Revealing reliable information from taxi traces: from raw data to information discovery

Anja Keskinarkaus
Center for Machine Vision and Signal Analysis
University of Oulu
Oulu, Finland
anja.keskinarkaus@oulu.fi

Ekaterina Gilman
Center for Ubiquitous Computing
University of Oulu
Oulu, Finland
ekaterina.gilman@oulu.fi

Lauri Loven
Center for Ubiquitous Computing
University of Oulu
Oulu, Finland
lauri.loven@oulu.fi

Satu Tamminen
Biomimetics and Intelligent Systems Group
University of Oulu
Oulu, Finland
satu.tamminen@oulu.fi

Marjo Hippi
Meteorological Institute
Helsinki, Finland
marjo.hippi@fmi.fi

Gang Xiong
Cloud Computing Center
Chinese Academy of Sciences,
Dongguan, China
gang.xiong@ia.ac.cn

Fenghu Zhu
Institute of Automation
Chinese Academy of Sciences
Beijing, China
fenghua.zhu@ia.ac.cn

Tapio Seppänen
Center for Machine Vision and Signal Analysis
University of Oulu
Oulu, Finland
tapio.seppanen@oulu.fi

Jukka Riekki
Center for Ubiquitous Computing
University of Oulu
Oulu, Finland
jukka.riekki@oulu.fi

Susanna Pirttikangas
Center for Ubiquitous Computing
University of Oulu
Oulu, Finland
susanna.pirttikangas@oulu.fi

***Abstract*— In this paper we present procedures for processing raw data collected with moving vehicles and for fusing this data with digital map data. The goal is to have a better understanding of the city traffic via quantitative research on collected taxi data in relation to digital map properties. Map attributes are provided by Digiroad, which is a database of Finnish road and street network. We define methods to clean up data that has been collected with taxis equipped with on-board vehicle tracking devices from real customer service situations. Consequently, the driving behavior may be inconsistent and sensor data can be limited and contain errors. We explain procedures of preparing data; filtering the most obvious errors from the data set, map-matching moving object data, and fetching map attributes along the routes of the moving vehicles. The fetched properties, as well as other measurement data, are used for deriving statistics and illustrations to study driving behavior in downtown Oulu, Finland.**

***Keywords*— digital maps, GPS data, regression analysis, taxi trajectories, data fusion**

I. Introduction

Traffic data collected by taxis can be utilized by various services, like route planning, traffic flow analysis and urban computing, as well as services promoting ecological driving habits. Zheng *et al.* [1] mine taxi trajectories for supporting urban planning, for sensing people's mobility in a city with taxis and detecting salient traffic problems. Zhu *et al.* [2] explore urban mobility by analyzing spatio-temporal patterns from taxi trajectories for better understanding of urban structures and movements. Zhou *et al.* [3] study functionally critical network locations from people's moving trajectories using taxi trajectories. Li *et al.* [4] use taxi trajectories and information around interesting intersections are segmented out to explain patterns behind traffic flows. Taxis serve a lot of customers, and do not follow predictable routes, opposite to public transportation. Moreover, private vehicles often follow similar routes due to daily routines. Consequently, taxi data offer rich information for multiple purposes.

In this study, we utilize traffic data collected by taxis to better understand traffic in urban areas, constructed and equipped with different landmarks, like crossings, traffic lights and bus stops. All applications using taxi data require data preparation as the initial step. Moreover, all services rely on the quality of the collected data. Low-quality data can easily lead to low-quality analysis and consequently produce flaws in services and applications. Furthermore, any explanatory analysis benefits from cleansing.

Specific problems related to the nature of the collected data need to be tackled when vehicle data is analyzed. Therefore, we also explain data gathering and preparation processes required to analyze traffic related data of real taxi trips. The methodology is developed for taxi data, but is targeted to generic features, like traffic speed. We are interested on city traffic's dependencies on the underlying road network and its' features (crossings, traffic lights, etc.). Here we take advantage of the Finnish national digital map database; we process this data to make it more suitable to our purposes and extract attribute data from it.

Overall, the contribution of this article is follows: The preprocessing step serves an important role before the actual analysis, as the range of actions performed at the preprocessing step filter out errors, as well as other

This research work has been financially supported by Academy of Finland UrBOT project (323630), by Academy of Finland 6Genesis Flagship (318927), NSFC (Natural Science Foundation of China) projects (71232006, 61233001, 61174172), and Chinese Dongguan's Innovation Talents Project (Gang Xiong)

properties otherwise effecting the analysis. We define an area of interest in an urban area; define important exit-entry point pairs in the selected area; preprocess the data, and analyze relations between gathered data and map features. Lastly, we apply mixed models to identify associations between map features and driving speed.

This study will contribute to our future work, which is part of a large effort on instrumenting City of Oulu in Finland, collecting rich data sets, and developing novel solutions for recognizing interesting phenomena from the collected data (Fig. 1).

The paper is organized as follows: Section II gives an overview of the related work. Section III describes the collected data, Section IV describes the overview for the study; initial data source and the map database used in the experiments as well as processes for cleaning the data. In Section V, we explain the related analyzing approach. In Section VI, we illustrate results mined from the example data. Finally, in Section VII, we conclude the article.

II. Related work

In many of the related studies, the aim is to make the taxi service more fluent and efficient. Li *et al*. [5] evaluate taxi traces to discover hotspots based on taxi pick-up and drop-off events, to help taxi drivers to find the next passenger. Yuang *et al*. [6] use taxi data analysis to suggest routes that are the fastest in practice. Gilman *et al*. [7] fuse real driving behavior data from taxicabs, weather, digital map, and traffic situation information to gain understanding of how the routes are selected and what are the effects in terms of fuel-efficiency.

