Transportation Research Procedia

Volume 15, 2016, Pages 513–524

ISEHP 2016. International Symposium on Enhancing Highway Performance

Spatio-Temporal Congestion Patterns in Urban Traffic Networks

Felix Rempe[1,2], Gerhard Huber[2] and Klaus Bogenberger[2]
[1]*BMW Group, Munich, Germany.*
[2]*University of the Federal Armed Forces Munich, Department of Traffic Engineering*
Felix.Rempe@bmw.de, Gerhard.Huber@unibw.de, Klaus.Bogenberger@unibw.de

Abstract
Traffic congestion in urban areas is a big issue for cities around the world. Thus, studying congestion and respective counter measures is of high importance for the increasing urbanization of society. Congestion analysis and forecast is most of the times done either on a link-wise network or on a network-wide level. Though, due to bottlenecks in the infrastructure and similar commuting patterns by road users, usually the same parts of an urban traffic network get congested. The idea is to observe and investigate primarily these most vulnerable parts of the network, which are denoted as congestion clusters, as they are crucial to both, drivers and operators. A methodology for determining congestion clusters is described, which provides a significant amount of flexibility to be able to meet different needs for different applications or cities. Based on a five months set of Floating Car (FC) data, the suggested methodology is tested. First analyses are conducted to understand up to which degree these clusters are able to represent the congestion level of the entire network. Besides, correlations between the clusters are investigated on a statistical basis and conclusions are drawn. The results provide a basis for potential traffic estimation and forecast systems.

Keywords: Probe data, network clustering, congestion analysis, traffic estimation, traffic prediction

1 Introduction

For traveller information systems as well as traffic control, it is fundamental to know where traffic jams occur and, in best case, to have reliable forecasts concerning their future development. Otherwise, neither can road users be informed adequately, nor can effective traffic management measures be executed to dissolve congestion.

However, monitoring an entire traffic network and providing forecasts for any possible location is very costly and, since significant parts of the network are never congested, also not necessary. This is also the fundamental motivation for the proposed approach: Start with identifying areas in which connected pockets of congestion typically emerge and reside, and then analyse primarily them instead

Selection and peer-review under responsibility of the Scientific Programme Committee of ISEHP 2016

doi:10.1016/j.trpro.2016.06.043

of the whole road network. These congestion-prone areas will from here on be denoted as congestion clusters. The hope is that congestion clusters can be understood as "neuralgic points" of the network. As such, they are most relevant to drivers and traffic management and possibly allow drawing conclusions on the traffic status of the whole network.

The paper is structured as follows: First, an overview on literature concerning urban traffic analysis and forecast will be provided. Then, the Floating Car (FC) data that are used to identify congestion clusters and the data preparation process are described. In a next step, it is explained which properties congestion clusters should show and how they can be computed algorithmically. Finally, a case study is executed. Munich (Germany) and its suburbs are used as test site for a five months period. The sensitivity of the cluster computation methodology on its input parameters is discussed. Then, the congestion in the clusters is analysed. First focus is laid on the typical congestion starts and ends, and second, the correlation between pairs of clusters is determined.

2 Related Work

Traffic forecasts provide valuable information for traveller as well as traffic management and control. With the rapid spread of mobile sensors the availability of traffic data increased dramatically and new method are developed that apply this type of data to problems such as traffic estimation and prediction (Corrado de Fabritiis, 2008), (Herring, Abbeel, Hofleitner, & Bayen, 2010). Short-term traffic prediction has been subject of many works. In (Vlahogianni, Karlaftis, & Golias, 2014) there is a detailed review of current approaches. Usually the methods are classified into parametric and non-parametric approaches. Parametric approaches define a model and fit the parameters of the model to the data. Non-Parametric approaches on the other hand have a flexible structure and a variable number of parameters. (Lippi, Bertini, & Frasconi, 2013) (Karlaftis & Vlahogianni, 2011) present comparisons of different methods and their accuracy in traffic forecasting. In (Vlahogianni, Karlaftis, & Golias, 2014) 10 mayor challenges for upcoming research are pointed out. One is to focus on network level spatio-temporal approaches. In the past year several works regarding that have been published. To mention a few, (Min & Wynter, 2011) apply a multivariate spatial-temporal autoregressive model on a sample network with different road categories. (Kamarianakis & Prastacos, 2005) model the traffic flow in space and time using a Space-Time Autoregressive Integrated Moving Average (STARIMA) model. (Yue & Yeh, 2008) analyse the spatio-temporal characteristics of flow on highways. (Cheng, Haworth, & Wang, 2012) compute correlations between links in the London traffic network in order to analyse required model complexities for models such as STARIMA. Most of the literature is based on small to medium sized networks that model dependencies between road links. For bigger networks, the computational expense increases dramatically to compute the correlations between all links. In order to deal with the computational costs, Neighbourhood Selection Techniques (STN) haven been developed. (Gao, Sun, & Shi, 2011) apply a graphical Lasso approach. (Haworth & Cheng, 2014) give a comparison about different techniques. However, those are still based on a link level approach. Road networks of bigger cities consist of a huge number of links. Furthermore, the traffic state at each time and each segment is often perturbed by nonlinear dynamics. Therefore, another approach is to analyse a traffic network from a macroscopic perspective. (Ji & Geroliminis, 2012) (Ji, Luo, & Geroliminis, 2014) partition a road network into clusters with similar properties in order to (1) determine a macroscopic fundamental diagram and (2) to observe congestion propagation in urban networks. The present work combines both approaches, which, to our knowledge, has not been done yet. First a network is partitioned into frequently congested clusters, followed by a spatio-temporal congestion analysis between the clusters. The aim is to show that the resulting clusters that simplify the complex network allow computing congestion estimations and predictions in a road network.

