Novel Road Classifications for Large Scale Traffic Networks

Werner Toplak, Hannes Koller, Melitta Dragaschnig, Dietmar Bauer, Johannes Asamer

Abstract—Establishing a highly sophisticated large-scale Traffic Information System (TIS) requires the creation and deployment of link travel time prediction models for large road networks. Due to the dimension of typical road networks and low coverage with Floating Cars (FC), data sets that can be used for prediction contain a large number of missing observations. Additionally, specifying prediction models for each link separately is impossible due to restrictions of both computational as well as modeling resources. This paper aims to improve the scalability of link travel time predictions by combining information from roads with similar characteristics. The Functional Road Class (FRC) is a widely accepted indicator for road similarity mainly based on static information from infrastructure planning. The coherence between the clustering introduced by the FRC and road dynamics measured by Floating Car Data (FCD) in the city of Vienna is discussed and analyzed. Clustering approaches that are based on indices characterizing speed measurement distributions are proposed as alternatives to the FRC system. It is demonstrated by way of examples that the new clustering is much more appropriate to provide predictions of link travel times.

I. INTRODUCTION

With the advent of large scale Intelligent Transportation Systems (ITS) and Traffic Information Systems (TIS) reliable real-time travel information is starting to become available for travellers and logistic companies. In order to provide relevant information to passenger and cargo transportation systems the estimation and prediction of future traffic states - in particular predicted travel times along a particular route - is of vital importance. Typically route travel time predictions are constituted by the summation of the predicted travel times along the links contained in the route.

A variety of prediction methods for link travel times have been proposed in the literature. In the recent survey [1] several approaches for short-term prediction have been distinguished: Naive methods are either based on the assumption that traffic will remain constant or use historical averages to calculate predictions. Although these methods generally show comparably poor performance they are often used in practice due to their low computational effort and ease of implementation. Moreover, these approaches constitute the only choice in situations where too many recent observations are missing.

Parametric models (in particular traffic simulations) use predetermined model structures to represent knowledge of traffic processes. The model parameters are usually found in a calibration phase during which an adaptation according to measured data is performed. Compared to naive methods these approaches typically demand data sets that do not contain missing observations. E.g. time series methods such as AR models require a number of consecutive time instants with observations. Hence on links with a large amount of missing data these methods are no option.

In contrast to parametric models in nonparametric models the model structure is flexible and not fixed in advance: model structure and model parameters are determined from the data. The advantage of these models is that the more flexible model structures adapt to the data set and hence reduce the need to impose prior knowledge of the underlying processes [1]. However, this comes at the cost of even higher data requirements compared to parametric methods. Moreover, most work on nonparametric models focuses on the prediction of a single location, link or route, the approach has rarely been applied network-wide [1]. The main obstacles in this respect again are a lack of significant amounts of available data for most links and the inherent computational complexity of creating prediction models for thousands of individual locations, links or routes.

Thus a number of different prediction models are available - each model with certain advantages depending on the data to be mapped, but none of them has up to date been established as a general best solution [1]. Ideally, several of these approaches can be combined to a single prediction model, preserving or even increasing model accuracy while at the same time increasing the requirements on computational resources. Prediction models for link travel times on a whole road network need to be based on large data sets. During the project FLEET (Fleet Logistics Service Enhancement with EGNOS and Galileo Satellite) such a data measurement system has been installed in Vienna, Austria, in 2004 which since then provides real-time link travel times based on the continuous position measurements of currently approx. 3,500 taxis. The graph representation of the urban road network which is covered by FLEET contains about 68,100 edges (about 4,945km of roads). Based on the measurement of the positions of the taxis current link
travel times are estimated at an aggregation level of fifteen minutes (see [2] for details). Hereby the raw measurements are projected onto the road network using map matching methods (again see [2] for details). Speed measurements are obtained by differencing consecutive projected positions. Outlier removal is performed by rejecting unrealistically high speed values (higher than 120km/h). Subsequently speed measurements are aggregated to fifteen minute time intervals. This results in a data set that contains link travel time information only for a subset of links even for the comparably large number of floating cars and extended observation periods.