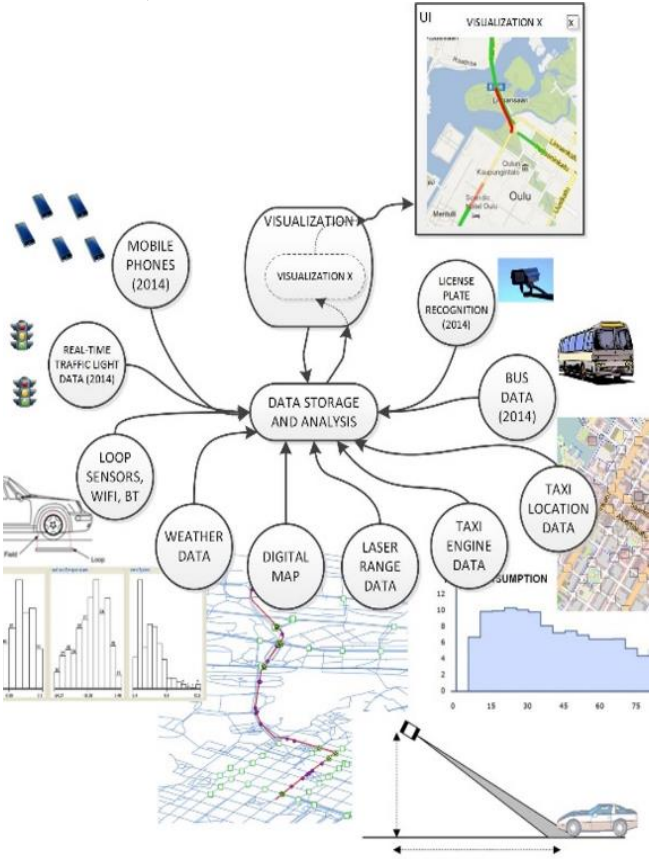

Fig. 1. Infrastructure of studies in City of Oulu region

Kong *et al*. [8] propose a service recommendation model; Time-Location-Relationship to improve taxi drivers' profits by integrating, processing and analyzing taxi trajectory data. Tang *et al*. [9] investigate spatial heterogeneity between travel demand and various variables using taxi GPS trajectory data. Phiboonbanakit and Horanont [10] concentrate on improving taxi driver profits with concentration on traffic congestion problem. Wang *et al*. [13] introduce an application to visualize hotspots using taxi trajectory data and information about passenger pick-up and drop-off points.

Taxi data is considered to be reasonable to represent intra-city spatial interactions and to reveal city structure. Liu *et al*. [11] use taxi data to gain better awareness of the city traffic; hotspots are detected on statistics derived from the gathered data. Liu *et al*. [12] explore travel patterns and city structure using taxi GPS trajectory data. Furthermore, Kong *et al*. [14] use GPS equipped taxis for traffic state estimation for urban road networks. In overall, the existing research is classified by Castro *et al*. [15] to three categories; social dynamics, traffic dynamics, and operational dynamics.

Alvarez *et al*. [16] propose a method to analyze trajectory data using geographical information. Moreover, Jiang *et al*. [17] study the type, characteristics and reasons for errors in traffic data collected by sensors, and a linear interpolation method is proposed to restore the lost data. Li *et al*. [18] analyze specifically the frequency of driving different routes. Zhang *et al*. [21] study a method to clean and repair probe vehicle data. Phiboonbanakit and Horanont [10] apply statistical rules for cleaning existing errors and outliers. Map-matching is one specific area of interest. Lou *et al*. [19] concentrate on improving map-matching of low-sampling-rate GPS data. In overall, as shown by Wang *et al*. [20], large-scale GPS probe data present challenges and opportunities for numerous applications.

Minett *et al*. [24] use Origin-Destination point analysis, the routes are preselected, with appropriate properties, like similar length and amount of data available. Our work differs from the related work in that taxi drivers freely selected the routes between origin and destination, based on their own silent knowledge and intuition in particular driving situations. Consequently, problem framework is realistic, providing valuable information on the actual process required to collect data from the real world, and to pre-process it to find useful knowledge from the data.

III. Data

In this study, we work with taxi traces data, as well as with geospatial information from digital map.

Taxis, with on-board Driveco devices (http://eco.driveco.fi/www/), have been used to collect data. The information from the measurement device, GPS locations and car's OBDII diagnostics are retrieved via HTTP interface from a server, check details from [7],[31]. The gathered data consist of trips, where each individual trip is defined as a run between two consecutive events of turning off the engine. A trip is identified with trip id and other measurement data including start and end time of the trip, start and end route point, total time (s), total distance (m), total fuel (ml). A trip data set includes a collection of route points, each containing the properties measured at a specified time. There is no specific sampling rate for the route points, but a route point is generated when some significant change in the driving behavior, such as a turn, is registered. The vector of properties related to route points stores point id, trip id, latitude, longitude and start time, to give examples. The data in this study contains the information gathered with seven taxis driving in Oulu area during 1.10.2012-31.9.2013, as a basis for the study almost 30000 taxi trips are considered.

We fetch the road geometry and attribute data from the Digiroad database of Finnish road network maintained by National Land Survey of Finland, the Finnish Transport Agency and individual municipalities [22]. The road network consists of traffic elements, which are the smallest units of centre line geometry of the road. These traffic elements have unique identifiers and characteristic attributes, such as coordinates, functional type, length, and digitization direction. Segmented line-like attribute data refer to the data objects that are described as line segments. Road address and speed restrictions are examples of segmented line-like attribute data in Digiroad. In overall, the map database contains rich information at three levels; the geometry of the roads, the objects of the transportation system, like bus stops and traffic lights, and the properties of the roads, updated several times a year [23].

IV. FUSING ROAD NETWORK AND DRIVING DATA

A. *Map preparation*

In the road network graph roads represent graph edges and road intersections represent vertices, formulated with G={V,E}, where *V* refers to intersections and *E* edges between intersection. Accordingly, we reconstruct the road network graph to have one traffic element within each edge.

First, we construct a table to identify the type of the endpoints of the traffic elements either as junctions (when at least three traffic elements have the same endpoint) or intermediate points (only two traffic elements have the same endpoint). In Digiroad map, the edges are a collection of traffic elements touching each other, and accordingly an edge may contain more than one traffic element. After this procedure, we can identify vertices (junctions) of the road network graph, as well as edges (sequences of traffic elements between two junctions).