3 Description and Preparation of Available Traffic Data

As test site, the city of Munich (Germany) is chosen (see Figure 2). FC data form the data basis for the following analysis. The FC data are collected by a fleet of private vehicles that frequently report their GPS positions. An internal logic in the vehicle compares the current velocity with an assumed free-flow speed that depends on the current location. If both differ, then the current position is sent to a server. The fleet delivers approximately 25,000 traces (sequences of GPS positions, one for each recorded trip) each day for Munich. The collected GPS positions are matched onto a digital map (describing the map-matching extents the scope of this paper. Please refer to (Quddus, Ochieng, & Noland, 2007) for details regarding usual map-matching procedures) and speeds are computed from the travel time between the positions. This digital map represents the considered part of the road network and can be understood as (directed) graph $\vec{G} = (V, \vec{E})$. The set of nodes V represents intersections and the set of edges $\vec{E}$ represents road segments. Here, it is only stated that at the end, a huge collection of link and time related driving speeds $\{v(e_k, t_k)\}_{k=1,2,\dots}$ is obtained for $e_k \in \vec{E}$ and t_k denoting a specific point in time for each k. These data are aggregated for each edge and discretized over time with a resolution of one minute. In other words: Each considered day is separated into $1440 = 24 * 60$ time intervals. Each time interval describes a timespan of one minute. For each link and each time interval, all available driving speeds are averaged arithmetically. The resulting aggregated speeds are denoted by $V_{Rec}(e, t)$. Thereby, $V_{Rec}(e, t)$ is understood as a function that returns for any $e_k \in \vec{E}$ and any time t the corresponding driving speed. Note that recorded driving speeds $v(e_k, t_k)$ are not available for each link and each time interval. However, for the further proceeding, it is necessary to have one unique estimation of speed for any point in time and any link in the network. Consequently, to fill the gaps, the existing data are extrapolated in time. This means that for each time interval, for which no recorded speed is available, the last measured speed is used. Certainly, this is only done if the timespan between last measurement and the considered point in time is not too long. For the described research, this timespan was set to 15 minutes. To fill the remaining gaps, one assumes free-flow traffic conditions. This is critical, since it is not possible to decide whether there is no car on-site or cars move with free-flow speed and simply do not report it. However, this is the typical proceeding in literature (Ji, Luo, & Geroliminis, 2014) (Saeedmanesh & Geroliminis, 2015), as it is the probably the best one can do.

4 Methodology: Congestion Cluster Identification

For the considered road network of Munich, as supposedly for many others, pockets of congestion typically appear in different regions. Here, a congestion pocket is basically defined as a connected part of the road network which consists (at a certain time) solely of congested segments.

These pockets change their shape over time as they propagate through the network, split up or merge with other pockets. Thereby, the critical areas, i.e., the parts of the road network where these pockets occur and reside, often remain the same. This regularity is a consequence of similar origin-destination relations among road users and the static nature of road networks (and their capacities).

The goal is to identify such areas, which are denoted as congestion clusters. Clusters are intended to fulfil three fundamental properties. First, it is postulated for congestion clusters that they are static, i.e., their shape does not change over time. This is a contrast to former work (Ji, Luo, & Geroliminis, 2014) (Saeedmanesh & Geroliminis, 2015), where congestion clusters are considered as dynamically changing parts of the network. Keeping them static will lead to issues concerning their computation. At the same, this property is fundamental from a practical perspective: Dynamically changing parts of the road network can hardly be observed steadily.

Second, congestion cluster need to span often congested parts of the network. This means that solely such road segments can be assigned to clusters that are frequently congested. The last property is compactness or connectivity of all edges belonging to the same cluster.

Other or further criteria are also possible. In (Ji, Luo, & Geroliminis, 2014) for instance, clusters are intended to achieve a high "intra-cluster similarity". This means that all segments belonging to the same cluster are intended to show a similar development of driving speeds over time. If one thinks of prediction procedures, it seems reasonable to demand high intra-cluster similarity. In this case, one could reduce prediction approaches to a cluster level (i.e., predictions are solely made for clusters and not on a link-level) without losing too much information, since all edges of the same cluster behave similarly.

It may also be interesting to check up to which degree the entire network can be represented by the union of all clusters. Especially for traffic management purposes, it would be very helpful to be able to concentrate observation on only a small part of the road network and, at the same time, to be able to draw conclusions on the entire network (or at least on the most critical parts of the remaining network).

The cluster computation is supposed to fulfil all these cluster properties. The algorithm is depicted in Figure 1. Other, potentially interesting cluster properties (intra-cluster similarity, the ability to represent additional parts of the entire network) are also ensured by the suggested approach - at least up to some degree. Certainly, especially these additional properties strongly depend on parameters that are applied within the approach. A corresponding analysis will be given in section 5.1. The understanding of the authors thereby is that the list of desired properties depends on the intended purpose (for instance, use cluster analysis as basis for forecasts or monitoring) and should not generally be predetermined.

