Reprint of publication from

IADIS European Conference on Data Mining 2011

“Interactive Anomaly Detection in time series resulting from local traffic measurements”

Andreas Theissler, Jochen Palmer, Hubert Rehborn, Ian Dear

In Proceedings of the IADIS European Conference on Data Mining 2011.

IADIS Press, Lisbon. 2011.
pages: 147-152

published July 2011
INTERACTIVE ANOMALY DETECTION IN TIME SERIES
RESULTING FROM LOCAL TRAFFIC MEASUREMENTS

Andreas Theissler, Jochen Palmer
IT-Designers GmbH, Germany
{andreas.theissler, jochen.palmer}@it-designers.de

Hubert Rehborn
Daimler AG, Germany
hubert.rehborn@daimler.com

Ian Dear
Brunel University, United Kingdom
ian.dear@brunel.ac.uk

ABSTRACT
The identification and optimization of problem sections in the road network, as well as the estimation and optimization of the time of travel is a challenge of great interest in the field of traffic engineering. The economic damage caused by congested traffic is tremendous. Complex and application-specific models reconstructing the current traffic patterns from measured traffic data are used in online-analysis applications like dynamically changing road signs. Yet, setting up a model requires its configuration, interfacing of the model to a specific data source and most likely some development effort. An application-independent visual analytics approach on the other hand can operate directly on the data, no configuration and no development time is required. We show how a visual analytics approach can be used for the offline-analysis of traffic data. We adapted and enhanced parallel coordinates [5] to cope with the spatio-temporal structure of traffic measurements, allowing to interactively explore the data. In this publication we focus on the detection of anomalies in long-term traffic measurements and refer to this as interactive anomaly detection. Using case studies based on real-world traffic measurements we show the effectiveness of the approach.

KEYWORDS
Visual analytics, traffic engineering, anomaly detection, visual data mining, temporal data mining

1. INTRODUCTION

Congested traffic on motorways causes both, a collective economic damage to the society, and a personal damage to individual drivers. In the field of traffic engineering there are two approaches to cope with this challenge:

1. Online-analysis, referring to the identification and prediction of congested traffic. Findings lead to short-term actions like dynamically changing road signs, which allows drivers to adapt to congested situations or to avoid them.

2. Offline-analysis, referring to the identification of infrastructure problems in the road network, which result in bottlenecks causing congested traffic. Findings lead to long-term actions like optimizations of the road network.

In this paper we focus on the latter one, the offline-analysis of traffic data. We first introduce the very sophisticated, but application-specific, models currently used in online-analysis applications and then discuss how essential questions about the traffic measurements can be answered using an application-independent visual analytics approach proposed in this work during offline-analysis.
State-of-the-art in traffic engineering

Based on the measurement of raw traffic data, modern traffic engineering uses traffic flow theories and traffic flow models which allow for the reconstruction of traffic patterns. One of the most well-known and successful theories used in online installations is Kerner's three-phase traffic theory and the accompanying models ASDA/FOTO based on this theory [7][8]. The models run in online operation mode and supply outputs used e.g. for dynamic route guidance on motorways. This scenario is in productive use on German motorways in the federal state of Hessen and in the free state of Bavaria.

Figure 1 shows two spatio-temporal traffic patterns, reconstructed from traffic measurements on a motorway section as we have shown in [10]. A red area corresponds to wide moving jam and a yellow area to synchronized flow [7] at the given point in time and motorway position.

Motivation

The models could also be used in the offline-analysis of traffic measurements. Yet, traffic models require computer processing time during their execution and configuration prior to their application like the number of lanes, the detectors’ positions, data rates and configuration of the road network. A visual analytics approach is more appropriate for the offline-analysis of traffic measurements in order to (1) evaluate the road network and (2) evaluate the detector network:

1. Evaluation of detector network: As any statement made about the road network relies on the data recorded by the detector network, an evaluation of the detector network, like e.g. querying for non-reliable detectors, is relevant for road authorities.
2. Evaluation of road network: Answering questions like where and when traffic jams occurred is essential. From the perspective of data mining we refer to traffic jams as abnormal traffic situations, i.e. anomalies.

Data analysis without the application-specific models is especially beneficial in environments, where it is not practicable to apply complex models from an economical point of view.

Our aim is to show that a visual analytics approach, as proposed in this work, can be used for the offline-analysis of traffic data. We adapted and enhanced parallel coordinates [5] to cope with the spatio-temporal structure of the traffic measurements, allowing to interactively explore the data and thereby detect anomalies. We refer to this as interactive anomaly detection.

2. INTERACTIVE ANOMALY DETECTION

The discovery of anomalies in multivariate time series data resulting from local traffic measurements of a motorway section is the aim of this work. From a general perspective, this can be done (1) in an automatic manner by applying data mining algorithms [4], (2) using application-specific models, or (3) in an interactive, visual way by integrating the user into the process. We use the latter one and refer to it as interactive anomaly detection.
In [11], Shneiderman introduced the term information-seeking mantra, which could be described as the steps to be taken to discover knowledge [1] from visualised data in general. Keim related this principle to visual data exploration in [6], where the exploration process takes place in an interactive way. Numerous different visualisation techniques are known. In the survey paper [2], the evolution from visual data exploration to visual data mining is discussed. In [13] the more recent term visual analytics was introduced, which is defined as the science of analytical reasoning facilitated by interactive visual interfaces.