A clip of this data is presented in Table 1. The table is enriched with a geometry line constructed from the contributing traffic elements, digitization information fetched from the map database, traffic flow direction and additional field information for map matching purposes. The final road network graph is constructed from this table, where edges are single elements created from an array of smaller traffic elements (elements column in Table 1) and vertices presented with edge intersections.

TABLE 1. EXAMPLE OF JUNCTION PAIRS (EPSG:4326 COORDINATE SYSTEM)

Junction1 geometry(Point,4326)	**elements integer[]**	**Junction 2 geometry(Point,4326)**
POINT(25.5244, 65.0252)	{121499}	POINT(25.524, 65.025)
POINT(25.4650, 65.0353)	{138854,138855, 122734}	POINT(25.464, 65.035)
POINT(25.4558, 65.0434)	{121427,121426}	POINT(25.460, 65.043)

B. *Data cleaning*

Taxi data is retrieved from real sensors and can contain errors, hence data pre-processing has to be conducted. A trip consists of route points where timestamp $T_{n-1} < T_n$ if no errors happen. However, due to occasional latency variation, the data obtained from the measurement device (id, timestamp) may arrive at the server in an incorrect order. To tackle this challenge, we sort the route points into two sequences: by their id and by their timestamp. Then, the overall distance of the trip is calculated for both sequences. The one with the smaller length is judged as the right sequence. Finally, all the corresponding properties are aligned with respect to the correct sequence to guarantee monotonic increase.

C. *Data segmentation*

Taxi drivers demonstrate quite different behavior in comparison with ordinary drivers, and accordingly we suggest methods to handle especially taxi data. One of the biggest differences is that they can drive almost the whole day without turning off the car engine. We are interested in individual trips, like taking customers to their destinations. We divide the overall trip into trip segments with an algorithm that applies time-based segmentation (Table 2).

The rationale behind these rules is as follows: The overall region is small and having long stops on traffic lights is a rare situation. The traffic lights are programmed so that in an error situation one has to wait at most 200 seconds in the traffic light. After this, the red light is switched to a blinking yellow. Otherwise, in an unfavorable situation, the waiting time is 50-60 seconds. Driving a long distance with a very low speed is unlikely. Finally, all trip segments containing less than five route points and longer than 30 km are removed from further analysis. Having five measurements for the whole run may give poor information to analysis. Trips longer than 30 km are unlikely in the local region.

D. *Origin-destination segment selection*

We follow the Origin-Destination based approach similar to that proposed by Minett *et al*. [24]. The difference is that we work with real taxi data where we don’t have any control over the route selection. For this analysis, all the trips obtained and processed as indicated in previous sections are used. We selected three locations (i.e. road segments) which we name by the region names with T, S and L. All of these are located at the key enter and exit points of downtown Oulu.

TABLE 2. SEGMENTATION RULES

1	If the distance between route points does not change within three minutes then it is a stop
2	If the distance change between route points is less than three km within the time more than seven minutes then it is a stop
3	If from one route point to another one the car has moved with the speed less than 0.002 m/s, then we consider this as a stop
4	If the car moved less than 3km in more than 15 minutes and the speed was more than 0.002 m/s we consider this as a stop
5	If after the first round, there are still trips longer than 40km, we try to split these with the rule 1, having 1.5 minutes’ interval

The red arrows in Fig 2. indicate the selected traffic flow directions of interest. We are interested in trips that contain route points near the origin and destination roads. In addition, the trips have to cross these roads in particular order in time (first origin, then destination). We chose to utilize a “thick geometry” approach (see Fig. 2) where the Origin and Destination roads are artificially made thicker to catch the routes significantly deviating from the original roads.

First, only the routes, which intersect the “thick roads” on an angle within a predefined range, are considered further. Table 3, second column, lists the number of trip segments for each taxi that fulfilled this condition. S-T transitions is expressed in Table 3, column 4. We ensure that the transitions happen within the central area. Hence, we filter out all the transitions, which occur outside (Table 3, column 5). We clean the trip segments to make sure that only the routes, which intersect two or more different roads, are considered (Table 3, column 3). A route between an Origin-Destination pair we call as a transition and the number of T-L, L-T, T-S, Finally, these T-L, L-T, T-S, S-T transitions within central area are map-matched and post-filtered to make sure that start and end route points of the transitions are close to the Origin-Destination roads of Fig. 2 (Table 3, column 6)

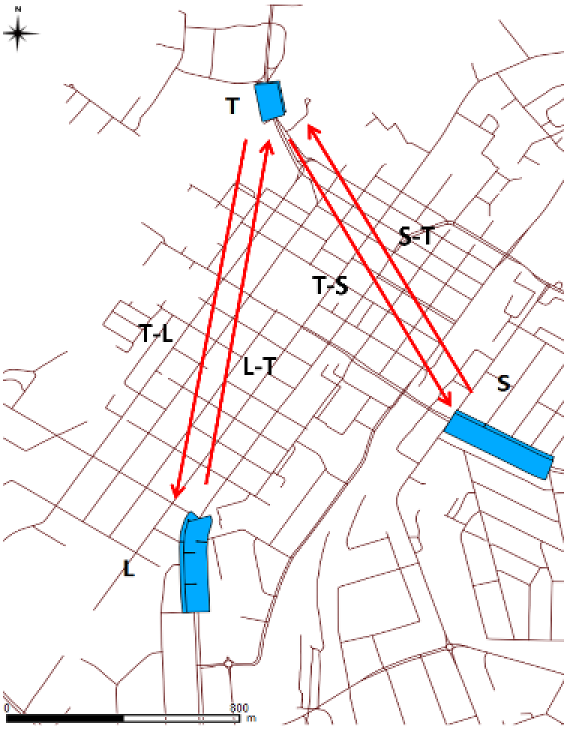

Fig. 2. Selected origin-destination pairs and thick geometry visualization.

