen et al. (2015) presents experiments using different error components to model reuse of links in different routes to show the versatility of the model, and performance impact of the different link reuse strategies.

From the review papers Prato (2009) and Prashker and Bekhor (2004), we have chosen to use the Path-size Logit method to model the influence of routes with overlapping links, due to good runtime and model performance. The previous traffic modelling research, as reported, has only been tried in small to medium sized cities (see Nielsen et al. (2002); Nielsen (2000); Bovy (2009)). The implementation presented in this paper makes it feasible to model the largest cities in the world as we show in the results section for Istanbul (25st largest city).

3. Methods and Materials

3.1. Congestion Modelling

The effect of congestion in road transport, is primarily that the travel time of a congested road segment increases as the traffic load approaches the traffic capacity of the road segment. The travel time increase can be modelled with the BPR (Bureau of Public Roads) Sheffi (1985) formula:

\[ t = t_0 \left( 1 + \alpha \left( \frac{f}{C} \right)^\beta \right) \]  

(1)

Where \( t_0 \) is the free flow travel time, \( f \) is the volume of traffic (traffic flow), \( C \) is the capacity of the road. The constants \( \alpha \) and \( \beta \) are country specific numbers capturing the way drivers react to congestion (the values of \( \alpha = 0.5 \) and \( \beta = 4 \) yield results that correlate well with observations). As can be seen from the formula, the travel time will stay at the free flow travel time until the flow is very close to the capacity. The travel time then increases rapidly as the flow increases above the capacity.

To model where congestion occurs, we are using the Route Choice Modelling framework. The Route Choice Model is based on the assumption that all travellers in a transport system are choosing the route in order to maximise the utility of the travel (or minimise the cost of their travel). The cost of the travel consists of actual costs for fuel, toll and wear of the vehicle, and perceived cost which is modelled as a Value of Time (VoT), associated with the travel time. The Route Choice Model seeks the equilibrium state, where all travellers travelling from the same origin to the

---

6https://en.wikipedia.org/wiki/Megacity
same destination have the same low travel cost. In this state of the traffic flow any change in choice of Route will increase the travel cost.

For the Deterministic User Equilibrium, the assumption is that all travellers have perfect knowledge of all routes and their associated cost. Each traveller will then choose the route with the lowest cost.

In the Stochastic User Equilibrium model, each traveller thinks that they choose the route with the lowest cost, but does not have perfect knowledge, and therefore there is a probability, that the chosen route is not the one with the lowest cost. This method leads to algorithms that consider a large number of routes with very low probability of being chosen, since all paths have to be considered.

The Route Choice in the Stochastic User Equilibrium is modelled by considering the travel times on the different routes and adding an error term, which models the uncertainty of the travel time for the route. A problem in the formulation of the SUE is that links can be shared between routes and this reuse of links must be taken into account. There has been proposed a large number of ways to model the reuse of links in the Route Choice literature Prato (2009). The probability of choosing the $k$'th route can be expressed by the Path Size Logit formula (2).

$$P_k = \frac{\exp(V_k + \beta_{PS} \cdot \ln(PS_k))}{\sum_{l \in C} \exp(V_l + \beta_{PS} \cdot PS_l)}$$

(2)

In equation 2 $V_k$ is the utility of the $k$'th route, $PS_k$ is a path size correction factor measuring the amount of reuse of links in route $k$. The $\beta_{PS}$ is a number controlling how much the Path Size similarity of the routes should affect the utility.

The $PS_k$ is given by :

$$PS_k = \sum_{a \in \Gamma_k} \frac{1}{L_a \sum_{l \in C} \delta_{al}}$$

(3)

In equation 3 $L_a$ is the length of the link $a$, $L_k$ is the length of route $k$, $\Gamma_k$ is the set of links making of route $k$, $C$ is the set of all links to chose from and $\delta_{al}$ is one if $l$ equals $a$ or zero otherwise. The number $PS_k$ is one, if none of the links in route $k$ is shared with other routes, and less than one for routes with shared links.

The Restricted Stochastic User Equilibrium in Rasmussen et al. (2015) combines the two mentioned equilibrium formulations by restricting the SUE to only consider a restricted number of possible routes.

The necessary data to model congestion with Route Choice is a digital road network, where each link (edge) is specified with free flow travel time and capacity. Further more an origin destination demand matrix, which specifies the travel demand in the network is needed.

