IDENTIFICATION OF SIMILARITIES AND PREDICTION OF UNKNOWN FEATURES IN AN URBAN STREET NETWORK

U. Feuerhake[1],*, O. Wage[1], M. Sester[1], N. Tempelmeier[2], W. Nejdl[2], E. Demidova[2]

[1] Institute of Cartography and Geoinformatics, Leibniz University Hannover, Germany
- (feuerhake, wage, sester)@ikg.uni-hannover.de
[2] L3S Research Center, Leibniz University Hannover, Germany - (tempelmeier, nejdl, demidova)@L3S.de

Commission IV, WG IV/3

KEY WORDS: urban traffic, traffic analysis, machine learning, clustering, spatio-temporal data, data integration, data mining, floating-car-data

ABSTRACT:

Accurate predictions of the characteristics of urban streets in particular with respect to the typical traffic situations are crucial for numerous real world applications such as navigation, scheduling of logistic and public transportation services as well as high-level planning of infrastructure which may include planning of construction sites or even changes of the road topology. However, this information may be hard to obtain, especially in complex urban road networks where interdependencies between roads are highly present. In addition, accurate and recent traffic data is not always available, especially for uncommon situations like large-scale public events, traffic accidents or construction sites. This work demonstrates how to employ historical traffic datasets in conjunction with other, infrastructure related data, to derive a deeper understanding of urban traffic behaviour. In particular this paper provides the following contributions: (1) the generation of meaningful features to describe the segments in urban road networks; (2) an unsupervised machine learning approach that identifies similar segments based on those features; (3) a supervised approach to predict unknown features of the segments and, finally, (4) an extensive evaluation of the extracted road characteristics and the proposed methods using real-world data. The resulting clusters reveal the similarities of the street segments and give a different perspective on the road network and the traffic situation, respectively. The experiments on the classification approach demonstrate that unknown features can be predicted with a good quality.

1. INTRODUCTION

Predictions of traffic situations in urban environments are crucial for many applications including individual navigation, mobility and logistics services as well as planning of urban infrastructure. To this end, historical traffic data, i.a. floating car data, is typically adopted to facilitate the predictions of re-occurring traffic situations in an urban environment (e.g. temporal fluctuations during rush hour traffic (Ehmke et al., 2010)). However, in practice these methods may not always be feasible. On the one hand, traffic data is typically not publicly available, i.e. difficult or expensive to obtain and can vary greatly with respect to its quality. On the other hand, even in cases where some historical traffic data is available, it does not sufficiently capture traffic conditions for unusual situations such as newly installed construction sites, incidents, or special events at a particular venue, a road or an urban area. In the literature, several aspects of traffic prediction have been addressed, e.g. traffic predictions in presence of public events (Kwoczek et al., 2014) or the evolution of urban congestion (Anwar et al., 2016). However, those approaches strongly depend on the availability of historical traffic data for the particular roads and segments being considered as well as comparable conditions.

Moreover, even if information is available it is not necessarily complete. For instance, the information about the amount or velocity of traffic in either minor or rarely used streets is difficult to obtain and therefore not available or only of poor quality. Additional traffic information, e.g. the fraction of different traffic types (individual mobility, public transportation, logistics, etc.), which is required for a more detailed analysis of the traffic related issues, is also missing.

Thus, the goal of this work is to better understand which features and metrics can be adopted to characterize the streets or infrastructural segments and estimate their similarity so as to enable accurate predictions of the feature characteristics that are not contained in the available datasets or do not have a sufficient quality. In particular, we assume that the similarity of street segments can allow the prediction of unknown characteristics of the streets. The contributions of this paper include:

- the definition of features that characterize street or infrastructure segments in an urban traffic network,
- the definition of similarity metrics to assess similarities across urban infrastructure segments using these features,
- the identification of similar areas and venues in an urban environment using a clustering approach while adopting the proposed similarity metric,
- the prediction of unknown features by applying a supervised learning method to the data and, finally,
- the evaluation of the proposed methods based on various experiments using real-world data and the discussion of the resulting clusters and the possibility to predict unknown features.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-185-2018 | © Authors 2018. CC BY 4.0 License.

The remainder of this paper is structured as follows: first, in Section 2 an overview over the most important related work is provided. Then, in Section 3 we describe the datasets adopted. Following that, in Section 4 the developed algorithms to identification of similar street segments are presented. In a subsequent evaluation presented in Section 5 the proposed approach is applied to the described real-world datasets. There, the influence of the calculated features and the similarity metrics on the quality of the resulting predictions is analyzed. The paper concludes with a summary and an outlook in Section 6, in which the portability of our approach to further scenarios, e.g. the prediction of the traffic load at the different times of a day or different days of the week, is also discussed.