TABLE 3. MAP MATCHING THE TRIP SEGMENTS

Car	Trip segments (total)	Filtered and cleaned	Transitions total	transitions within city centre	Post-filtered T-L,L-T,T-S,S-T transitions
1	2409	636	89	79	65
2	3068	1282	172	156	128
3	1790	447	44	32	19
4	2486	622	102	93	73
5	2429	616	88	75	65
6	1815	625	113	108	96
7	4080	1109	162	131	98

E. Map-matching

Map matching is a way to align collected GPS locations on a digital map. There exists a big difference on the matching algorithm performance depending on the GPS accuracy and the sampling rate. In our case, the sampling rate is not even, but location information is only stored when significant changes occur. For map-matching, we use incremental map-matching algorithm [25] enhanced with information retrieved from the digital map (like road directions). Furthermore the Dijkstra Shortest Path algorithm from pgRouting is utilized to fill the gaps, when data points are too far from each other, check details from [7],[31]. Only cleared and filtered transitions going through the city centre are map-matched (Column 6 of Table 3).

F. Fetching attribute data from digital map

The digital map attribute data may explain the driving behavior. Accordingly, we use a trip identifier (trip id) together with the start time of the trip as a unique identifier of a transition. The road network traffic elements of the transitions are identified by map-matching procedure, hence respective digital map attribute data for transitions can be retrieved. We extract the information about the number of junctions, number of pedestrian crossings and number of traffic lights for transitions within the region of interest.

G. Tools

Driving data from taxi cars is retrieved with a program written in Java that pools information from the dedicated service. Retrieved data are stored in PostgreSQL 9.1 DBMS having PostGIS extension that allows manipulation with geospatial information. The road network graph is stored in the same database. SQL and PL/pgSQL languages of PostgreSQL DBMS are used to pre-process and manipulate the data. To visualize results, Quantum GIS is used.

V. DATA ANALYSIS TECHNIQUES AND METHODS

We use 200 m x200 m grid with even sized grid dimensions, as related to the nature of selected map features, it would be expensive and time-consuming to create a more complex spatial element structure. Furthermore, the dimension has been selected to have enough measure points on the individual cells, as well as to be meaningful to capture effects of multiple map features.

To identify associations between map features and driving speed, we apply mixed models. Mixed models extend linear regression, a method commonly used to seek for a model to describe relationship between response variable and a predictor value. The formulae below [26] characterizes the linear relationship

$$Y = Xb + \varepsilon, \quad \varepsilon \sim MN\left(0, \sigma_{\varepsilon}^{2} I\right) \tag{1}$$

,where Y is the response variable and X the model matrix consisting of the intercept and the values of the explanatory variables. Here, b is here the vector of regression coefficients to be estimated. In our case, each Y_i is the point speed and in addition to the intercept, X may include the respective 200 m x200 m cell identification factor as well the map features such as the number of traffic lights, bus stops, pedestrian crossings or crossings for the cell.

Mixed modelling sets further Gaussian priors on some or all the coefficients, treating them not as fixed but random. In effect, it thus regularizes the model, borrowing information from the cells with a lot of data to those with little data [26]. The model equation thus reads

$$Y = Xb + Zu + \varepsilon,\ u \sim MN(0,\sigma_u^2 I),\ \varepsilon \sim MN\left(0,\sigma_\varepsilon^2 I\right) \qquad (2)$$

,with Z denoting the random effect values and u the vector of random effect coefficients.

VI. RESULTS AND INDICATIONS

The retrieved and processed data allow us to inspect the driving behavior in downtown Oulu. Intuitively we expect that there is a relation between map attributes and driving behavior, like the location of the traffic lights and average speed.

A. Features analysis

First, we studied some features expressing the cleaned and preprocessed data. In overall, the data contains 30469 measured point speeds. The overall speed figures for taxi 1 (4186 measured speed points) are shown in Fig. 3, with colours illustrating speed values. Furthermore in Fig. 4 is illustrated the effect of driving direction to the results. The speed data results for taxi 1 when data is categorized according to the direction (T-S, S-T, T-L, L-T) of the route. Especially in northern countries, there exist clearly separate seasons. Accordingly, we studied the seasonal variations in the collected data. Similarly to directional differences the seasonal variances in speed are illustrated in Fig. 5 for taxi 1. We calculated statistics on the routes by fetching attribute data from the prepared Digiroad digital map along the cleaned and preprocessed driven routes (See Table 3). We studied low speed, as it is one of significant factors affecting vehicles' fuel consumption and gas emissions [22].

The statistics on the route distance, normal speed (speed at the speed limit), low speed (less than 10km/h) and fuel consumption are derived from the pre-processed trip data. The number of bus stops along routes is not calculated because the current map does not give information about the direction of a particular bus stop. The results of our inspections are presented in Table 4.

It can be derived that routes S-T and T-S contain a greater proportion of low speed than T-L and L-T. Proportion of normal speed is contrariwise. Low speed also correlates to fuel consumption, supporting findings in literature [28]. In Table 4, the mean value of traffic lights and junctions is almost the same for each Origin-Destination pair. Accordingly, the count of traffic lights, does not itself explain the difference in low speed for T-L, L-T versus T-S, S-T directions.

To study the issue further, we calculated the average speeds within cells of size 200 m x 200 m to find out areas where the average speed is considerably lower than elsewhere and whether this reflects the appearance of four selected features: traffic lights, bus stops, pedestrian crossings or crossings in overall.