In this paper the road network is created from the Open Street Map data, by converting the GIS data into a topology, to ensure that the routing functions can be used.

The demand data can be obtained in a number of ways. The use of surveys to create the demand data has been used Nielsen et al. (2002), as well as interviews combined with traffic counts in Nielsen (2000). We propose to use smartphone apps to facilitate creating the demand data. By having users participating in the generating of the demand data, the hope is to get more accurate and up to date information.

The users have to install a smartphone app, which will record GPS and accelerometer data and send the data to be stored on our servers.

Since it is not possible to get a complete coverage of travel data, for a specific area, due to not all travellers participating in the data gathering process, methods for generating the Origin Destination matrix from a small number of respondents, need to be created.

For Istanbul, where no smartphone apps were deployed, synthetic data was used. By using demographic data on the population densities of the municipalities of the Istanbul megacity, origins were chosen as single points near major roads, in the different municipalities. The destinations were chosen at the same points, and the origin destination matrix was created as a symmetric matrix with zeroes in the diagonal. The demand level were chosen to be the same for each origin destination pair.

For the danish cities of Aarhus and Herning we have collected data from participating users. In Herning users would use an app to measure their bike travels to participate in a municipality driven contest of "Biking to the moon". The sum of all bike rides in the municipality should reach the distance to the moon. Users would also allow the
Table 1: Data collected in Herning

<table>
<thead>
<tr>
<th>Number of data points</th>
<th>1.2 million</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>17</td>
</tr>
<tr>
<td>Number of trips</td>
<td>2179</td>
</tr>
<tr>
<td>Number of flights</td>
<td>11</td>
</tr>
<tr>
<td>Number of slow trips (less than 25 km/h)</td>
<td>1251</td>
</tr>
<tr>
<td>Number of map-matched trips</td>
<td>833</td>
</tr>
</tbody>
</table>

Researchers access to the data data from their other travel activities. The data from Herning is divided into single trips and the trips are truncated at the start and end to the closest traffic junction, to anonymise the user. The trips are then grouped by start area and end area, to create the origin-destination matrix. Within each origin-destination group the trips are sorted in time buckets to find rush hour patterns. To account for the incomplete coverage of data, the origin-destination data thus created has to be complemented with other data sources as census data, traffic counts and household surveys.

In Aarhus a project called "Are you electric vehicle ready" 7, participants first use our app to register the driving demand. After a month’s use of the app participants can borrow an electric car for a month to see how it is to drive an electric vehicle. The data from the app is used both to grade the suitability of an electric vehicle, but also for our research purposes.

Due to the nature of collecting data from users, without screening or surveying the objectives the users want to achieve by using the apps, we collect data for a variety of user scenarios. In the dataset collected in Herning, a large proportion of the trips are quite slow. As seen in table 1, 1251 out of 2179 trips have a maximum velocity less than 25 km/h. Some of these slow trips also goes through parks and can be discarded when we are looking for data about vehicle traffic. However the slow trips can still be valuable for determining the travel demand in the city, as they represent real user movements.

In the other end of the speed spectrum our data set also contains trips with very high speeds, which does not seem to follow the ordinary roads. As one can see from figure 2 these trips start or end at the local airport used by small amateur airplanes. We have filtered these flights out of the data set, as amateur flights are not relevant when considering either pollution from traffic or traffic demand.

Figure 3 shows the the position and speed data collected by our app in the municipality of Herning. The position data, shown as dots in the figure, has been partitioned into trips, and sorted to exclude data that does not concern road transportation. The average speed for the trips is shown for the cells in a one by one kilometre grid, by applying different background colours for the cells.

---

Figure 4 shows only data from trips, which are neither too slow (walking/running/biking) nor too fast (flights). The coloured boxes in the one by one kilometer grid shows the max speed. From the two figures 3, 4 it can be seen that the difference between average speed and max speed is small at highways (towards east and south-east), and rural roads (towards north, west, and south-west), and some what larger in the city proper which can indicate that congestion is a problem at times in the city proper.

The data collected from our crowd sensing apps, can be used to examine qualitative properties of the transport system. Examples of what low coverage data can be used for are:

- The origin and destinations of the collected trips are realised travel routes, and can be used as a seed for generating the origin destination demand matrix.