In order to obtain predictions for each link in any given fifteen minutes time interval it is necessary to combine the measurements of similar links by way of clustering in order to specify one prediction model for each cluster. This drastically reduces the number of required models and simultaneously increases the amount of data available for prediction model training. Obviously the quality of the resulting models strongly depends on the internal cohesion of the clusters found. Traditionally link similarity is often determined by parameters such as spatial distance and road classifications like FRC [3]. In this paper we propose and apply a clustering technique that specifies similarity of links via indicators characterizing the speed distribution. It will be demonstrated by example that this procedure leads to superior link travel time predictions.

The paper is organized as follows: The next section (II) provides detailed insights into link characteristics of the street graph of Vienna. Subsequently a set of indicators suited to describe characteristics of speed distributions which is used in order to define link similarity from Floating Car Data (FCD) is proposed (section III).

In section IV a Self-Organizing-Map (SOM) is used in order to achieve a reduction of statistical patterns of speed distributions on each link. On the corresponding reduced number of patterns various clustering algorithms are applied including different ways for autonomous estimation of optimal cluster numbers. Results will be benchmarked against a hierarchical clustering of the unreduced patterns which have also been used to train the SOM.

A brief demonstration of the findings in travel time prediction is given in section V. Finally the outlook in the last section focuses on the use of the obtained results for prediction purposes in large scale traffic networks (section VI). The limitations of today’s approaches and in particular the opportunities arising from better link clustering methods are discussed.

I. THE ROAD NETWORK

The road network is represented in a graph consisting of nodes and links. The description of the links is specified in the conceptual and logical data model ISO 14825:2004 (ISO GDF 4.0) [3]. This norm also determines the exchange format for geographic data bases for ITS applications. Major input on this ISO standard was provided by the Geographic Data Files (GDF) standard (CEN GDF 3.0) [4]. GDF 3.0 provides rules on how data is captured as well as how the features, attributes and relations of a street graph are defined.

According to [4] GDF is primarily used for car navigation systems but has also found wide acceptance in other transport and traffic applications such as fleet management, traffic analysis, traffic management and automatic vehicle location.

One of the road attributes proposed by [4] is the Functional Road Class (FRC), described as a classification based on the importance of the role that a road element performs in the connectivity of the total road network.

FRC is considered an appropriate classification scheme for use in simulation studies supporting for example traffic management [5] or routing decisions in navigation systems. As such the grouping in FRC is assumed to be appropriate to predict link travel times by grouping links with similar speed characteristics. It provides a convenient approach to road classification since it is available within many types of road map representations. However, it is not based on actually observed link travel times but only on geometric properties of the links and their planned usage.

The Tele Atlas™ street graph of the city of Vienna [6] as used in the FLEET system is shown in Fig. 1. Dark edges represent FRC 0 to 3 which are treated within this paper, light gray edges are related to higher FRC.

![Fig. 1. Street Graph of Vienna (including all FRC) visualized in QGIS](image)

Table I shows the composition of this street graph, listing the number of links per FRC and the classes proportion of the total number of links. About 2.67% of the links are defined as road bound elements (i.e. FRC is set to -1).

In this paper a database of historical single speed measurements of the first half of 2008 has been used for the analysis. FCD obtained on road links belonging to FRC 0-3 has been selected as they are sufficiently covered by floating cars.
Table II lists the number of edges per FRC (FRC 0-3, a total of 9,300 edges 13.65% of the whole graph) for which at least one speed measurement was obtained in the observation period (99.56% of the 9,300 links).