2. RELATED WORK

Recently a number of studies addressed several prediction tasks at the interface of the urban infrastructure and mobility, e.g. the prediction of travel times and traffic characteristics in the context of the individual transportation. TomTom (2009) employs speed profiles for travel time predictions in routing applications. The speed profiles are extracted by performing a cluster analysis which is followed by several quality and alignment steps. The profiles are then used to build time-dependent speed maps that enable time aware routing algorithms. Lécué et al. (2014) employ semantic technologies to develop STAR-CITY, a system for traffic prediction and reasoning, used for spatio-temporal analysis of the traffic status as well as for the exploration of contextual information such as nearby events.

Another line of research aims to analyze urban network infrastructure using mobility data. Wang and Li (2017) leverage taxi flow data to learn vector representation of city regions. The representations are then used to make predictions about the regions such, e.g. crime rate, average income or average house prices.

Pan et al. (2013) make use of taxi GPS-trajectories to classify the land use of urban areas. They propose an iterative DBSCAN algorithm to cluster regions with respect to the frequency with which passengers are picked up or set down. They make use of the same information to classify the land use of regions, e.g. the land use for hospitals or commercial districts. The analysis of trajectories in the traffic context has been conducted with respect to movement behaviour (Sester et al., 2012) or travel mode (Bohte and Maat, 2009). This is done with different foci, e.g. for usage in travel planning and the identification of travel mode, e.g. for later usage in travel planning (Zhang et al., 2013). Trajectories can also be analyzed concerning anomalies using Bayesian networks Huang et al. (2014). Further approaches utilizes location based social network (LBSN) data. Song et al. (2017) leverage LBSN data to identify functional urban regions. They employ latent Dirichlet allocation and unsupervised machine learning algorithms to determine the regions. Yin et al. (2017) make use of LBSN data to infer boundaries of functional regions in urban environments. They construct a mobility network from spatial user interaction and delineate boundaries by identifying strongly connected communities within the network space.

Furthermore, several approaches target the identification of problematic segments and areas of urban networks under specific conditions (e.g. planed special events). Kwoczek et al. (2015) propose the use of an artificial neuronal network to identify road segments that are typically affected by public special events that take place in a particular venue. Moreover, Rodrigues et al. (2017) investigate the affect of public events on the public transportation network. They propose a Bayesian additive model that can be employed to gain an understanding of public transportation demand in the presence of events. I.e. the model is able to predict the number of public transportation trips to the venues where the respective event takes place.

Urban road networks have been subject to several studies aiming at identifying problematic areas and inter-dependencies. Anwar et al. (2016) proposed a method to keep track of the congestions in urban road networks to identify unstable road segments. Jin et al. (2016) make use of context-aware tensor decomposition to identify so called *urban black holes*, i.e. traffic anomalies with a greater inflow than outflow. Jing et al. (2018) analyzed the correlation of node degrees within a road network. They find that existing measures are inconsistent when the road representation is changed and propose an own road network ratio.

Whereas the above mentioned approaches only focus on a specific prediction problems, in this work we also aim to analyze general characteristics of the urban street segments and in particular features that can contribute to the identification of their similarities and transfer of prediction results.

3. DATASETS

This work is based on data which is collected in the urban area of Hanover, Germany, and is taken from different sources. On the one hand, open geo-spatial data, the street network and additional features from OpenStreetMap[1] (OSM), as well as supplemental internal authoritative LHH[2] datasets (including inhabitants and buildings distribution and additional street attributes) are used. On the other hand, a proprietary traffic dataset consisting of aggregated floating car data (FCD) is employed. In the following paragraphs the different datasets are described in more detail.

OSM is the most common platform for open geo data and provides a comprehensive amount of spatial objects and additional attributes. Despite general rules for tagging (attaching attributes to objects), due to the data collection and maintaining by volunteers, possible inhomogeneity, incompleteness and even faultiness of the given information has to be considered. Features like traffic signals (point object tag *highway=traffic_signals*) can be easily extracted for further processing. The further mentioned aggregated FCD is referenced to OSM street segments.

The **LHH** dataset provided by the city administration of Hanover is also used by this approach. Originally it is collected for administrative and traffic management purposes and by this it is not adapted to our purposes. One part of it consists of the location of all buildings and the number of their residents. The second provided dataset consists of a street network, which differs geometrically from OSM; in particular, it includes only one edge for bi-directional streets. The segments are enriched by different attributes like the speed limit or the number of tracks. Other attributes, such as the average daily traffic on each segment are the result of a simulation and thus are available area-covering.