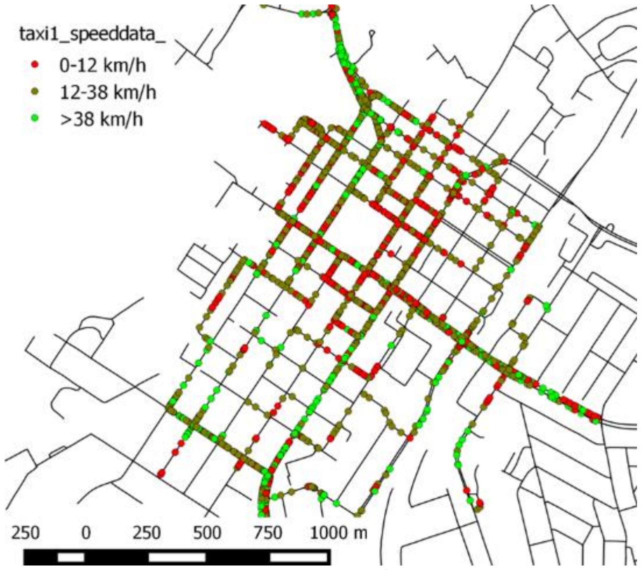

Fig. 3. Cleaned and preprocessed speed data for taxi1

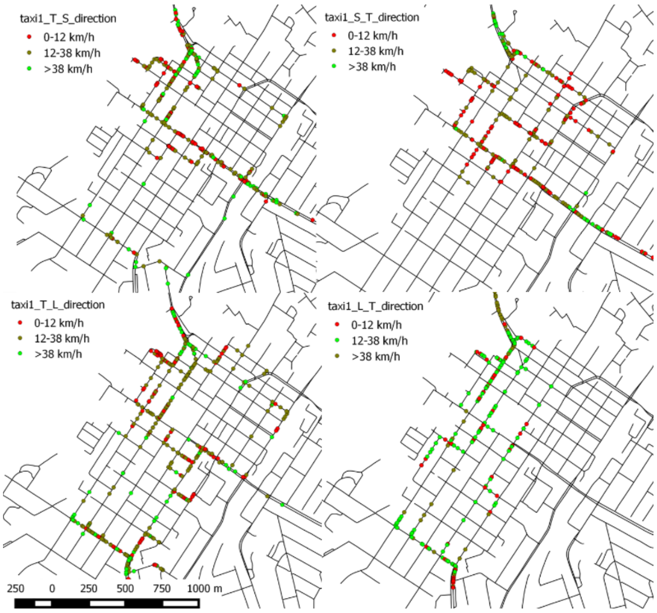

Fig. 4. Taxi 1 data categorized according to the direction.

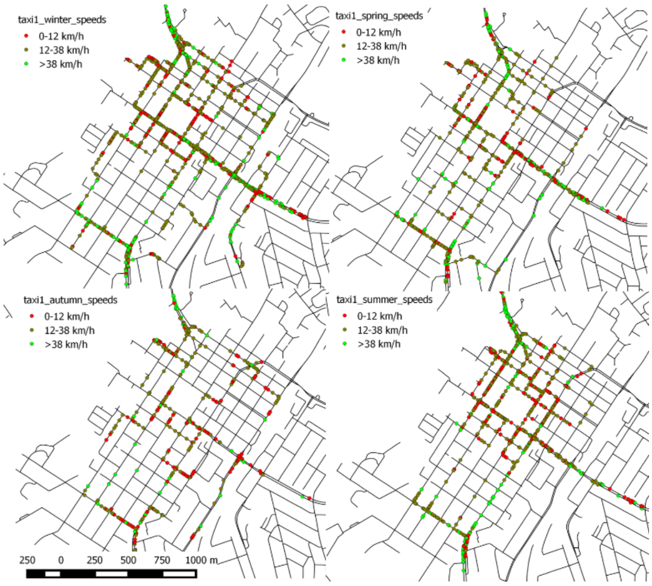

Fig. 5. Taxi 1 data categorized according to the season.

The speed data is illustrated with statistics of the number of the four features in the map in Fig. 6 for L-T direction as an example. In the study area the overall number of traffic lights, bus stops, pedestrian crossings, and non-pedestrian crossings is {67,48,293,271} respectively. The illustration shows that for example traffic lights affect the speed values, but simultaneously it can be seen that the relations are not self-evident. This is furthermore inspected in Table 5. The results show that traffic lights decrease the average speed, similarly to bus stops. In cells with no traffic lights or bus stops the variance of values is much higher. In the evaluation, there was no particular differences for L-T, T-L, S-T, T-S directions. In winter season the average decrease of speed was -0.07 km/h, in spring 0.46 km/h increase, in summer 0.70 km/h increase and in autumn 1.38 km/h increase when measured against the averages over the whole year.

TABLE 4. SUMMARY STATISTICS OF THE SELECTED FEATURES.

	Route	Min.	1st Q.	Med.	Mean	3rd Q.	Max.
Route time (h)	T-S	0.058	0.089	0.120	0.153	0.188	0.458
	S-T	0.047	0.106	0.124	0.135	0.152	0.347
	T-L	0.041	0.071	0.083	0.107	0.127	0.393
	L-T	0.041	0.078	0.100	0.114	0.136	0.388
Route dist. (km)	T-S	1.457	1.742	1.968	2.377	2.448	7.079
	S-T	1.664	1.987	2.148	2.319	2.465	5.708
	T-L	1.721	1.932	2.070	2.214	2.267	5.189
	L-T	1.849	2.131	2.266	2.377	2.447	6.089
Low speed %	T-S	0.2	24.2	38.3	38.2	51.8	72.6
	S-T	0	23.2	34.0	33.3	42.3	71.0
	T-L	0	7.5	20.9	23.3	37.0	76.6
	L-T	0	5.6	24.2	24.2	38.3	64.8
Norm. speed %	T-S	0	1.5	3.7	6.4	8.0	32.9
	S-T	0	3.8	6.6	8.8	11.4	40.0
	T-L	0	2.4	7.4	14.7	19.8	91.1
	L-T	0	4.2	8.5	14.5	18.6	98.6
Traffic lights	T-S	2	5	7	8	9	22
	S-T	1	3	5	5	7	13
	T-L	2	5	7	7	9	18
	L-T	1	4	8	7	9	15
Junction	T-S	16	18	21	23	24	52
	S-T	17	20	22	23	24	47
	T-L	15	18	20	22	23	47
	L-T	19	22	23	24	26	43
Pedestr. crossings	T-S	4	6	7	9	10	25
	S-T	4	6	7	8	8	21
	T-L	5	6	7	8	9	18
	L-T	5	6	8	8	10	23
Fuel cons. (ml)	T-S	106.3	176.5	216.9	264.9	302.3	823.9
	S-T	107.9	178.9	217.3	239.8	302.2	536.3
	T-L	76.68	1361	177.3	212.	2502	989.2
	L-T	92.3	168.2	210.6	231.2	273.1	766.1