Table II

<table>
<thead>
<tr>
<th>FRC</th>
<th>No. of Edges</th>
<th>Contingent [%]</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>930</td>
<td>1.36</td>
<td>Motorway, freeway and other major road</td>
</tr>
<tr>
<td>1</td>
<td>651</td>
<td>0.96</td>
<td>Major road less important than a motorway</td>
</tr>
<tr>
<td>2</td>
<td>3,331</td>
<td>48.9</td>
<td>Other major road</td>
</tr>
<tr>
<td>3</td>
<td>4,388</td>
<td>6.44</td>
<td>Secondary road</td>
</tr>
<tr>
<td>4</td>
<td>6,534</td>
<td>9.59</td>
<td>Local connecting road</td>
</tr>
<tr>
<td>5</td>
<td>4,809</td>
<td>7.06</td>
<td>Local road of high importance</td>
</tr>
<tr>
<td>6</td>
<td>16,664</td>
<td>24.46</td>
<td>Local road of minor importance</td>
</tr>
<tr>
<td>7</td>
<td>29,007</td>
<td>42.57</td>
<td>Other road</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0.01</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Total</td>
<td>68,136</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

During the observation period FCD was collected for \( N = 9,259 \) (99.56%) of the 9,300 links.

II. INDICATORS USED FOR DEFINING SIMILARITY

For each link \( E_i \) with \( i = 1..N \) the data set contains a set \( V_i \) of \( n_i \) aggregated speed measurements \( v \). These observations will be characterized using a set of indicators which subsequently will be used to define the similarity of the links.

The complete time series for one link for half a year (roughly 182 days with 96 fifteen minutes intervals per day) then consists of \( n_i = 17,472 \) measurement values. If the whole network of Vienna was fully covered with FCD, about 1.2 billion measurements (~10GB) were available for the calibration of prediction models. While complete FCD coverage is of course not available in real-world applications, the availability of FCD is expected to further increase e.g. by incorporating cell phones to provide additional data [7]. The FLEET system in Vienna generated a total of 74,343,965 measurements for FRC 0 to 3 in the first half of 2008.

The resulting empirical cumulative distribution function shows the distribution of the number of observations and hence the FCD coverage per link (Fig. 2). This indicates that for more than half of the links on average less than half of the measurements are available. Since the quality of prediction models is directly related to a sufficiently large number of observations for model calibration, the proposed clustering of similar edges compensates for incomplete FCD time series on some links as well as reduces the number of gaps in predictions.

![Empirical Cumulative Distribution of Measurements per Link](image)

Fig. 2. Empirical Cumulative Distribution of Measurements per Link

The main idea for the clustering consists of defining similarity of links based on indicators that characterize the speed distribution of each link. The set of indicators is constituted by nine quantities including the minimum, the maximum, the sample mean and median, the sample variance, the mode, skewness, the mean global Lyapunov exponent (MGLE) and the Shannon entropy. For a detailed description see Table III. Here \( P \) denotes the empirical cumulative distribution function and \( p_i \) the point mass of discretized (into 100 levels) speed measurements.

The indicators contain four measures of location (mean, mode and median, MGLE), three measures for spread (minimum and maximum combined, sample variance, Shannon entropy) and one measure of shape of the underlying probability density function (skewness). While one can argue for the in- or exclusion of any of these indices they appear to be useful for prediction purposes as is documented in section V.

Minimum speeds are between approx. 0 and 18 km/h, maximum values lie between 60 and 120 km/h. Mean values are mostly between 20 and 60 km/h.

Two indicators deserve separate mentioning: The MGLE compares the average mean of the logarithm to the logarithm of the first sample. As such it is more sensitive to small values than the sample average and hence provides a good indication of the frequency and depth of speed reductions.

The Shannon entropy finally measures the concentration of a time series which is related to the information content. It takes on its minimal value of zero if all observations are equal and its maximal value for a time series in which every speed measurement occurs the same number of times (i.e. a discrete uniform distribution).

These indices potentially are hampered by systematically missing observations. In this case similar roads according to our criterion also relates to similar systematic of missing observations. Indices that take such effects into account are...
subject of further research.