The **FCD dataset** provides traffic speed records for each street segment among the most important OSM road types (from tertiary to motorway level). Note, that due to this limitation of road types small roads which are typically found in residential areas are excluded. In fact, the dataset only contains major roads of

[1] https://www.openstreetmap.org

[2] [ger.] Landeshauptstadt Hannover - [engl.] state capital Hanover

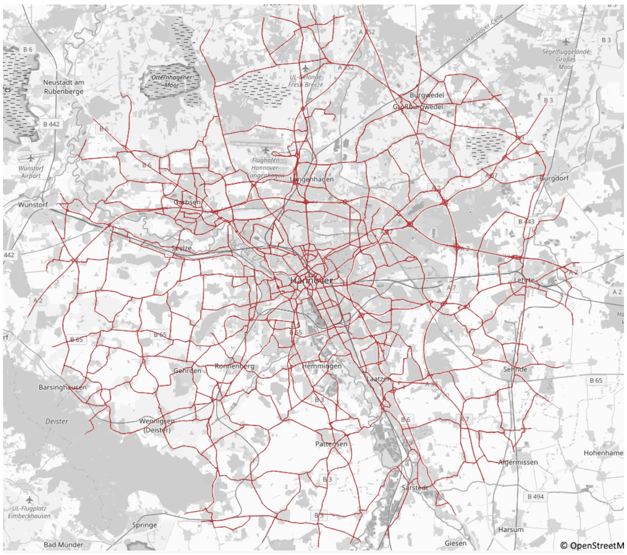

Figure 1. An overview of the considered road segments of the FCD set. [source: Stamen map tiles and OSM]

Hanover city and we are not able to sense traffic flowing into residential blocks. Figure 1 provides an overview of the road segments contained in the dataset. The records contain information regarding the measured traffic speed on the individual street segment of the street network at discrete time points i.e. the speed is recorded every 15 minutes. Note that this kind of information can easily be extracted from more commonly available floating car data. The dataset covers the time span from October 2017 to January 2018 and all street segments of the aforementioned category that are located within a distance of 20 km from the centre of the city of Hanover, Germany. The dataset contains approximately 195 million records in total. Note that given the scale and costliness of real-world traffic data, our experiments are limited to the reported duration of four months. According to our experience this duration is sufficient to capture the typical traffic patterns.

4. APPROACH

Claiming the assumption that similar street segments also have similar traffic feature characteristics, the overall approach of this work is to use machine learning algorithms to either identify similar street segments based on their features or to predict features which are not available on all segments. In order to be able to apply the corresponding unsupervised (section 4.2) and supervised methods (4.3), features describing the segments sufficiently are generated in a preprocessing phase (4.1).

4.1 Preprocessing of the Street Segments and Calculation of Features

Street Segments. The original OSM street segments used in the FCD dataset strongly vary with respect to their length, i.e. the smallest segments are shorter than 0.5 meters. Moreover, this dataset contains a large number of these segments, and thus is difficult to handle. E.g. the region described in Figure 1 contains over 23,000 road segments. To allow an efficient use of the dataset, in the experimental settings the total number of segments needs to be reduced, e.g. by merging multiple short segments to form longer segments. To this end, short street segments are merged if they satisfy the following conditions: (1) Merged segments have the same type according to the OSM taxonomy. (2) Merged segments are directly connected. (3) Merged segments are not separated by crossings or intersections. (4) Merged segments have the same direction, i.e. the segments are not allowed to form circles. (5) The summed length of the merged segments must not exceed 500 meters. Following this procedure the total number of street segments in the dataset was reduced from approx. 23,000 to approx. 10,000. Finally, the measured FCD speed values of the merged segments are averaged. The resulting street graph will be used as reference in this work and additional features will be assigned to it.

FCD-based Features. Based on the speed information per segment for the complete time span of this dataset we determine typical **profile types**, which describe the development of the traffic load over a week. The calculation of the traffic load TL is done by combining the actually driven speeds v and the official speed limits of the segments v_{limit}.

$$TL = 1 - \frac{v}{v_{\text{limit}}} \tag{1}$$

Please note that negative values for the traffic load mean that people drive faster than allowed. This can more frequently be observed at night when less people are on the way (see Figure 2). Using the averaged week values per time step over the whole time span of the dataset we cluster the averaged and normalized profiles of the segments to extract profile prototypes. For this purpose, we use the k-means algorithm MacQueen (1967) with the parameter $k = 4$. This value has been determined using the Davies-Bouldin-Index Davies and Bouldin (1979), which is a cluster evaluation metric that indicates how well the clusters are separated from each other. The resulting prototypes for the traffic load profiles are shown in Figure 2.

Prototype 0 (blue) is characterized by the largest changes of amplitude in a day including a prominent PM-peak on weekdays. Obviously, the opposite can be observed for prototype 3 (yellow), as is has more prominent AM-peak. There, Prototype 1 (orange) represents segments with only a minimal change in daily traffic around rush hours. Finally, type 2 (grey) also indicates a larger PM-peak than in the morning, but in total less strong and with flatter slopes than the previous two. Except type 1, the weekend peaks seem to merge to a wide raising, but less strong than formally. To summarize, street segments belonging to type 3 seem to be used mainly by people going to work in the morning, whereas segments belonging to type 0 are used on the way back home. Please note that profile types 1 and 2 are nearly about four times more frequent in number (1600 and 1531 segments) and total length (192 and 205 km) than types 0 (443 segments and 59 km) and 3 (406 segments and 49 km).