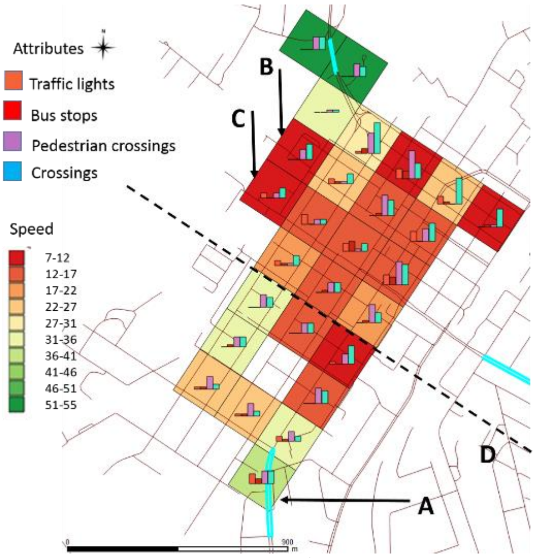

Fig. 6. Average speed and map properties for L-T direction.

TABLE 5. SUMMARY STATISTICS OF THE EFFECT OF MAP TRAFFIC LIGHTS AND BUS STOPS ON AVERAGE SPEED.

	Number of traffic lights =0	**Number of traffic lights and bus stops =0**	**Number of traffic lights and bus stops >0**	**Number of traffic lights>0**
min	11.9600	11.96 (B*)	9.26	9.26 (C*)
max	53.2700	53.27	32.09 (A*)	32.09
mean	25.5273	29.2486	18.78	18.71
var	231.4873	303.49	49.8995	47.898

*Map attribute statistics for areas illustrated in Fig. 6 with A,B and C; the number of traffic lights, bus stops, pedestrian crossings, crossings is {4,2,5,5} for A, {0,0,4,6} for B and {2,0,2,4} for C.

B. *Mixed models analysis*

Moving to mixed models, we take a look at the average point speeds for each cell. In the regression analysis we have excluded all the cells having no measurement points. The model we are considering is now

$$Y_i = \alpha_{Cell_i} + \varepsilon_i, \alpha_{Cell_i} \sim N(0, \sigma_\alpha^2), \;\; \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right) \qquad (3)$$

,with α_{Cell_i} indicating the random intercept value for each cell. As seen in the QQ-plot for cell intercepts in Fig. 7, with the exception of only the far edges, the Gaussian regularization indeed seems justified. Variances estimated by REML, the BLUP predictions for the intercepts for each cell appear as strong evidence of the effect of geography on the point speeds, with coefficients varying between ca. -15 and +20 km/h. As seen in Fig. 8, while the variation is large for some cells, for most cells the result is solid.

Plotting the results on map in Fig. 9, we can clearly see how the proximity of dead end roads areas reduces speeds on cells. However, the most interesting effects appear at the very center of the map (and the Oulu city), with speed reductions up to -8km/h. These are likely the result of both pedestrian movements and static map features such as traffic lights.

Referring to Table 5, for the case when the number of traffic lights is greater than zero, the average speed is in general lower in the corresponding cells. Accordingly, we expect that when a number of traffic lights grows, it evidently leads to higher fuel consumption on the driven route. The experiments showed that in some cases this holds true, as illustrated in Fig. 10. When the number of traffic lights is greater than 9 (an experimentally chosen boundary), in general there is an increase of low speed, also independent of the weather conditions. Some traffic lights are in average passed without having to stop, which is also reflected on the average values (Fig. 3).

From Fig. 6 it can be seen that compared to S-T and T-S routes, the routes L-T and T-L go through areas where more cells contain less features{10,14,110,82} (Fig. 6. below line D) This leads to proportion of normal speed to grow. Also, the hotspots, crowded areas with a lot of pedestrians moving, have an effect to the results. In Kostakos et al. [29], pedestrian movements were studied by calculating number of clients connected to a particular WiFi access points in City Oulu area.

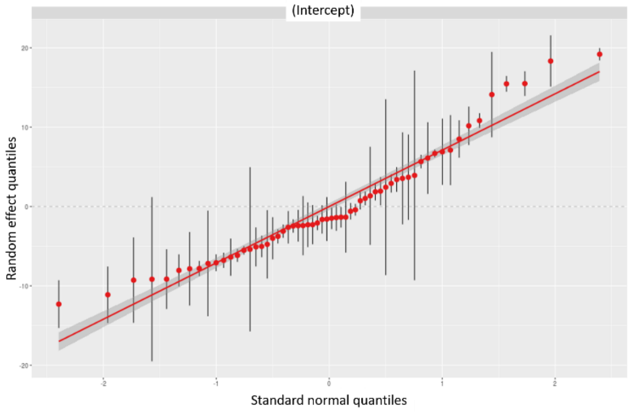

Fig. 7. qq-cel-1 title: Cell intercept regularization QQ-plot..

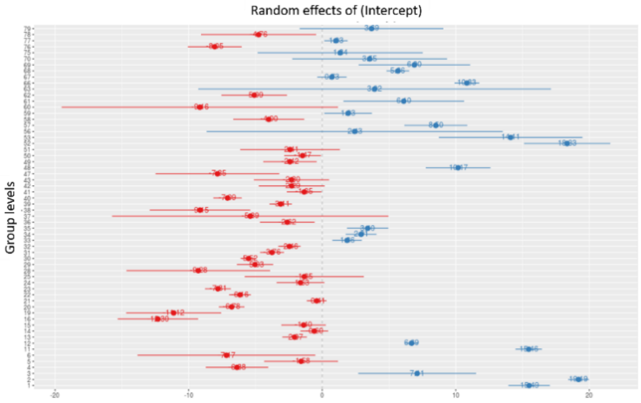

Fig. 8. cel-random-1 title: Cell intercepts with confidence limits.