4.1 Dynamic Congestion Pockets

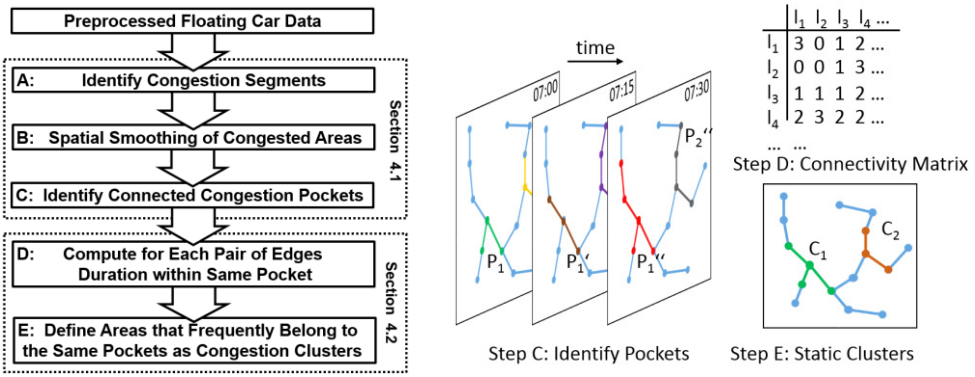


Figure 1: Flowchart and schema describing the methodology of computing congestion clusters

Before describing the clustering procedure, first a formal definition of congestion pockets is given. Let for some time t and each edge $e \in \vec{E}$ a recorded driving speed $V_{Rec}(e,t)$ be given. It is assumed that a free-flow driving speed $V_{max}(e)$ is available for any edge e. Then, relative driving speeds are defined as

$$V_{Rel} := \frac{V_{Rec}(e,t)}{V_{max}(e)} \tag{1}$$

The definition of relative driving speeds is necessary to describe under which conditions an edge is interpreted as congested. For this purpose, let a parameter $v_{crit} \in [0.0, 0.1]$ be given that defines up to which relative driving speed an edge e is seen as congested. Furthermore, let with J ("J" for "jam") a function be denoted that assigns for any time t the value of one to an edge e if and only if this edge is congested at time t:

$$J(e,t) \coloneqq f(x) = \begin{cases} 1, & if\ V_{Rel}(e,t) \leq v_{crit} \\ 0, & else \end{cases} \tag{2}$$

A similar definition of congested edges can be found in (Ji, Luo, & Geroliminis, 2014). It allows providing a formal introduction of congestion pockets: For some time t and a congested edge e^*, a congestion pocket $P(e^*,t)$ is defined as the set of all edges $e \in \vec{E}$ that have the subsequent properties:

1. $J(e,t) = 1\ \forall e \in P(e^*,t)$
2. Either there exists a path from $e^*$to e on $\vec{G}$ or a path from e to e^* on $\vec{G}$ that consists solely of edges to which J assigns a value of one.

To be able to differentiate between pockets, they are associated with single edges. Each pocket could be associated with any of the edges that belongs to it. An edge, on the other hand, always belongs solely to one pocket. Note that pockets are defined dynamically, i.e., they depend on time. This is their main difference to congestion clusters, which will be constructed as time-independent (i.e., static) parts of the road network. The definition of congestion pockets as connected subgraphs of $\vec{G}$ is rather intuitive. However, when considering real traffic data, the number of congestion pockets can become very high. An often occurring problem is that areas, which visually seem to belong to the same congestion pocket, are separated by single, non-congested edges. In many cases, this is rather a lack of data than a real separation between congestion pockets. Hence, a spatial smoothing is carried out as it suggested in (Ji, Luo, & Geroliminis, 2014). There, after determining the set of congested edges for some time t, all edges which have more congested neighbours than non-congested neighbours are understood as congested edges, too. This means that function J is redefined as stated below:

$$J(e,t) \coloneqq f(x) = \begin{cases} 1, & if\ V_{Rel}(e,t) \leq v_{crit} \\ 1, & if\ |\{e' \in N(e): V_{Rel}(e',t) \leq v_{crit}\}| > 0.5\,|N(e)| \\ 0, & i_{Rel} \end{cases} \tag{3}$$

Thereby, $|A|$ denotes the cardinality of set A and $N(e)$ denotes the neighbourhood of e, i.e., the set of all edges in $\vec{E}$ that share at least one node with e (except for edge e itself). Having function J adjusted, the definition of congestion pockets remains basically the same.

4.2 Construction of Static Congestion Clusters

To generate static congestion clusters, one proceeds as follows: First, information on driving speeds depending on time and location is prepared (as it is sketched in section 3). This information is again represented by $V_{rec}(e,t)$. In a next step, relative driving speeds V_{Rel} are computed and a critical relative threshold v_{crit} speed is defined. Based hereon, congestion pockets are computed for each time $t \in T$ as explained in section 4.1. Thereby, T describes the set of all time intervals for which traffic data are available. Now, the actual cluster computation can take place. It starts with computing for each pair of edges $e_1, e_2 \in \vec{E}$ the number of times $D(e_1,e_2)$ ("D" for duration) for which these edges were part of the same congestion pocket:

$$\begin{aligned} D&: \vec{E} \times \vec{E} \to \{0,1,\dots,|T|\}| \\ D(e_1,e_2)&:= |\{t \in T : e_1 \in P(e_2,t)\}| \end{aligned} \tag{4}$$