As seen in, the entropy and the MGLE are almost normally distributed.

Fig. 3. Distribution of the indicators for all links

The information content with respect to 100 subspaces of the value range of all available measurement values per edge is mostly between 4 and 6 bit, the mean entropy is above 5 bit at about 5.12 Shannon (Sh). Combining all indicators leads to a 9-dimensional representation (Degree of Freedom DoF = 9) of the speed dynamics for each link, a feature vector $x_i, i=1...N$.

$$x_i = \{ \text{Min } v, \text{Max } v, \text{Mean } v, \text{Median } v, \text{Var}(v), \text{Mode}(v), \text{Skew}(v), \text{MGLE}(v), E(v) \}$$

In order to account for different units of measurement between the coordinates of the feature vectors an adequate linear norming of indicator values to the interval [0.1, 0.9] was conducted to get normalized feature vectors $\tilde{x}_i$ for clustering. Fig. 4 shows a plot of these feature vectors, coloured according to FRC. The distributions shown in are now depicted in each column of the abscissa. It can be seen that links of FRC 0 tend to show the highest values for several indicators and lower skewness. Links of FRC 3 tend to show higher variances. FRC 1 to FRC 3 blur for nearly every indicator, except for a few links. Consequently Fig. 4 makes it obvious that the FRC classification does not lead to clusters of homogenous maximum, minimum, median or mean speed.

III. SELF ORGANIZING MAP AND NUMBER OF CLUSTERS

The Self Organizing Map (SOM) was developed by [8]. It is an unsupervisedly trained Artificial Neural Network (ANN) representing a method for exploratory visual data analysis, clustering and classification. The SOM is defined by identifying a set of weight vectors $w(i, j)$, called neurons, in feature space (in our case 9-dimensional) with a location $(i, j)$, $(i=1...p, j=1...m)$ in 2D, the so called Kohonen layer. The collection of all weight vectors is called codebook. Points in the Kohonen layer are arranged either on a hexagonal or orthogonal grid. The learning procedure is competitive and distributed; consecutively the feature vectors $x_i$ are presented to the algorithm for training. The Euclidean distances between the presented feature vector and the neurons in the Kohonen layer are computed. The neuron with smallest distance is called winner neuron and is moved into the direction of the presented vector. The neurons in the vicinity are also moved into this direction. In this way at the end of the training similar neurons are located close to each other on the 2D grid such that Euclidean distance in 2D is an approximation of Euclidean distance in feature space.
Accordingly each neuron already represents a cluster center for all feature vectors closest to this neuron. Further details on this ANN architecture and the SOM Toolbox for Matlab are given in [9]. Table IV shows the configuration parameters of the SOM used to cluster the feature vectors defined in Section III.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>35</td>
</tr>
<tr>
<td>p</td>
<td>50</td>
</tr>
<tr>
<td>Initialization</td>
<td>linear</td>
</tr>
<tr>
<td>Neighborhood grid</td>
<td>hexagonal</td>
</tr>
<tr>
<td>Kohonen Layer</td>
<td>sheet</td>
</tr>
<tr>
<td>Training</td>
<td>batch</td>
</tr>
<tr>
<td>No. of rough Tuning cycles</td>
<td>120</td>
</tr>
<tr>
<td>No. of fine Tuning cycles</td>
<td>120</td>
</tr>
</tbody>
</table>

For each cluster one prediction model is to be built. The number of clusters found either based on the whole data set or the preprocessed SOM codebook hence determines the amount of short term prediction models and the quality and amount of available historical data for prediction model calibration. This constitutes two conflicting aims which are to be balanced in the selection of the number of clusters.

To answer the question of how many clusters to choose for an optimal mapping of speed observations per edge, several clustering methods have been investigated. The goal is to minimize the overall errors between the cluster centers and the corresponding feature vectors while at the same time limiting the number of clusters in order to keep the complexity of modeling manageable.