In addition to the profile types, we also extract the **time of minimal flow (total, AM, PM)** for each segment based on an averaged week. The minimum for each day of the week is determined, and further split into before- and afternoon. Because many segment’s flow does not varies much over day, a significance threshold of 15% over the daily mean is implemented. Figure 3 illustrates the situation for a work day. The highlighted places show an interesting traffic behaviour. In A the times of the peaks suddenly change on the junction. This is caused by the people coming from the outside going to city center (on the left) in the morning (the green color indicates an early time of the day) and leaving in the evening (the orange color indicates a later time of the day) in the opposite direction. Thus, this place can be seen as sink where incoming streams meet. In place B the situation is similar as there

https://doi.org/10.5194/isprs-archives-XLII-4-185-2018 | © Authors 2018. CC BY 4.0 License.

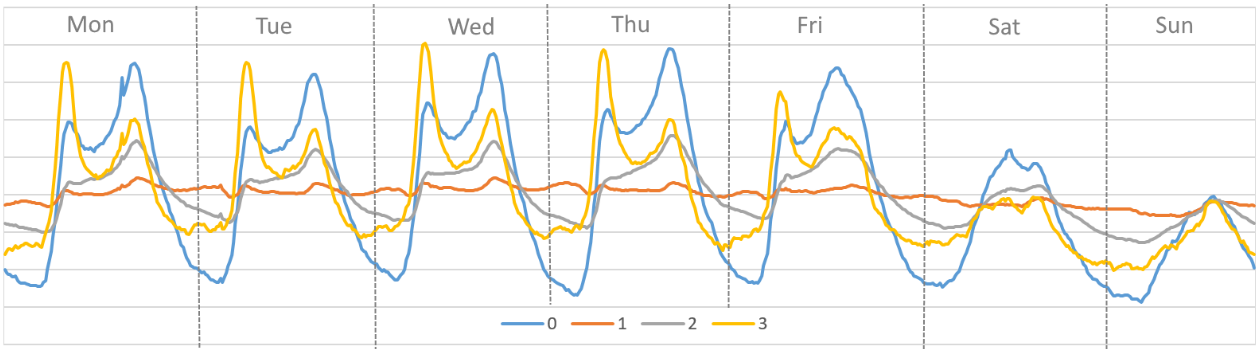

Figure 2. The resulting traffic load profile prototypes' mean week starting on Monday. They differ with regard to their amplitude, their shape and the number of their peaks.

are peaks at both direction of the street, which end at same location. This means that many people enter and leave this street at a the same times in the morning and evening. This is good indicator that there must be a point of interest (POI) for those people. In this case, there is a children hospital, which attracts a lot of people. The *time of minimal flow* feature can thus also be used to identify POIs.

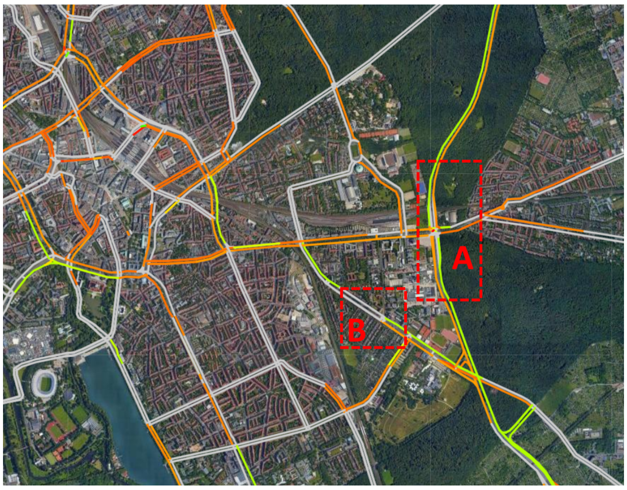

Figure 3. The times of the daily traffic load peak per segment indicate when the traffic load is at its maximum, the velocity is minimal, respectively. The spectral color scheme encodes the time of the day. Green means early (usually in the morning), red late (afternoon/evening). Gray segments do not have a significant peak at all. The marked places A and B show two situations of interesting behaviour. Map source: GoogleEarth

OSM-based Features. From OSM traffic signals are extracted. They are assigned to nearby segments. To prevent double assignments from different directions, signals were clustered spatially. The feature **#traffic lights** value contains the number of intersecting clusters.

Based on the network graph, two features were calculated to take its topology and geometry into consideration. First, the **centrality** is calculated using the Floyd-Warshall algorithm (Floyd, 1962). The edge weights are based on a optimistic travel time (speed limit times length). The resulting feature consists of the summed number of 'visits' on each edge and represents the topological importance in the street network. The second feature, called **direction**, gives the relative direction in relation to the city centre (Kröpcke square). By this measure each street segment can be discerned as outgoing, ingoing or in radial direction.