The more often two edges are part of the same pocket, the more similar their congestion behaviour typically is. Besides, being often part of the same congestion pocket ensures also spatial proximity. Afterwards, one assigns edges that are often part of the same pocket to the same cluster. Thereby, one proceeds iteratively. In each iteration, the pair of edges (e_1^*, e_2^*) is considered that shows the highest D-value, i.e.:

$$(e_1^*, e_2^*) \coloneqq argmax(D(e_1, e_2): e_1, e_2 \in \vec{E}, e_1 \neq e_2 \tag{5}$$

Both edges are then assigned to the same cluster. If exactly one of both edges has already been assigned to a cluster, then the other edge is assigned to this cluster, too. If none of both edges is assigned to a cluster, a new cluster is generated. It consists at that time solely of e_1^* and e_2^*. If both edges have already been assigned to clusters (and not to the same cluster), then a new cluster is generated by fusing both existing clusters. At the end of each iteration, in order to exclude this pair of edges from further considerations, the corresponding value $D(e_1^*, e_2^*)$ is set equal to zero. To assign only frequently congested edges to clusters, a minimum congestion duration $D_{min} \in \{1, \dots, |T|\}$ is introduced for this iterative procedure. This means that one stops the iteration as soon as the highest remaining D-value falls below D_{min}. Consequently, in the end one receives a set of clusters that contains exactly the set of edges for which another edge exists that shared at least D_{min} time periods within the same congestion pocket:

$$\vec{C} \coloneqq \bigcup_{i=1,2,\dots,nc} \overrightarrow{C_i} = \left\{ e \in \vec{E}: \exists e' \in \vec{E} \; with \; D(e, e') \geq D_{min} \right\} \subseteq \vec{E} \tag{6}$$

Here, $nc \in \mathbb{N}$ denotes the number of computed clusters and $\overrightarrow{C_i}$ with $i \in \{1,2, \dots, nc\}$ denotes the i-th cluster. For practical use, the parameter $\alpha \in [0,1]$ is introduced that chooses D_{min} (independent of T) as the α - quantile of all $D(e, e^*)$:

$$\begin{aligned} D_{min} &= quantile(D, \alpha) \; with \\ D &:= \{D(e_1, e_2): e_1, e_2 \in \vec{E} \; and \; e_1 \neq e_2\} \end{aligned} \tag{7}$$

The suggested clustering approach is designed in such a way that the three fundamental cluster properties can be expected. The suggested method obviously ensures that any of the computed clusters is static. Furthermore, high congestion rates of all edges belonging to clusters are guaranteed by an appropriately chosen value α (or D_{min}, respectively) and step E. Note that connectivity of clusters cannot be guaranteed by the described proceeding. However, for the analysis which is described in section 5 this turned out to be no issue. Steps C and D indirectly achieved cluster connectivity in almost all cases.

The congestion probability of the clusters can be regulated by adjusting α accordingly. Returning to the two "optional" cluster properties, it can be observed that a certain level of intra-cluster similarity is implicitly achieved by the suggested methodology. Edges within the same cluster are in most cases frequently part of the same congestion pocket and hence need to be congested during the same time periods. The idea that clusters may be used to represent larger parts of the road network, on the other hand, seems not to be mirrored by the design of the suggested procedure. With α and v_{crit}, however, two degrees of freedom exist within the clustering approach. One could use them to achieve or strengthen additional properties. For this purpose, indices that allow quantifying up to which degree a specific set of clusters fulfils these properties are introduced. Intra-cluster similarity, for instance, can for each cluster $\overrightarrow{C_i}$ be quantified by its counterpart, the *dissimilarity*:

$$\Delta(\overrightarrow{C_i}) := \frac{\sum_{t \in T} \sum_{e_1 \in \vec{C}_i} \sum_{e_2 \in \vec{C}_i} \left(V_{Rel}(e_1, t) - V_{Rel}(e_2, t)\right)^2}{|T| * \left|\overrightarrow{C_i}\right| * \left|\overrightarrow{C_i}\right|} \tag{8}$$

This measure of dissimilarity is oriented on the one that was introduced in (Ji & Geroliminis, 2012). However, here a temporal component has been included, whereas in (Ji & Geroliminis, 2012) speeds for a single point in time have been considered. High $\Delta(\overrightarrow{C_i})$-values indicates low intra-cluster similarity.

One possibility to quantify the ability of a set of clusters $\overrightarrow{C_1}, \dots, \overrightarrow{C_{nc}}$ to represent the traffic situation of the entire network is the so-called Pearson correlation:

$$D(\vec{N}, t) \coloneqq \frac{1}{L(\vec{N})} * \sum_{e \in \vec{N}} J(e,t)L(e) \; with \; \vec{N} \subseteq \vec{E} \tag{9}$$

$$\rho(\vec{E}, \vec{C}) \coloneqq \frac{cov\left(D(\vec{C}, t), D(\vec{E}, t)\right)}{\sigma\left(D(\vec{C}, t)\right) * \sigma\left(D(\vec{E}, t)\right)} \tag{10}$$

where $L(e)$ is the length of edge e, $L(\vec{N})$ is the total length of all edges in an edge set $\vec{N}$, cov is the covariance and $\sigma(\dots)$ are the standard deviations of functions $D(\vec{C}, t)$ and $D(\vec{E}, t)$ over time t, respectively. Thus, $D(\vec{C}, t)$ describes up to which extend all clusters C_i combined are congested at time t. $D(\vec{E}, t)$ correspondingly describes the overall congestion percentage of the network. Then, $\rho(\vec{E}, \vec{C})$ describes how strong the level of congestion inside the clusters $\vec{C}$ correlates with the level of congestion in the entire network.