- A speed much lower than the speed limit, will indicate a congested link in the road network, and we can use occurrences of this low speed pattern as a check, since the model also must predict that the link is congested.

- The crowd sensed data also presents actually chosen routes, and thus these routes should be present in the chosen routes in the model.

The gathered data is biased towards passenger traffic due to the way participants are recruited. To persuade users to use the apps there must be a benefit for the user. This perceived benefit will create a bias in the group of people choosing to use a data providing app. As an example the app "Herning cykler til månen" (Herning bikes to the moon), will have a bias towards people who like to use their bicycles. The car driving data we receive from these participants, might not be representative for the population of Herning.

The low coverage is a problem for estimating travel demand. The data is very accurate and detailed. The origin and destination of a trip is clearly discernible, as is the chosen route, but the low number of users, does not provide enough data to generate the origin destination demand matrix directly. In combination with other data, i.e. census data, survey data, or traffic counts, it will be possible to generate the origin destination matrix.
4. Method for extracting OD demand matrix from smartphone data

4.1. Trip detection

To extract an origin destination from the acquired smartphone data, the data is segmented into trips. The acquired data is stored on a central server in a format resembling the packet format which the smartphone sends. The observations from the smartphone are collected into packets with multiple timestamped observations in one packet. The trips are found by assembling the data from all packets sent from a client and search for time gaps in the observations.

4.2. Filtering the data

The users in the experiment has applied the data collecting smartphone app to many different activities, beyond the original bike tracking purpose. This allows the researcher to obtain information on some other transportation behaviour. In this paper we focus on road transport, thus the obtained data set has to be filtered to exclude activities other than road transport in vehicles. The trips are filtered so that trips with a low maximum speed, which correspond to walking, running, and biking, and equally trips with very high maximum speed and a maximum altitude above 200m above the surface of the earth, are excluded.

4.3. Map-matching

In order to remove some of the noise in positioning introduced by the GPS receiver of the smartphone, and to transform the trips to routes in a road network, a map-matching algorithm is applied. The map-matching is a topological map-matching (see Quddus et al. (2007)), where the topology of the road network is guiding the choice of which road segments are part of a route.

The algorithm start by selecting a nearest road segment to every GPS position in a trip, into a primary set of road segments. Many of the chosen road segments are not part of the true route as every road segment in an intersection tend to be chosen. In figure 5 is an example of an intersection, where the red dots visualise GPS points. As some of the GPS points are in the middle of the road, the crossing road segment will be selected with a high probability.
Parallel road segments can also be selected to the choice set. The true route is found by finding the road segments that are connected to other segments in the set of chosen segments in both ends of the road segment, and exclude the rest. This will exclude all segments which are only connected to other segments in one end (orphaned segments) or not connected at all (parallel roads). The first segment and the last segment is also removed from the set of segments, but in order to ensure the privacy of the participants in the experiment, no further work has been done to include the true start and end segments. In figure 7 a small illustration is shown, for the intersection from figure 5. In this example both three green road segments and the four red segments has been selected by the initial nearest neighbour algorithm. The node are shown as black dots. The interior select algorithm selects three green road segments as they are the only segments which connect in both ends. For origin destination matrices the exact location of the origin and destination points are less interesting, as the origin destination matrix is used to investigate congestion in the road network between the origin and destination.
parallel roads are connected to the true route in both ends (see figure 6 where both ways around the round-about will be selected, and one way will later be rejected). The map-matching algorithm will then choose one of the alternatives by one-way attributes (useful for roundabouts), different length between the alternatives, different travel time, or the route which minimises the distance between the route and the acquired GPS locations.

4.4. Preparing the road map data

The road map data from OSM is converted in to a topology ready to import into a database by the program OSM2PO. The created file to be imported into the database, contains SQL commands to create the necessary tables and at the same time creates useful indexes.

4.5. Trip detection performed in a relational database

As explained above the first process is to assemble the observations across the received packets. This is performed by inserting each timestamped observation into a row in a relational database with the client id as field, and then performing a search with results grouped by the client id and sorted by timestamp. The result of this search is then joined onto itself on the condition that the two results are compared with one item shifted. Then a search for time gaps is performed by subtracting the two neighbouring timestamps and measure the size of the difference. If the difference is larger than five minutes a gap is recorded by storing the two positions (end of the current trip and the start of the new trip). The start and end trip for each client has to be manually added to the result table. The trip location data can now be sorted into a table of trips by once again join a table on itself shifted one row, this time the table containing start and end point.

