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 in to 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.

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:

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

(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 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.

\[1\]https://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\textsuperscript{th} route can be expressed by the Path Size Logit formula (2).

\[
P_k = \frac{\exp(V_k + \beta_{PS} \cdot \ln(P_{Sk}))}{\sum_{i \in C} \exp(V_i + \beta_{PS} \cdot P_{Si})}
\]  

In equation (2) \(V_k\) is the utility of the k\textsuperscript{th} route, \(P_{Sk}\) 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 \(P_{Sk}\) is given by:

\[
P_{Sk} = \sum_{a \in V_k} \frac{L_a}{L_k} \sum_{l \in C} \delta_{al}
\]  

In 3 \(L_a\) is the length of the link \(a\), \(L_k\) is the length of route \(k\), \(V_k\) 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 \(P_{Sk}\) 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 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 [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 3, with the extensions PostGIS and pgrouting. The road data for the model is from the Open Street Map project 4(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 PostgreSGL, the OSM2PO 5 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.

3http://www.postgresql.org
4http://www.openstreetmap.org
5http://osm2po.de
The runtime of the algorithm is visualised in figure 4.

![Figure 4. Runtime versus iterations](image)

The figure shows the runtime for one iteration as a function of the iteration number.

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 `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.

![Figure 5. Convergence of the model for the Istanbul case](image)

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.

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

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 Øresund-Kattegat-Skagerrak, EU Interreg (NYPS 20200437).
REFERENCES