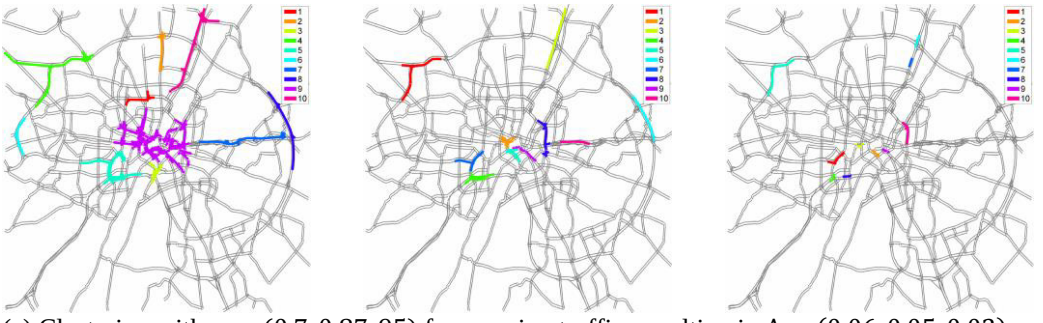
5 Case Study: The Munich Road Network

The clustering algorithm is fed with traffic data that are available for the Munich road network. The traffic in Munich is dominated by commuters that head towards the city in the morning and leave the city in the evening. This increased demand causes congestion on a very regular basis. The goal is to identify these regularities and find patterns regarding congestion states in everyday traffic. Therefore, we first present the results of the clustering algorithm based on different parametrizations.
Afterwards, we analyse the behaviour of an exemplary set of clusters for a duration of 5 months. Thereby, all "non-typical" days will be excluded from consideration to minor scatter.

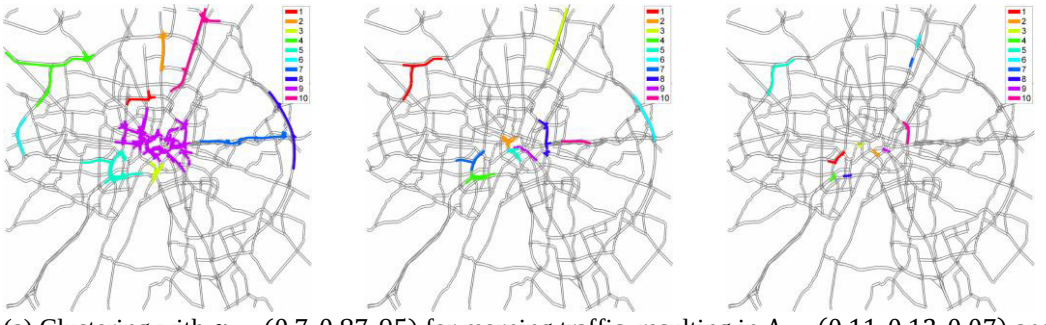
5.1 Static Clusters in Munich Road Network

To generate clusters, the cluster generation procedure is executed on a basis of four weeks of data from the 3rd of November 2014 to the 30th, a time without vacation. v_{crit} is kept fix with a value of 0.5 and several different values for α are tested (0.7, 0.87 and 0.95). Besides, the data is split up into a morning period (0 - 12am) and an evening period (12am - 12pm). This is up to some degree a contradiction to the idea of static clusters, but the suggested methodology does not allow an edge to be part of two different clusters. If the data was not split, the location and size of many clusters would blur, since the congestion behaviour of many edges prompts that they are part of a morning and an evening cluster. Another important observation is that the number of the resulting cluster becomes very high (more than 100). To counter this, only the biggest cluster in terms of covered length are chosen for further considerations. For the described analysis, their number was reduced to ten. All other clusters

were discarded. For each of the three example parametrizations, the aforementioned intra-cluster similarity and the Pearson correlation are computed.



(a) Clustering with $\alpha = (0.7, 0.87, 95)$ for morning traffic, resulting in $\Delta = (0.06, 0.05, 0.03)$ and $\rho = (0.97, 0.97, 0.93)$



(a) Clustering with $\alpha = (0.7, 0.87, 95)$ for morning traffic, resulting in $\Delta = (0.11, 0.12, 0.07)$ and $\rho = (0.97, 0.93, 0.89)$

Figure 2: Static clusters of morning and evening traffic depending on different parameters α. The number of clusters is set to 10 and the threshold v_{crit} to 0.5.

Figure 2 shows the resulting morning and evening clusters for different parametrisations. Moreover, also the corresponding Pearson correlation $\rho(\vec{C}, \vec{E})$ of the edges inside the clusters $\vec{C}$ and the entire network $\vec{E}$ as well as the averages of all $\Delta(\vec{C}_i)$ for $i = 1, \dots, 10$ are displayed.