There is a variety of visualisation techniques, but the visualisation of multivariate time series data puts special requirements on the visualisation technique. Many of the current visualisation techniques were not especially developed to be applied on time series data or use only a few time-dependent variables. For interactive anomaly detection in multivariate time series, enhanced visualisation techniques are required.

An enhancement of parallel coordinates [5] is proposed to cope with the time series data resulting from traffic measurements. Using parallel coordinates, relating n time series of equal length \( d \) is done by drawing \( n+1 \) parallel axes: \( n \) axes for each of the time series values and one axis for the vector of timestamps. Each time series’ value is mapped to the position on its axis, starting with low values at the bottom of the screen. The values on the axes are then connected by line segments which results in \( d \) lines. Information about the location of detectors is expressed by the arrangement of the axes.

User-driven highlighting of a value range, referred to as brushing [3], allows to focus on portions of the data set. Individual brushing operations work on selected axes. In this work, the enhancement was made that consecutive brushing operations can be linked by Boolean operators. The Boolean operators are applied to the set of currently brushed data items and the data items contained in the current brushing operation. The history of sequential brushing operations can be visualised. This way, sophisticated queries can be formulated on a time series data base, as we have shown in [12]. Brushed or unbrushed data items can be removed in order to filter the data set. The temporal information is added using additional axes, which allows deducing the timestamps of brushed data items. In addition, as opposed to visual analytics tools for the analysis of unstructured data, working on time series allows to offer functions that are defined for time series: differentiation, integration, point-wise arithmetic like addition or subtraction and time series distance measures [9] are integrated.

3. CASE STUDIES

The case studies in this section are based on real-world traffic data. Analysing the reliability of detectors and detecting abnormal traffic situations will be shown based on traffic measurements recorded by stationary loop detectors. The data was recorded between November 2009 and January 2010 by 33 detectors positioned non-equidistantly at a 40 km long section of the motorway A5 between Butzbach and Westkreuz Frankfurt intersection in Germany. The detectors measure the velocities and quantities of passing vehicles in one direction of the motorway 24 hours a day with a sample rate of one minute. The authors would like to thank the traffic control center of Hessen, Germany for supplying the data set.

Mapping space and time to parallel coordinates

The mapping of the spatio-temporal data to the parallel coordinates plot will first be shown in an introductory example. It lies in the nature of interactive visualisation approaches, that they rely on the following two basic conditions: (1) the bigger the available space for visualisation, the more information can be displayed and (2) the approaches offer the best benefit, when integrating the user into the exploration process, i.e. having the user interact with the graphical user interface. Both conditions can obviously not be fully met in a printed publication. In order to overcome the space restrictions, an example on a reduced data set is shown, allowing for the identification of individual axes and values. The missing interactivity is addressed by specifying the consecutive user operations (e.g. queries) in the case studies, where appropriate.

In Figure 2 a small subset of the data is visualised using the enhanced parallel coordinates. The velocity values of only three of the 33 detectors are visualised over a period of one hour. Additional information was added in dark blue indicating how (1) the time, (2) the location and (3) the vehicle speed are illustrated in the plot. The timestamps (8.00 am - 8.59 am) of one pre-selected day are mapped to the first vertical axis in the exemplary plot in Figure 2. Each of the remaining vertical axes shows the velocities measured by one
detector during this period. The location of the individual detectors is shown in horizontal direction. The axes holding the velocities are ordered in a way, that a vehicle would pass the detectors from the most left axis to the most right axis. Each of the axes is labelled according to the detector's position, so for example the axis labelled '081.9_v' shows the velocities measured by the detector at motorway position 81.9 km.

Each of the lines connecting the axes from left to right corresponds to the vehicles' velocities at one point in time. The three axes holding the velocities are aligned to have a common minimum and maximum value, i.e. one horizontal line connecting the axes would correspond to equal velocities at all detectors. In order to be able to distinguish the temporal and spatial information, the detector axes are indicated by a black square at the bottom of the axis. Axes without a black square hold the temporal information. Due to the fact that the detectors output one velocity value per minute and one hour is contemplated, there are 60 lines connecting the axes in this example.

![Figure 2: Mapping of spatio-temporal traffic measurement to parallel coordinates showing one hour and three detectors](image)

One of the major benefits of the parallel coordinates plot is the possibility to easily formulate queries by highlighting selected data items. In Figure 2, the data set was queried to show velocities below 30 km/h at the detector at 81.9 km, indicated by the rectangle on the bottom of the second axis. The brushed data items are coloured in green. Following the green lines, two statements can be deduced in this example plot: (1) at which points in time the low velocities occurred and (2) what the velocities were at that time at the remaining two detector positions. In all further plots, the brushing operations will be indicated by red rectangles.