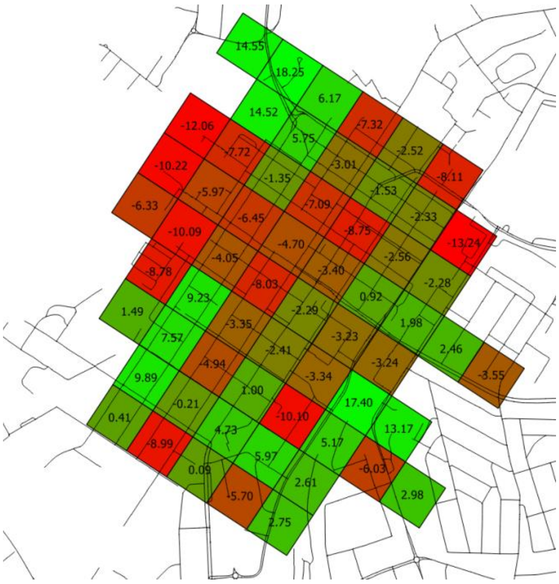

Fig. 9. Cell intercept predictions on map

In Fig. 6 the marked area B belongs to the location detected in the study as a crowded area, explaining the low average speed. Also, number of the pedestrian crossings (Table 5) support the observation that the effect of pedestrian crossings on low speed proportion is not direct, but related to real movements of people, similarly to the observation made from the average speed. Our findings support the need for data fusion from heterogeneous sources for data analysis as well as the need for data pre-processing. The analysis would benefit from crowdsourcing.

VII. CONCLUSIONS

In this paper we explained how traffic-related data needs to be prepared to proceed with map-matching and fetching attribute data. The noise present in sparse vehicle data measurements has to be removed to gather reliable data, as well as to align the data on a digital map. Furthermore, the characteristics of taxi data present problems that we discussed. We use time-based segmentation to divide long trips to sub-trips in order to make data analysis more reliable. By studying Origin-Destination pairs we are able to restrict the analysis on a specific area of interest. Free flow from origin to destination is allowed to get more realistic information about the relations between city traffic and attribute data.

The experiments show that relations can be found by information fusion, between average speed and map attributes. However, the experiments also show that special care should be taken in data analysis. External factors, like real movements of pedestrians may play important part in driving behaviour and consequently relations are not self-evident.

The results of mixed modelling show clearly the difference between the average point speeds in cells. This should provide a good platform for further investigation on what constitutes this effect. Also, in data analysis, accuracy and correctness of the digital map information is important. In our prior work, we have incorporated the preprocessing, map preparation, filtering, map-matching and feature extraction properties to a Driving coach prototype, suggesting post-driving analysis of the trips driven [31]. Map context of the driven routes can be very useful in analysis of driving patterns and may reveal the reasons behind such patterns, consequently providing important information for applications like personalised route recommendation systems. Instructing driver for fuel-efficient driving is of great interest [32]. In future work, more heterogeneous context information will be provided through data suppliers and stored in a common database enabling more thorough analysis of the data dependencies.

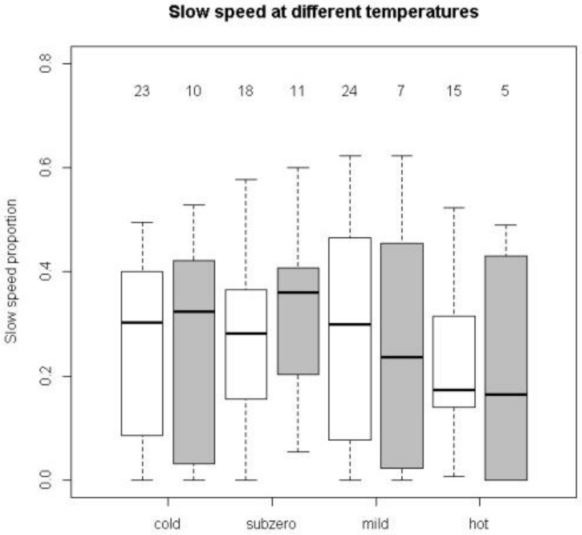

Fig. 10. Low speed % with different temperature classes, number of traffic lights<9 (white) or >=9 (grey). Weather information provided by a road weather model, supplied by FMI (Kangas et al., [30]).

REFERENCES

[1] Y. Zheng, Y. Liu , J. Yuan, X. Xie,"Urban computing with taxicabs," Proceedings of the 13th international conference on ubiquitous computing, pp. 89-98, 2011.

[2] D. Zhu, N. Wang, L. Wu, & Y. Liu, "Street as a big geo-data assembly and analysis unit in urban studies: A case study using Beijing taxi data," Applied Geography, vol. 86, pp. 152-164, 2017. doi:10.1016/j.apgeog.2017.07.001

[3] Y. Zhou, Z. Fang, J.-C. Thill, Q. Li, and Y. Li, "Functionally critical locations in an urban transportation network: Identification and spacetime analysis using taxi trajectories," Computers, Environment and Urban Systems, vol. 52, pp. 34–47, Jul. 2015.

[4] X. Li, D. Tang, X. Xu, "Deriving features of traffic flow around an intersection from of vehicles," International conference on geoinformatics, pp.1-5, 2010.

[5] X. Li, G. Pan, Z. Wu, G. QI, S. Li, D. Zhang, W. Zhang, Z. Wang, "Prediction of urban human mobility using large-scale taxi traces and its applications," Frontiers of computer science, vol. 6, no. 1, pp. 111-121, 2012.