It can be observed that α has a significant influence on the appearance of the clusters. The higher the value of α, the smaller the clusters get. Furthermore, it can be observed that together with α, also the intra-cluster dissimilarity Δ increases. (With one exception for $\alpha = 0.87$ for the evening clusters). The Pearson correlation, on the contrary, gets reduced. That shows that smaller the clusters are less sensitive to changes of congestion in the entire network. Different values for v_{crit} have a comparably strong influence, but this won't be considered here.

5.2 Distinguishing between Regular and Irregular days

The sets of morning and evening clusters that result when setting α equal to 0.87 are from here on considered. The behaviour of these clusters will be analysed in the following for a time period of five months, from August 1st, 2014 until December 31st, 2014. To reduce scatter, only "regular" days will be considered. Thereby, a day is denoted as regular if the set of clusters shows an average congestion behaviour with regard to the corresponding weekday.

For the described study, altogether 20 weeks (j is used as index for weeks) of data are considered to analyse the behaviour of the aforementioned set of clusters. The weekdays are encoded via an index $k \in$

{1,2, ... 7}, where 1 encodes Mondays, 2 encodes Tuesdays and so on. Furthermore, it is assumed that t denotes the considered minute of day, i.e., $t \in \{1,2, \dots 1440 = 24 * 60\}$. Then, $t_{k,j}$ encodes the t-th minute for the k-th weekday in the j-th week. Hence, $420_{2,5}$ encodes the time span 06:59 - 07:00 on the Tuesday of the fifth week. Based hereon, $\widetilde{D}_k(\vec{C}, t)$ is defined as the median congestion percentage over all clusters at time t for weekday k. The median is here used instead of the arithmetic average to achieve robustness against outliers.

Figure 3 illustrates the median of each weekday over time. We can see that in median case the weekdays Monday-Thursday show a very similar congestion state having peaks at approximately 8am and 6pm. Fridays have lower peaks and the congestion level is higher in the early hours of afternoon than on any other day. Saturdays and Sundays show low median congestion. These results match with the commuting behaviour in the city: For most workers Monday-Thursdays are similar days, on Friday fewer people commute to work, plus, the leave work earlier than on the other days.

We define a day (k, j) as *regular* if the following regularity index does not exceed the threshold $RI_{thres} = 0.5$:

$$RI_{k,j} = \frac{\sum_{t=1,\dots,1440} |\, \widetilde{D}_k(\vec{C}, t_{k,j}) - D(\vec{C}, t_{k,j})|}{\sum_{t=1,\dots,1440} \widetilde{D}_k(\vec{C}, t_{k,j})} \tag{11}$$

The resulting set of N_R regular days is denoted by R. From the originally 153 days in the period of 5 month, a total of 32 was filtered out using the regularity index. The analysis of the filtered days revealed the most probable reasons for irregular congestion behaviour: missing data due to technical issues, official holidays on weekdays and winter vacation.

5.3 Cluster Congestion Analysis

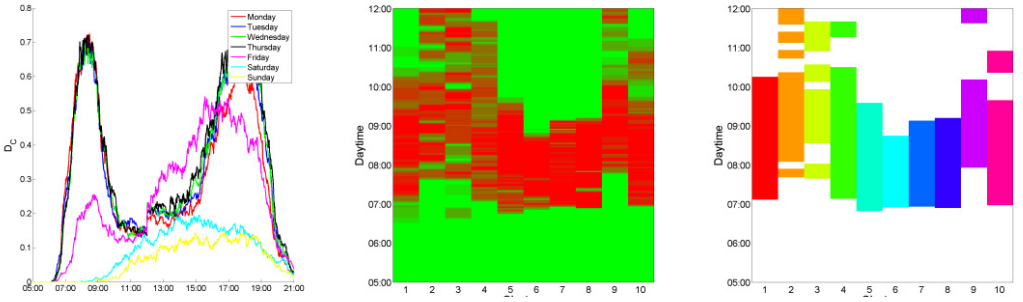


Figure 3: Hourly median total congestion level for each weekday (left), individual congestion level of each cluster for half a day (mid), binary congestion level of each cluster after hysteresis-filtering (right).

Subsequently, typical times when clusters usually get congested and when congestion dissolves intra-day are analysed. For this purpose, the level of congestion $D(\vec{C}_i, t_{k,j}) \in [0,1]$ is computed for each cluster separately for all regular days. To remove perturbations and transform the time series $D(\vec{C}_i, t_{k,j})$ into a binary signal, we introduce two thresholds: c_{low} and c_{up} and perform a hysteresis-based processing of $D(\vec{C}_i, t_{k,j})$ into the states 0 (not congested) and 1 (congested). Thus, a cluster $\vec{C}_i$ can only change its status from zero to one if $D(\vec{C}_i, t_{k,j}) \geq c_{up}$. The other way round, it can only switch to status zero if $D(\vec{C}_i, t_{k,j}) < c_{low}$. Besides, a minimal duration of states is postulated to smooth peaks and drops in the signal. This means that if a cluster is congested for less than 15 consecutive one-minute time intervals, then its status is set back to zero throughout this period. In a second step, all non-congested states that persist less than 15 minutes are set to congested. The influence of this process can be seen

exemplarily in the middle and the right part of Figure 3 for the morning traffic of one day. On the x-axis, the ten clusters are listed and the y-axis contains the time. In the middle figure, red areas indicate time intervals for which the congestion percentage of a cluster $D(\vec{C}_i, t_{k,j})$ is close to one at that specific time, green areas correspondingly indicate low percentages. In the right figure, the congestion states (*white* means state zero, *colour* means state one) that result from the smoothing process for one example are shown.