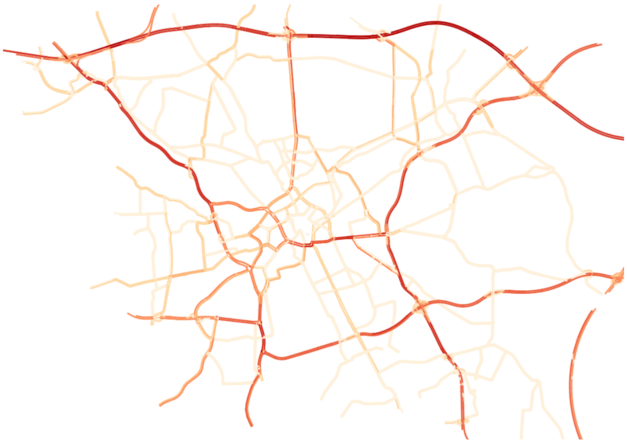

Figure 4. The centrality features shows how often a segment is used in fastest paths (required time considering the length and the speed limit). The intensity of red symbolizes the number of usages.

LHH-based Features. To also use information provided by buildings in the street segment's vicinity, the number of inhabited (**#build.inh.**), not inhabited buildings (**#build.n.inh.**) and the sum of inhabitants (**#inhabitants**) is aggregated in the cells framed by the street network. Street links and small interspaces are ignored. The street segments get the total of neighbouring cell values, shown in Figure 5. By this, a kind of commuting area of the segments is modeled. The differentiation of buildings by inhabitants is used to include rough information about industrial and commercial areas, which is not consistently available in OSM.

The LHH-features are linked to the street graph provided by the city of Hannover. In order to use it in our target system - the OSM graph - they have to be transferred. However, they are geometrically different. Figure 6 shows the area of the inner city ring as an example. Different levels of granularity (street type) are evident. The main roads of it are included in both graphs (and two directions in OSM). By this, the administrative features need to be assigned by a matching to the OSM graph. For each OSM edge one of the city graph in the surrounding is selected by searching for smallest sum of distance between both combinations of start and end point. Transferred features attributes by this matching are the average daily amount of traffic (**DTV**[3]), number of lanes (**#lanes**) and type of (**land use**).

[3] [ger.] Durchschnittlicher Tagesverkehr

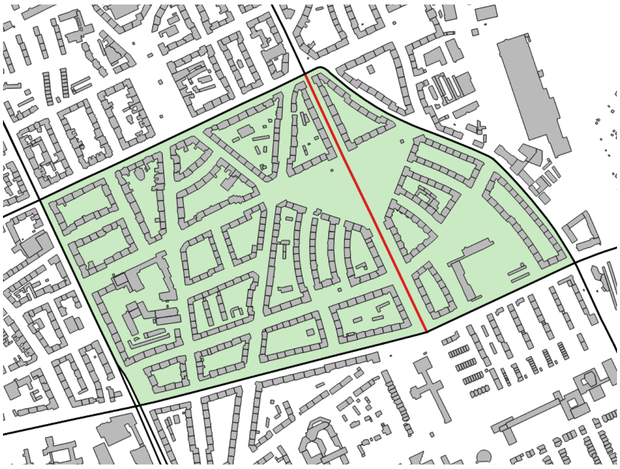

Figure 5. Number of buildings and inhabitants are aggregated on cells, which are formed by the street network (black). A street segment (red) receives from adjacent cells (green) their sum.

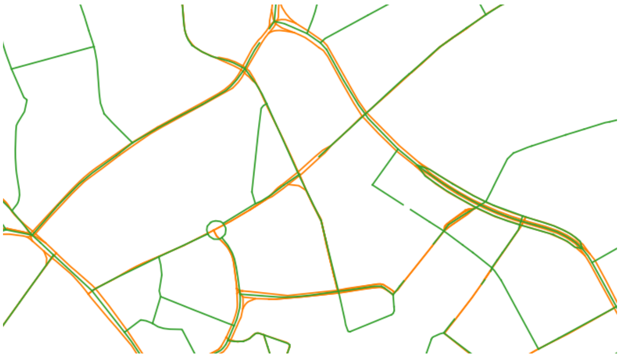

Figure 6. Geometrical differences of OSM (orange) and administration (green) street graph. In the snipped is shown the area of inner city ring road. The administration graph includes hierarchical smaller streets, but only one edge per street.