[6] J. Yuang, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, T. Huang, "T-Drive: Driving directions based on taxi trajectories," IEEE transactions on knowledge and data engineering, vol. 25, no. 1, pp. 220-232, 2013.

[7] E. Gilman, S. Tamminen, A. Keskinarkaus, T. Anagnostopoulos, X. Su, S. Pirttikangas, J. Riekki, "Fuel consumption analysis of driven trips with respect to route choice," IEEE 36th Int. Conf. on Data Engineering Workshops (ICDEW), pp.40-47, 2020. https://doi.org/10.1109/icdew49219.2020.000-9

[8] X. Kong, F. Xia, J. Wang, A. Rahim and S. K. Das, "Time-Location-Relationship Combined Service Recommendation Based on Taxi Trajectory Data," IEEE Transactions on Industrial Informatics, vol. 13, no. 3, pp. 1202-1212, June 2017, doi: 10.1109/TII.2017.2684163

[9] J. Tang, F. Gao, F. Liu, W. Zhang, and Y. Qi, "Understanding Spatio-Temporal Characteristics of Urban Travel Demand Based on the Combination of GWR and GLM," Sustainability, vol. 11, no. 19, p. article 5525, 2019. doi: 10.3390/su11195525.

[10] T. Phiboonbanakit, T. Horanont, "Analyzing Bangkok city taxi ride: reforming fares for profit sustainability using big data driven model," J Big Data 8, vol. 7, 2021. https://doi.org/10.1186/s40537-020-00396-5

[11] S. Liu, Y. Liu, L.N. Ni, J. Fan, M. Li, "Towards mobility-based clustering," Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 919-928, 2010.

[12] X. Liu, L. Gong, Y. Gong, and Y. Liu, "Revealing travel patterns and city structure with taxi trip data," Journal of Transport Geography, vol. 43, pp. 78–90, Feb. 2015.

[13] H. Wang, H. Zou, Y. Yue, Q. Li, "Visualizing hot spot analysis result based on mashup," Proceedings of the international workshop on location based social networks, pp. 45-48, 2009.

[14] Q. J. Kong, Q. K. Zhao, C. Wei, and Y. C. Liu, "Efficient traffic state estimation for large-scale urban road networks," IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 1, pp. 398-407, 2013.

[15] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi GPS traces to social and community dynamics: A survey," ACM Computing Surveys, vol. 46, issue 2, article 17, pp. 1-34, 2013.

[16] L. Alvares, V. Bogorny, B. Kuijpers, J.A.F Macedo, B. Moelans, A. Vaisman, "A model for enriching trajectories with semantic geographical information," In proceedings of the 15th annual ACM international symposium on advances in geographical information systems (GIS'07), article 22, pp.1-8, 2007.

[17] Jia Jiang, Hong Li, Rui Xiao, "Error processing on the real-time traffic data," International conference on intelligent system design and engineering application, pp. 680-682, 2010.

[18] Q. Li, Z. Zeng, B. Yang, T. Zhang, "Hierarchical route planning based on taxi GPS-trajectories," 17th international conference on geoinformatics, pp.1-5, 2009.

[19] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, Y. Huan, "Map-matching for low-sampling rate GPS trajectories," Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 352-361, 2009.

[20] Y. Wang, Y. Zhu, Z. He., Y. Yue, Q. Li, "Challenges and opportunities in exploiting large-scale GPS probe data," Tech. rep. HPL-2011-109, HP Laboratories, 2011.

[21] Z. Zhang, D. Yang, T. Zhang, Q. He, X. Lian, "A study on the method for leaning and repairing the probe vehicle data," IEEE Transactions on intelligent transportation systems, vol. 14, no. 1, pp. 419-427, 2013.

[22] "Digiroad –National road and street database," https://vayla.fi/en/transport-network/data/digiroad

[23] https://vayla.fi/vaylista/aineistot/digiroad/aineisto/aineistojulkaisut

[24] C. F. Minett, A. M. Salomons, W. Daamen, B. van Arem, S. Kuijpers, "Eco-routing: Comparing the fuel consumption of different routes between an origin and destination using field test speed profiles and synthetic speed profiles," IEEE Forum on Integrated and sustainable transportation system (FISTS), pp. 32-39, 2011.

[25] S. Brakatsoulas, D. Pfoser, R. Salas, C. Wenk, "On map-matching vehicle tracking data," In Proceedings of the 31st international conference on Very large data bases (VLDB '05), pp. 853-864, 2005.

[26] B. Jœrgensen, "The Theory of Linear Models," Chapman & Hall, London, UK, 1993.

[27] C.E. McCulloch, S.R. Searle, J.M. Neuhaus, "Generalized, Linear and Mixed models," John Wiley & Sons, Hoboken, New Jersey, 2008.

[28] E. Ericsson,"Independent driving pattern factors and their influence on fuel-use and exhaust emission factors," Transportation Research Part D: Transport and Environment, vol. 6, no. 5, pp. 325-345, 2001.

[29] V. Kostakos, T. Ojala, T. Juntunen,"Traffic in the smart city: Exploring the potential of city wide sensing to augment a traffic control center," IEEE Internet Computing, vol. 17, no. 6, pp. 22-29, 2013.

[30] M. Kangas, M. Hippi, J. Ruotsalainen, S. Näsman, R. Ruuhela, A. Venäläinen, M.Heikinheimo,"The FMI Road Weather Model," HIRLAM Newsletter no. 51, October 2006. Available from: http://hirlam.org/index.php?option=com_docman&task=doc_details&gid=47

[31] E. Gilman, A. Keskinarkaus, S. Tamminen, S. Pirttikangas, J. Röning, J. Riekki, "Personalized assistance for fuel-efficient driving," Journal of Transportation Research Part C: Emerging Technologies, vol. 58, part D, pp. 681–705, 2015.

[32] E. Gilman, G.V. Georgiev, P. Tikka, S. Pirttikangas and J. Riekki, "How to support fuel-efficient driving?," IET Intelligent Transport Systems, vol. 12, no. 7, pp. 631-641, 2018.