4.6. Filtering of trips

The filtering of trips is performed by finding all trips with low max speed or high max speed and high max altitude, store these trips in a table and remove the trip locations from the trips table.

4.7. Map-matching database algorithm

The map-matching algorithm first select the nearest road to each position in each trip. The nearest neighbour algorithm is implemented as a function installed into the database system. As the nearest neighbour algorithm compares geometry fields in the roadmap table with the geometry object of the current position it is very important to have created an index (GIST index on PostgreSQL) on the geometry field of the road map table.

The selection of nearest road segments from the road map will create many orphaned segments, segments which only are connected to other segments in the trip at one end of the segment. By selecting only those segments which are connected in both ends, the interior part of the trace route is selected. The selection of interior segments is quite
complex as the interior segment under consideration has to be compared with two other segments. This can be solved by a three way join of segments chosen for each trip for each client. As many comparisons involve both client id and trip id, an unique trip id is created by concatenating the client id and trip number as strings.

The above map-matching algorithm can produce faulty routes for trips. The selected interior segments should be checked for gaps, which can occur if the sample rate is too low. The gaps can be found by starting at the road segment with the lowest timestamp, and find the segment in the interior segment set connected to the first segment. By continuing this process of finding connections, thus finding the transitive closure of the connection (in PostgreSQL this can be performed with a recursive query). If the transitive closure is not equal to the interior set one or more gaps exist which have to be closed by adding road segments to close the gap. By closing the gaps with shortest path search between segments which have timestamps close to each other a chronological route creation is ensured.

4.8. OD matrix generation

When the chosen interior segments of each trip has been selected, checked, and completed, the origin point can be found as the unconnected end of the interior segment with smallest timestamp and the destination point can be found in the unconnected end of the interior segment with largest timestamp.

If there are origin and destination points in the vicinity (the vicinity is dependent of the size of the city, for Herning the vicinity is within a block, for Istanbul it would be ten to twenty blocks) of each other they should be combined to one point preferably place in a node in the road map as close to the centre of the points. After all origin destination points have been found the demand can be distributed to the origin point. The demand of each origin can be estimated with power law related to the distance (see Liang et al. (2013)).

5. Results

5.1. Dataset

The smartphone data used in this paper, was collected from users of the "Herning cykler til månen" app (see figure 1). The data set contains 1.2 million position points from 17 users covering over 2000 trips. The road map used is from the OSM project, and it contains almost 56 thousand road segments for Herning. The two-way streets are converted into one-way streets which brings the number of road segments to almost 106 thousand road segments. To help the performance of the queries performing on the road data a geometry index is created on the geometry column, and btree indexes for the of source and target nodes.

The Istanbul road network consists of over 300 thousand bidirectional road segments. These segments are expanded to unidirectional arcs to make sure that the routing algorithm does the routing adhering to normal traffic rules. For the transport model there is no need for an index for the geometry column, because it is only used to drawing maps. The transport model performs the shortest path searches on a graph representation of the road network, which the pg_routing extension builds internal in the database system.

5.2. Extracting OD matrices

In figure 5.2 the result of the map-matching algorithm is shown. The map-matched trips are drawn in different colours. The important thing to notice is that almost all of the minor roads in the map are not coloured, which they were before the selecting of the interior connected road segments. The selection process eliminates all the falsely selected side roads and cross roads. As the sample rate of the data in the experiment is rather high, there are no gaps in the finally selected roads. The map-matched trips seem to cover most of the city even though the trips are acquired from only 17 users. All the major residential areas seem to be covered by trips. A special case has to be made with respect to the highway in the right hand side of the map since the map is cut off in the middle of the highway. There is no population density at the cut off points, we can use as an estimate of the transport demand along the highway. But there are more trips going along the highway, which propose a high demand.

The property that data from a small fraction travellers can add qualitative knowledge about the condition of a road network is inline with the claim of Herrera et al. (2010), data from only 2-3% of travellers was enough to detect congestion. This is also a possible conclusion of the difference between figure 3 and figure 4. The clear difference between the average speed in the city and the maximum speed in the city could be interpreted as the max speed is
measured in non congested periods of the day, while the average speed will be sensitive to slow down during congested periods.