4.2 Identification of Similar Street Segments

Since the similarity of the street segments depends on the evaluated features, we generate feature vectors FV for the street segments s_i containing the precalculated features of section 4.1.

$$FV(s_i) = \begin{pmatrix} \text{profile type} \\ \text{time min flow (total)} \\ \text{time min flow (AM)} \\ \text{time min flow (PM)} \\ \text{centrality} \\ \text{direction} \\ \text{\#traffic lights} \\ \text{\#build. inh.} \\ \text{\#build. n.inh.} \\ \text{\#inhabitants} \\ \text{\#lanes} \\ \text{DTV} \\ \text{land use} \end{pmatrix} \quad (2)$$

If a feature should be estimated, it must not be included in this vector. For instance, if we want to predict the traffic profile types of road segments, the corresponding feature vectors for the clustering obviously must not contain the *profile type* and the three derived *time min flow* features. The vector from Equation (2) will be reduced by the feature *profile type*. One might miss the road type attribute from OSM as a strong feature for similarity. Please be aware that we neglected it as we determined it a too strong a-priori expert knowledge. Further, it describes one possible classification of the segments which can be used later to evaluate our cluster results.

To obtain those clusters of similar segments we apply the *k-means algorithm* to the data using the described feature vectors in combination with the Euclidean distance metric. As the latter is not capable of dealing with nominal features, e.g. the profile type and land use, we have to transform them into numerical features using *one-hot-encoding* Harris and Harris (2012). To determine the most suitable clustering parameter the same mechanism based on the Davies-Bouldin-Index as described in section 4.1 is used.

4.3 Prediction of Street Segment Features

In contrast to the unsupervised approach of the previous section the second analysis in this work is based on a supervised learning and aims on predicting unknown features of the street segments. To this end, besides the previously described preprocessing phase, in which mainly the features are prepared, a second preparation of the dataset for the training of the classification model is required. It consists of a rebalancing of the dataset to have the same number of samples for all classes. In addition to that, numerical features are transformed to nominal ones to enable algorithms which are only capable to handle nominal features. After having prepared the dataset the *random forest algorithm* Breiman (2001) is used as classification method to predict the unknown feature of the street segments. Excluding the feature which should be predicted and its derived features, the feature vector is the same as given in Equation (2).

5. EXPERIMENTS AND DISCUSSION

In the following section the results of both, the similarity analysis and the feature prediction, are evaluated and discussed. For this purpose experiments based on the described datasets are performed.

5.1 Identification of Similar Street Segments

The experiment based on the application of the k-means method to identify similar segments in the road network provides the resulting clusters illustrated in Figure 7. The required clustering parameter is set to $k = 10$, which is the number of resulting clusters. It has been determined by the described procedure using the Davies-Boulding-Index.

In order to evaluate the resulting clusters, we first inspect them visually. It turns out that in most cases the differently directed lanes of the same street are assigned to the same cluster. However, looking at the bigger streets, e.g. motorways and trunks, the different directions fall into different clusters (see example regions *A* in Figure 7, left). This can be explained by the different peaks in the daily traffic behavior mainly caused by the people’s way to work. For instance, in the morning many people drive into the city, in the evening they use the opposite direction. Further, multiple adjacent segments along a street are relatively uniformly clustered. Only in junction areas segments are sometimes assigned to different clusters (see regions *B*). A reason for this are the very different traffic profiles of the street segments. While on non-junction segments there is often moving traffic with relatively high velocities, on junctions there regularly are low velocities caused by cautiously turning cars.

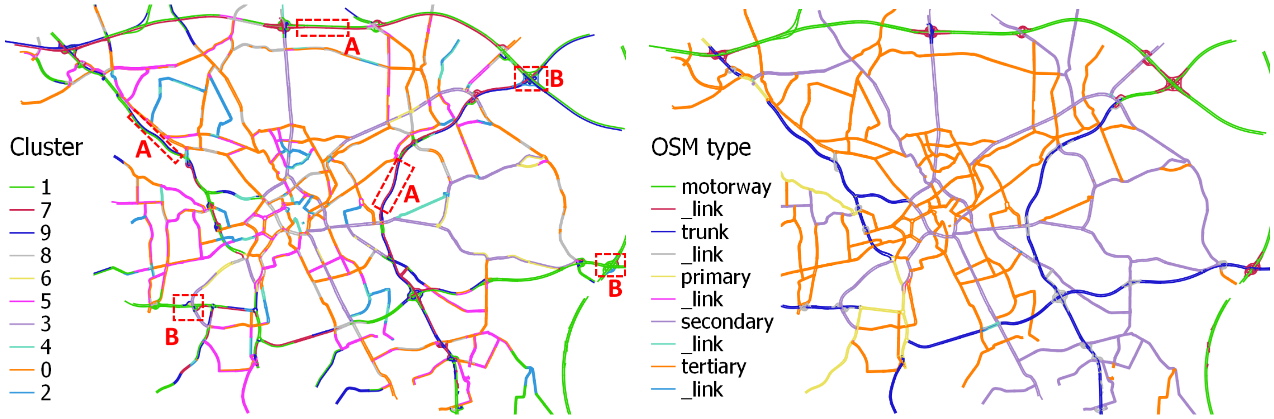

Figure 7. The resulting clusters of similar street segments (left) and OSM road types (right) for visual comparison.