In order to map the full data set of traffic measurements to the parallel coordinates plot, the time span between 6 am - 10 pm was chosen to be visualised, corresponding to approximately 2.8 million data points. In addition, as opposed to plotting the raw time stamps, the temporal information was pre-processed, allowing for more goal-oriented queries. Mapping of the data to the vertical axes is done as follows: (1) index, (2) year and month, (3) day, (4) time of day, (5) weekday, (6..39) one axis for each of the 33 detectors in the range of 81.9 - 120.4 km. The axes' positions from left to right correspond to the detectors' positions on the motorway in ascending order.

**Evaluating the detector network**

Following the first question formulated in the motivation section, the evaluation of the underlying detector network is essential. In a pre-processing step, invalid or unavailable detector values were replaced with an error value of -10 km/h, the error value was chosen arbitrarily. Selecting the erroneous values of the fourth detector positioned at 93.7 km reveals a correlation between erroneous or missing values between detectors 4 ... 33 as shown in Figure 3a. It is highly unlikely that the problem in this case is due to erroneous detectors because of the coinstantaneous occurrence. Road authorities would use this information to track down the error cause.

A second example shown in Figure 3b is the removal of one unreliable detector: A high number of erroneous values become obvious at the axis of the detector at position 117 km. The erroneous detector values were highlighted using the brushing operation. The number of erroneous values can be obtained this way: 3.9% or an overall of 37 hours over the contemplated period of three months. As can be seen from the plot, other detectors show abnormal values as well, but to a much lesser extent. The unreliable detector was removed from the data set prior to further analyses, resulting in 32 remaining detectors.

As shown by these two examples, visual analytics can be used to evaluate the reliability of detectors without applying any complex traffic model.
Detection of abnormal traffic situations

We refer to traffic jams as abnormal traffic situations, i.e. anomalies [14]. Interactive querying for those types of anomalies allows to determine where and when they were observed. This is the first step to deduce statements about why traffic jams occur at specific subsections. Hence, the capability to browse the data for specific traffic situations is beneficial, which is supported by the described brushing operations. Thereby queries for traffic jams at specific sections can easily be formulated. In Figure 4a all traffic jams between positions 95.4 km and 98.8 km over the period of three months were highlighted by incrementally querying. This allows first, to evaluate their frequency and second, to obtain their points in time.

Contemplating the axis holding the weekday in Figure 4a reveals the following: a small portion of the highlighted traffic jams occurred on a Saturday, which is unexpected. Enhancing the query in order to only highlight the subset of jams that occurred on a Saturday (AND: weekday = Saturday), results in Figure 4b. From the plot two occurrences can be detected, both on Saturday 30th January 2010 between (1) 3.03 pm and 4.44 pm and (2) at 9.57 pm. The second occurrence was only one minute long and is therefore viewed as an irrelevant outlier. Traffic jams on a Saturday are atypical on the contemplated motorway section. Research on the internet showed that on the 30th January 2010, there were dramatic road conditions in Germany due to snow and ice.

Figure 3: (a) Correlation of erroneous detector values (left) (b) Identification of one unreliable detector

Figure 4: (a) Interactive query for traffic jams (left). (b) Refined query for abnormal traffic situation (right)

Figure 5: Reconstruction of the traffic situation identified in parallel coordinates using the models ASDA and FOTO
As a test of plausibility, Figure 5 shows the same situation reconstructed using the models ASDA/FOTO with the query results from the parallel coordinates marked in the plot of the reconstructed traffic pattern. It should be noted, that Figure 5 shows a time span of six hours (distance plotted w.r.t. time) while the parallel coordinates in Figure 4 shows the measured data of three months.

The anomalies were detected by incrementally refining queries, which is an unsupervised, user-driven process. As opposed to the relatively easy detection of the erroneous detector values in Figure 3a, the detection of this abnormal jam is more complex, as it integrates several variables. The detection of anomalies in the traffic measurements can have several goals: either to investigate atypical traffic behaviour in a goal-oriented manner or to remove the anomalies in order not to consider them in the deduction of global statements about the traffic behaviour.

4. CONCLUSION

In this paper we first introduced a state-of-the-art approach for the online-analysis of traffic measurements. We stated that this model-based approach could in theory be used for the offline-analysis of traffic measurements as well, but due to the required effort for configuration and their computational effort, we propose to use visual analytics approaches instead. We focused on the interactive detection of anomalies in long-term traffic measurements and showed how an adaption of parallel coordinates can be used to (1) visualise the spatio-temporal information present in the traffic measurements and (2) to interactively query long-term traffic measurement, allowing for the detection of anomalies in the traffic flow as well as in the underlying detector network.

An approach of this type can be used by road authorities during optimization of existing road networks. Tailoring the visual analytics approach to the problem-domain of traffic engineering would make the approach even more powerful. So for example visually integrating the road network together with the detector positions would be beneficial. We found that in the field of traffic engineering, model-based approaches are used and we state that the use of visual analytics as proposed in this paper is helpful to answer essential questions based on long-term traffic measurements. A combination of a model-based and a visual analytics approach, e.g. mining the models’ outputs, could be one direction for further research.
REFERENCES