5.3. Modelling traffic

All model simulations were run on a MacBook Pro 15” with Intel 2 GHz Core I7 processor. The database system was chosen to be PostgreSQL version 9.4.1, with the extensions PostGIS and pgrouting. The road data for the model is from the Open Street Map project (OSM). OSM is a crowd sourced map database. Anyone can create a user account and start create or update map data. When data is added to the database the committed changes will be made visible at the next update, and other OSM users and automatic rule checkers will review the changes. OSM provides several different ways to add and retrieve map data.

To convert the data from OSM to a topology importable by PostgreSGL, the OSM2PO was used.

The result of the route choice model for Istanbul can be seen in figure 9, the road chosen by the route choice model is shown in light green. The Bosphorus strait is visible in the centre with the two bridges connecting Europe and Asia (a third opened to north outside of the map in 2016). The demand is simulated, by choosing 46 points in Istanbul as origins. The points were chosen at central locations in different parts of the city, where the traffic could easily be dispersed without creating congestion at the origin. The points represent the traffic demand from an area around the chosen point. Each origin creates traffic demand to all other origins, with a constant demand, to create an origin destination matrix with 2070 non zero entries.

The runtime of the algorithm is visualised in figure 10.

The runtime of an iteration increases as the number of iterations increases. The main contributors to the runtime are the shortest path searches (we are using the \textit{kDijkstra} function from pgrouting), and the test to see if a new found

---

8\url{http://www.postgresql.org}
9\url{http://www.openstreetmap.org}
10\url{http://osm2po.de}
shortest route is already in the set of possible routes. This test compares the hash value of the new route to the hash values of already found routes. As the number of routes increases the number of tests for determining if the found shortest routes are already in the choice set also increases.

There is a possible performance improvement by adding indexes to some of the tables in the database.

In figure 11 the convergence of the algorithm is shown. The error is a measure of the differences in travel times for the different routes for each origin destination pair.

The x-axis is the number of iterations, the model has been run through, and the y-axis is the logarithmic view of the error.

The figure shows how the error decreases fast in the initial steps of the algorithm. The error further decreases as the algorithm progresses. The sudden increases in the error is when a new route is found, with a lower travel time. The new route is added to the set of routes without any traffic assigned. The traffic distribution for that origin destination pair is thus unbalanced, and hence contribute with a large error. As the traffic volumes for the different routes are redistributed, the travel times of the routes move closer to each other, thus reducing the error.

6. Further work

To further improve the congestion model, we plan to expand the model to more kinds of traffic, by integration of commercial transport of goods, and public transport into the traffic mix.

To create the origin destination demand matrix from the collected data, we want to create a stochastic model for generating the traffic demand, guided by knowledge gained from local observers, census data, transit data, and crowd sensed data.

Including turn delays and intersection modelling Nielsen et al. (1998) would further increase the accuracy of the model.

There is a possibility for improving the runtime of the model, by utilising a recent feature of pgrounting. With this feature the shortest path for the origin destination matrix can be found in a single query to the PostgreSQL database system. As implemented in this work the shortest paths are found for one origin to all destinations at a time. This forces the pgrounting function to rebuild the graph of the road network for each origin in each iteration.
7. Conclusion

In this paper we have presented the efficient database implementation of a congestion model based on the Restricted Stochastic User Equilibrium algorithm. We propose to extend this model by adding routes derived from crowd sensed data to the starting choice set, instead of relying solely on shortest path algorithms.

Using a database has been a fortuitous choice, since the performance of the database has resulted in acceptable run times of the algorithm, without having to look for advanced performance improving techniques. This also means that it is highly likely that the performance can be drastically improved, by a careful examination of the queries used.

The congestion modelling can be used for a number of purposes: for transport and urban planning, the model can be used to show effects of new roads, planned road works, or increased traffic. The model result can also be used as input to Environmental and Air Pollution models (see Berkowicz (2000)), to further increase the accuracy of such models.

Acknowledgment

This work has been supported by The Danish Council for Strategic Research as part of the EcoSense project (11-115331), and the BIOGAS2020 project supported by the European Regional Development Fund for Oresund-Kattegat-Skagerrak, EU Interreg (NYPS 20200437).
References