Comparing our solution to the original OSM road types (see Figure 7, right) a general correspondence between both can be recognized, after a rough resorting of the color assignment. The calculated overlap between both is about 42%. The top three clusters 1, 7 and 9 (light-green, red, dark-blue) are basically motorways, trunks and partly complemented by primary streets. Secondary roads are especially inner city main axes, which are covered by cluster 3 (pale-purple). Tertiary streets seem to be split into clusters 0 and 2 (orange and light-blue).

However, there are some deviations in this assignment, which can be observed, when, for example, segments of minor OSM road types are assigned to the motorway cluster, too. The reason for this is the way the segments are classified. The OSM road types are based on traffic relevance, development state, administrative assignments and are further tried to keep constant for one road, whereas our clusters have been determined using the generated features for each segment individually and thus are showing the actual traffic behavior and urban surrounding. Resulting, our results look less homogeneous than the OSM types (see Figure 7).

5.2 Prediction of Segment Features

In several experiments the performance of the prediction is exemplarily evaluated based on the nominal features *profile type*, *land use* and *#inhabitants* (ten discrete intervals). For the evaluation the classification task is interpreted as multi-class and binary-class problem, respectively. The latter is done to find out whether certain feature characteristics are better predictable than others. The different classes are balanced for each scenario. The resulting performance is shown by the precision and recall indicators in Table 1.

The prediction of the *land use* feature, performs very well, even as multi-class problem (precision and recall values of about 95%). The classification of the data with respect to the *profile type* feature is in the range 80% for the binary scenarios. The precision and recall values for the multi-class classification are only 65 and 64%, respectively. Even if the values are low, they are considerably higher than a random guess, which would be in the range of 25

The corresponding values for the multi-class classification of the *#inhabitants* feature are in both cases at 98%, using all available features. Ignoring the features *#build. inh.* and *#build. n.inh.*, as they are certainly related to *#inhabitants*, the values decrease to 71% in both cases.

Predicted feature	Precision	Recall
profile type {0, 1, 2, 3}		
Multi-class	0.65	0.64
0 vs. all	0.79	0.74
1 vs. all	0.78	0.78
2 vs. all	0.72	0.72
3 vs. all	0.83	0.80
land use {commercial, residential, radial, none}		
Multi-class	0.95	0.95
commercial vs. all	0.94	0.94
residential vs. all	0.97	0.97
radial vs. all	0.97	0.97
none vs. all	0.94	0.93
#inhabitants (10 discrete intervals)		
Multi-class	0.98	0.98
Multi-class (without *#build. inh.* and *#build. n.inh.* features)	0.71	0.71

Table 1. The evaluation of the segment feature prediction is performed by a multi-class- and one-against-all-classification.

Feature	Information gain [bit]		
	profile type	*land use*	*#inh.*
profile type	n.a.	0.05	0.03
time min flow (tot.)	n.a.	0.04	0.03
time min flow (AM)	n.a.	0.08	0.02
time min flow (PM)	n.a.	0.02	0
centrality	0.04	0.12	**0.22**
direction	0.03	0.04	0
#traffic lights	0.04	0.11	0.03
#build. inh.	**0.05**	**0.16**	**2.18**
#build. n.inh.	**0.07**	0.15	**2.12**
#inhabitants	0.04	**0.16**	n.a.
#lanes	0.03	**0.17**	0.07
DTV	**0.05**	**0.18**	**0.55**
land use	**0.05**	n.a.	0.08

Table 2. The information gain of the features in relation to the corresponding classification task. Not evaluated features are marked by ’n.a.’.

In order to explain the varying prediction qualities, we analyze the influence of the evaluated features in relation to the corresponding classification task. To this end we calculate the information gain of each feature to describe its relevance. The results are shown in Table 2. For each classification task the most relevant features are highlighted. In case of the *profile type*-classification the influence of the *#build. n.inh.* feature can be explained by the fact that it indicates typical destination hot spots for the daily way to work of many people. Those areas favor the profile types showing the typical rush hour behaviour. Resulting from the visualization of the profile types (Figure 2), work traffic seems to be the most characteristic component of traffic load for type 0 (blue) and 3 (yellow). In this way, a dependency between the profile types and those hot spots can be concluded. Comparing the information gain values between the different classifications scenarios, relatively low values for the *profile type*-classification can be noticed. This is a good indicator for the absence of a really significant feature, which would improve the overall prediction quality. The same relationship can be observed for the *#inhabitants* feature. If the two dominant and, of course, somehow related *#build. inh.* and *#build. n.inh.* features are ignored, the precision and recall will decrease to only 71%.

6. CONCLUSION AND OUTLOOK

In this paper we addressed the problem of identifying urban street segment similarities and predicting segment features by applying machine learning algorithms on real-world data. We presented an unsupervised approach, in particular the k-means algorithm, that clusters the segments based on pre-calculated meaningful features describing the segments with respect to the associated actual traffic behavior. Further, a supervised learning method, i.e. the random forest algorithm, was used to predict unknown features. In this case, we exemplarily showed the prediction of the *profile type*, *land use* and *#inhabitants* by using the remaining features.