The starting time $tsm_{i,k,j}$ of the congestion in cluster i for the morning data is defined by the first time, for which cluster $\vec{C}_i$ reaches state one (congested) on that half-day. The ending time $tem_{i,k,j}$ correspondingly is the last time (up to 12am for the morning period) for which state *congested* is obtained. In case no congestion happened on that day, the values are set to zero. The times $tse_{i,k,j}$ and $tee_{i,k,j}$ analogously denote the starting and ending times of congestion in the evening clusters, i.e., for the second half of the day.

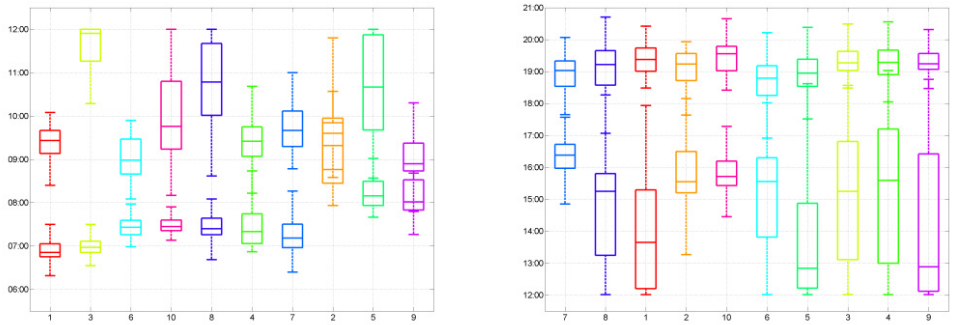


Figure 4: Distributions of starting and ending times for each cluster in the morning (left) and in the evening (right). The boxes describe the $q_1 = 25$ and $q_2 = 75$ percentiles of the data, the whiskers include all data in the range of $[q_1 - 1.5(q_3 - q_1), q_3 + 1.5(q_3 - q_1)]$

Figure 4 shows the distributions of the variables tsm, tem, tse and tee based on data collected on regular days and sorted in descending order by the distance to the city centre from the left to the right. We observe several things: (1) The starting times of congestion in the morning depends on the distance to the city centre. Especially clusters 1 and 3 emerge early, while clusters 2, 5 and 9, which are part of the city centre, get congested comparably late. (2) The congestion in all evening clusters dissolves without exception at around 7pm. (3) The spread of congestion starting times in the morning and congestion ending times in the evening is considerably smaller for all clusters than for congestion ends in the morning and congestion starts in the afternoon.

The preceding analysis shows the behaviour of individual clusters. However, also the relations between different clusters could reveal spatio-temporal characteristics of congestion in an urban network. For this purpose, we define the cluster correlation $\theta_m(\vec{C}_1, \vec{C}_2)$ between cluster $\vec{C}_1$ and $\vec{C}_2$ as:

$$\theta_m(\vec{C}_1, \vec{C}_2) := \frac{1}{|R|} * \sum_{(k,j)\in R} S_m(\vec{C}_1, \vec{C}_2, k, j) \tag{12}$$

$$S_m(\vec{C}_1, \vec{C}_2, k, j) := \begin{cases} 1, & if\ tsm_{1,k,j} = tsm_{2,k,j} = 0 \\ 1, & if\ tsm_{1,k,j} * tsm_{2,k,j} > 0 \\ 0, & else \end{cases} \tag{13}$$

The index m thereby indicates that the corresponding quantities are computed for morning periods and can analogously be computed for evening periods. $\theta_m(\vec{C}_1, \vec{C}_2)$ describes the probability that two clusters behave similar during the morning period of a day. "Similar" in this context means that there exists either a time in the morning where the two clusters are congested (condition two in equation (**13**)), or both clusters are inactive for the complete morning (condition 1).

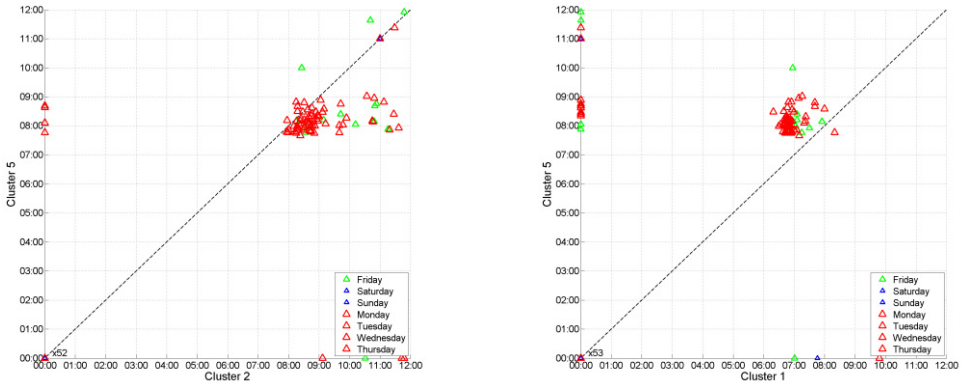


Figure 5: Congestion starts for the clusters with the highest correlation (left); Congestion starts for a motorway cluster (1) and a cluster in the city center (5).