A clustering algorithm can be applied either on the whole data set or on the reduced codebook of the SOM to make further investigations directly in this ANN. It has to be considered that the higher the amount of patterns, the higher is the computational effort of achieving clustering results. The building of a SOM offers the use of memory intensive clustering methods on the smaller representative codebook set. The following clustering methods have been used to estimate optimal numbers of clusters and their member patterns:

The neighbour distance matrix approach generates a cluster set of optimal size based on topological similarities with predefined neighbourhood size in the Kohonen layer [9].

K-means clustering, hierarchical clustering (ward) and Neural Gas [10] were used for comparing clustering results with the same number of clusters as found by autonomous approaches [9]. A k-means approach for the prediction of motorway speed patterns on link basis is described in [11].

A further approach for SOMs automatically detects clusters based on local minima of the U-matrix [9]. Hierarchical clustering has been used on all feature vectors, where average and ward linkage measures were used. An automated approach has been integrated for the use on the whole data set to determine the optimal number of clusters based on the in-sample classification error: If the cophenetic distance gradient falls below a threshold, no more increase in diversification will be reached by incrementing the number of clusters by one. In this case the classification error of the increased cluster set would be larger than of the original cluster set.

After clustering, the in-sample classification errors were estimated based on all feature vectors $X$: The quantization error $QE$, defined as the mean of all Euclidean norms between feature vectors $X_i$ and the corresponding vector for the cluster centre $c_{j(i)}$ (where $j(i)$ denotes the cluster to which feature vector $i$ corresponds) is a measure of overall classification quality:

$$QE = \frac{1}{N} \sum_{i=1}^{N} \| X_i - c_{j(i)} \|$$

Table V shows the quantization errors obtained for each method. Bold font for $n_C$ indicates that the number of clusters has been obtained automatically. The neighbour distance matrix approach with neighbourhood size N3 and ward linkage automatically produced $n_C = 24$ clusters. This is an acceptable number in terms of individual specification of prediction models. The approaches k-Means (5000 iterations), Hierarchical ward and Neural Gas (50 iterations) were manually set to produce 24 resp. 55 clusters for a direct comparison of cluster properties. Here Neural Gas showed the best results. The U-matrix local minima approach shows the lowest quantization errors based on 93 clusters and the SOM codebook. The automated hierarchical clustering approach used on all feature vectors produced 55 resp. 146 clusters (using ward resp. average linkage). It is expected that an increase of cluster numbers for k-Means and Neural Gas will show more competitive results in $QE$. A classification based on FRC showed the worst results.

### IV. Link Travel Time Prediction

In this section the average speed mid term prediction is investigated in the situation where no resp. only few measurements of a considered link are available. In this case the prediction is formed based on a model for the average
speed values. Since the average speed values will largely depend on the day as well as on seasonal effects (e.g. holidays) for many links insufficient data for the estimation of average values for each time interval will be available. In this situation combining the data from similar links provides a solid data base on which predictions can be based for the whole network. Thus in this section three different approaches to obtain predictions of average speed values are compared with respect to their accuracy:

- Predictions based on a historical record of the previous month for the considered link (denoted as single link in the following). This applies only if there are sufficient measurements (1 link).
- Predictions based on measurements for the whole FRC class (651 links).
- Predictions based on measurements for the corresponding cluster. The clusters generated by the U-matrix minimum approach have been chosen since it has shown best results (93 links).

In this respect a single link of FRC 1 and one cluster are used as an example. Time series of one month (04/2008) have been used for calibration of the conditional mean model. The model here separates weekdays (where public holidays are treated as Sundays) and an average speed value is estimated for each 96 fifteen minute time interval. The estimation is performed using linear local polynomial regression with manually tuned bandwidth of 30 minutes. Fig. 5 provides the mean curves estimated from the data of one of the considered links. The pronounced decreases of speed during the workday rush hours are clearly visible.