The results of the clustering reveals the similarity between street segments. After a comparison to OSM they do not fully match to the original road types. However, in contrast to the OSM road classification, which mainly results from relevance, our clusters also include the actual traffic behaviour. In further experiments, related to the supervised learning, we obtained different qualities for the prediction of the *profile type* and *land use* feature. Whereas the latter worked well, the profile types could be improved, especially if the learning task is interpreted as multi-class problem (all against all).

Besides those findings there are several open issues, which can be addressed in future work. An issue is still the prediction of the traffic load profiles. There, additional features or different learning techniques, e.g. methods from the deep learning field, could further improve the prediction quality. Also, it has to be investigated, which applications need which quality traffic load profiles. Nevertheless, this work can be considered as a basis for further analyses. The prediction of the traffic load at different times of a day or different days of the week is one of them as it provides information, which, for example, will be useful for navigation purposes, in particular, if no FCD is available. Moreover, the calculated features can also be used for other purposes. For instance, the *time of min flow* feature can be used in the context of detecting POIs automatically.

ACKNOWLEDGEMENTS

This work is partially funded by the research initiatives "Mobiler Mensch" and "Urbane Logistik Hannover" as well as the associated BMBF projects "Data4UrbanMobility" (grant ID 02K15-A040) and "USEfUL" (grant ID 03SF0547).

REFERENCES

Anwar, T., Liu, C., Vu, H. L. and Islam, M. S., 2016. Tracking the evolution of congestion in dynamic urban road networks. In: *CIKM'16*.

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

Breiman, L., 2001. Random Forests. *Mach. Learn.* 45(1), pp. 5–32.

Davies, D. L. and Bouldin, D. W., 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1(2), pp. 224–227.

Ehmke, J. F., Meisel, S. and Mattfeld, D. C., 2010. *Floating Car Data Based Analysis of Urban Travel Times for the Provision of Traffic Quality*. Springer New York, New York, NY, pp. 129–149.

Floyd, R. W., 1962. Algorithm 97: Shortest path. *Commun. ACM* 5(6), pp. 345.

Harris, D. and Harris, S., 2012. *Digital Design and Computer Architecture, Second Edition*. 2nd edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Huang, H., Zhang, L. and Sester, M., 2014. A recursive bayesian filter for anomalous behavior detection in trajectory data. In: *Connecting a Digital Europe Through Location and Place*, Springer, pp. 91–104.

Jin, L., Feng, Z. and Feng, L., 2016. A context-aware collaborative filtering approach for urban black holes detection. In: *Proc. of the CIKM 2016*.

Jing, T., Huaqiang, F., Yiheng, W. and Chang, R., 2018. On the degree correlation of urban road networks. *Transactions in GIS* 22(1), pp. 119–148.

Kwoczek, S., Martino, S. D. and Nejdl, W., 2014. Predicting and visualizing traffic congestion in the presence of planned special events. *J. Vis. Lang. Comput.* 25(6), pp. 973–980.

Kwoczek, S., Martino, S. D. and Nejdl, W., 2015. Stuck around the stadium? an approach to identify road segments affected by planned special events. In: *IEEE ITSC 2015*.

Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M. L. and Tommasi, P., 2014. Star-city: Semantic traffic analytics and reasoning for city. In: *Proc. of IUI '14*.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA., pp. 281–297.

Pan, G., Qi, G., Wu, Z., Zhang, D. and Li, S., 2013. Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems* 14(1), pp. 113–123.

Rodrigues, F., Borysov, S. S., Ribeiro, B. and Pereira, F. C., 2017. A bayesian additive model for understanding public transport usage in special events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sester, M., Feuerhake, U., Kuntzsch, C. and Zhang, L., 2012. Revealing underlying structure and behaviour from movement data. *KI-Künstliche Intelligenz* 26(3), pp. 223–231.

Song, G., Krzysztof, J. and Helen, C., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21(3), pp. 446–467.

TomTom, 2009. White paper how TomTom's HD Traffic^{TM} and IQ Routes^{TM} data provides the very best routing travel time measurements using GSM and GPS probe data.

Wang, H. and Li, Z., 2017. Region representation learning via mobility flow. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, ACM, New York, NY, USA, pp. 237–246.

Yin, J., Soliman, A., Yin, D. and Wang, S., 2017. Depicting urban boundaries from a mobility network of spatial interactions: a case study of great britain with geo-located twitter data. *International Journal of Geographical Information Science* 31(7), pp. 1293–1313.

Zhang, L., Dalyot, S. and Sester, M., 2013. Travel-mode classification for optimizing vehicular travel route planning. In: *Progress in Location-Based Services*, Springer, pp. 277–295.