The correlation reveals several cluster combinations that correlate well. Figure 5a) shows the starting and ending times of clusters 2 and 5, which have a strong correlation (> 0.95). Since both of them are located in the city centre, a high correlation is very likely. The data are mostly located right of the bisecting line. This means that in most cases cluster 2 got congested later than cluster 5.

Another example, where the spatio-temporal dependency is more distinctive, is given in Figure 5b). It shows how cluster 1 (motorway) and 5 (city centre) behave. There is a high data density around (7am, 8am), several data points are on the y-axis, there are many days without any congestion (counted by the number at the bottom left of the scatter plot) and two days where there was only congestion in cluster 1 but not cluster 5 (this is indicated by the two triangles on the bottom of the scatter plot; remember that $tsm_{5,k,j}$ is equal to zero if no congestion occurred for cluster 5 on day (k, j)). Thus, on all days (except 2) it holds that if cluster 1 got congested, also cluster 5 got congested, but approximately one hour later.

The reason for this dependency is probably the fact that commuters heading towards the city centre cause first a high demand and thus congestion on the motorways leading and later reach the city centre causing congestion there.

Applying findings like these in an online traffic forecast means to observe the congestion state of each cluster. In the case one cluster gets congested, the dependencies with other clusters are checked. If there is a high correlation between them, it is probable that the other cluster will get congested, too. The analysis of the distribution of samples possibly allows predicting the approximate time of congestion in other clusters. However, for an online prediction system, the presented methodology has to be adapted, in order to identify irregular days in real-time.

6 Conclusion and Outlook

Traffic congestion in urban areas is an increasingly severe problem for major cities. Estimating historic, current and prospective traffic is crucial to take measures against traffic jams in road networks. The quick spread of mobile traffic sensors allows to gather network-wide traffic data und thus, allows to analyse traffic congestion on a wider scale and in more detail than it has been possible before.

In this paper, we first presented an algorithm to reduce a complex traffic network to the parts that are most frequently congested (clusters). The clustering method has been applied to the city Munich, where the traffic conditions can be estimated using a huge set of FC data over five months. The subsequently applied analyses of the congestion behaviour of the resulting clusters allows (1) identifying weekdays that do not behave regularly and, thus, can be classified as outliers, (2) estimating the times and variances of the congestion in each cluster and (3) quantifying how strong different clusters correlate with regard to their congestion behaviour. Applications of the methods and results can be used to control

urban traffic in order to avoid or reduce the negative effects of congestion. Furthermore, the statistical results can be used as traffic forecasts for road users such that they can react and take other routes, drive at other times or take other transportation modes.

Among the presented, more experiments can be and should be done. For now, the clustering has only been tested on the network of Munich but needs to be applied to other cities that may have differing traffic patterns. Finally, the obtained results provide a promising basis for traffic forecasts, but developed method still need to be implemented and tested in an online traffic forecast system.

References

Cheng, T., Haworth, J., & Wang, J. (2012). Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems, 14*(4), pp. 389-413.

Corrado de Fabritiis, R. R. (2008). Traffic Estimation and Prediction Based on Real Time Floating Car Data. *Intelligent Transportation Systems (ITSC), 11*, pp. 197-203.

Gao, Y., Sun, S., & Shi, D. (2011). Network-Scale Traffic Modeling and Forecasting with Graphical Lasso. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, . . . H. He, *Advances in Neural Networks – ISNN 2011* (Vol. 6676, pp. 151-158). Berlin, Heidelberg: Springer Berlin Heidelberg.

Haworth, J., & Cheng, T. (2014). A Comparison of Neighbourhood Selection Techniques in Spatio-Temporal Forecasting Models. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2*, pp. 7-12.

Herring, R., Abbeel, P., Hofleitner, A., & Bayen, A. (2010). Estimating arterial traffic conditions using sparse probe data. *IEEE Transactions on Intelligent Transportation Systems*(13), pp. 929-936.

Ji, Y., & Geroliminis, N. (2012). On the spatial partitioning of urban transportation networks. *Transportation Research Part B: Methodological, 46*(10), pp. 1639-1656.

Ji, Y., Luo, J., & Geroliminis, N. (2014). Empirical Observations of Congestion Propagation and Dynamic Partitioning with Probe Data for Large-Scale Systems. *Transportation Research Record: Journal of the Transportation Research Board, 2422*, pp. 1-11.

Kamarianakis, Y., & Prastacos, P. (2005). Space–time modeling of traffic flow. *Computers & Geosciences, 31*(2), pp. 119-133.

Karlaftis, M., & Vlahogianni, E. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies, 19*(3), pp. 387-399.

Lippi, M., Bertini, M., & Frasconi, P. (2013). Short-Term Traffic Flow Forecasting: An Experimental Comparison of Time-Series Analysis and Supervised Learning. *IEEE Transactions on Intelligent Transportation Systems, 14*(2), pp. 871-882.

Min, W., & Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies, 19*(4), pp. 606-616.

Quddus, M. A., Ochieng, W. Y., & Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies, 15*(5), pp. 312-328.

Saeedmanesh, M., & Geroliminis, N. (2015). Clustering of heterogeneous networks with directional flows based on "snake" similarities. *94th Annual Meeting of the Transportation Research Board*.

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies, 43*, pp. 3-19.

Yue, Y., & Yeh, A. G.-O. (2008). Spatiotemporal traffic-flow dependency and short-term traffic forecasting. *Environment and Planning B: Planning and Design, 35*(5), pp. 762-771.