A comparison of the conditional means for Tuesday is shown in Fig. 6 for the three estimators described above (including estimated confidence bands – dash-dot lines – which are very small for the cluster and the FRC and thus very close to the mean). From these plots it can be seen that the FRC class 1 contains many links that show on average larger speeds resulting in a gross overestimation of typical link speeds. Since the clustering proposed in this paper is based on speed measurements the links contained in the cluster more closely correspond to the typical speeds of the considered link. However, the drop in speed during rush hour is less pronounced in the cluster than in the example link.

Fig. 7 provides mean absolute deviations (MAD) for the predictions for May 2008 (i.e. out of sample). From these plots it is clearly visible that using the FRC for the prediction results in less accurate predictions on average for all time instants. The mean absolute prediction errors from using the cluster means to represent the average speed measurements are lower during the whole day. The same is true for six out of seven links, when regarding the average MAD (see Fig. 8).

These seven links are those with FRC 1 which are also contained in the investigated cluster, which typically consists of similar links from various FRC. For the fifth link the speeds during the morning hours are comparatively higher than for other links inside the cluster, such that the higher average speed values corresponding to FRC 1 result in a smaller error.
V. CONCLUSION AND OUTLOOK

In this paper the first steps toward link travel time predictions for a whole network have been proposed. The main idea in this respect is to combine the data from similar links in order to provide prediction models to counter sparse data sets, due to the inevitable lack of full coverage of the whole network on the one hand and excessive modeling work on the other hand. In order to cope with the increasing computational complexity introduced by upscaling a localized approach to a whole network it is suggested to define the similarity of the links needs to be investigated on the one hand. In order to cope with the increasing computational complexity introduced by upscaling a localized approach to a whole network it is suggested to create clusters of similar roads for which predictions can be calculated using a single model.

The quality of such predictors depends on the amount of available historical data. The quality of clustering depends on a good choice of indicators for road similarity. In this respect the suitability of the FRC has been investigated. The results show that the FRC is not well suited to cluster similar roads with the purpose of using the cluster to define link travel time prediction models. The clustering approach proposed in this paper leads to better prediction accuracy while at the same time suggesting a higher number of clusters, which can be used to provide statistical mid-term predictions in real time and to calibrate simulation models. The proposed indicators have been selected to capture the dynamic properties of a link from speed time series obtained by FCD analysis. The same indicators could easily be used on data from loop detectors or other local detector technologies. Currently most indicators are general, statistical indicators which were selected for computational simplicity and performance reasons. In the future additional, traffic specific indicators could be added. For example such an indicator could characterize the relationship between free-flow speed and rush-hour speed.

The presented approach could be further improved by taking the properties of the measurement fleet into account. The probe vehicle coverage of a road could serve as an additional indicator, or it could be used to rate the reliability of the statistical indicators.

Clusters of similar roads can be useful for a variety of applications in traffic information processing besides reducing the number of nonparametric prediction models in large road networks. In [1] it is observed that no single best prediction approach exists and different approaches work better for different kinds of situations. Following this observation, the most suitable prediction approach for each cluster could be determined during a training phase to build hybrid predictors by means of methods such as model averaging. This would make it possible to combine the strengths of different predictors for a link-cluster based network-wide traffic prediction. Since these methods are computationally very expensive applying them for each single link is out of reach. A better road classification could also serve as an input to parametric traffic simulations and route planning. Clusters of similar roads could also be used to improve the data base for time series prototypes for medium-term prediction. In real-time traffic state estimation the data base could be used to provide more accurate fallback values on roads with low FCD coverage.

As a concluding remark, the results in this paper are seen just as the beginning of the analysis of the rich FC data set. In particular the modeling of one-step prediction models for each cluster is a logical next. In this respect it has to be mentioned that the selection of the best set of indicators to define the similarity of the links needs to be investigated further. Due to the size of the network and the required computational power this is a nontrivial task. The set proposed in this paper is seen as a good starting point in this direction.

REFERENCES