

Topographically-Based Real-Time Traffic Anomaly Detection in a Metropolitan Highway System*

Rajmonda Sulo[†]
University of Illinois at Chicago

Anushka Anand[‡]
University of Illinois at Chicago

Leland Wilkinson[§]
SPSS Inc.

Robert Grossman[¶]
University of Illinois at Chicago

Stephen Eick^{||}
University of Illinois at Chicago

ABSTRACT

In the wake of recent terrorist threats, where major highways could be the target of attack, the need for real-time traffic management is crucial. Real-time traffic management encompasses many aspects of highway traffic, such as the analysis of congestion levels, incident detection and classification, traffic forecasting, and visualization of all the above.

The work presented in this paper focuses on real-time detection and visualization of unusual changes in traffic in the Chicago metropolitan area. Our Chicago Alert System (CAS) considers both the spatial and temporal aspects of the data by clustering sensors that report similar traffic flows and by building baselines that capture the seasonality and variation of data over the period of a year. Outlier values of the traffic flow are then detected using the baseline models. Real-time alerts are visually displayed through an online Web Service. We discuss analytic refinements of the system, including continuous updating of baselines and modeling to include weather, holidays, and other exogenous variables.

CR Categories: H.5.2 [User Interfaces]: Graphical User Interfaces—Visualization; I.3.6 [Computing Methodologies]: Computer Graphics—Methodology and Techniques;

Keywords: visualization, statistical graphics

1 INTRODUCTION

Francis Bacon wrote in his *Novum Organum* about 400 years ago:

Errors of Nature, Sports and Monsters correct the understanding in regard to ordinary things, and reveal general forms. For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways.

Visual analytics offers a new opportunity to test Bacon's observation. By incorporating appropriate statistical models into data viewers, we can more easily recognize both regularities and deviations. The severest challenge of visual analytics is to apply these models and views to massive sources of data. An example is high-volume highway traffic in a large metropolitan area, How do we

display critical aspects of highway traffic data in order to recognize and respond intelligently to sudden or unexpected changes in traffic patterns? Our problem is not simply to recognize accidents shortly after they occur, but also to flag and analyze gradual changes in behavior that signal other abnormal processes. For this kind of sensitivity, we need to ground our visualizations on appropriately developed statistical models.

There are several interesting aspects to this project. First, in contrast to other traffic studies that focus on congestion, our focus is on change and anomaly detection. The reason for this is that congestion is widely studied and easy to detect. The more interesting aspect and open research problem is to detect sudden *and* gradual changes in sensor behavior.

Second, our approach to change detection is nonparametric. We attempt to minimize the distributional assumptions required for conventional parametric models. Furthermore, our methods are highly adaptive to local data features. We smooth the data but we do not oversmooth with more inflexible parametric models.

Third, our analysis and display system is grounded on the *topological* aspects of the transportation system. Other researchers have developed network flow models, spatial models, temporal models, and other models that ignore specific aspects of road topology. Our system focuses, by contrast, on the similarities in topography (terrain, pavement, lane width, merging) that lead to similar traffic behavior. We do not model these features explicitly, but instead expose them through the adaptive clustering and smoothing procedures we have developed.

Fourth, there is a significant visual analytics focus to this effort. We have been able to summarize a massive number of sensor observations (approximately 73 million) over a geographically dispersed area in a coherent and concise tableau of displays on a single screen in a way that captures the complexity of the system without confusing the viewer.

Fifth, we handle missing values in the sensor stream appropriately. Our approach imputes missing values from patterns of observed values and allows us to develop baselines despite sensor malfunctions and other system errors.

2 RELATED WORK

Traffic modeling and incident detection are not new research areas. The literature is extensive and our review covers only some of the more important algorithms.

2.1 Comparative Algorithms (California Algorithms)

These algorithms compare current traffic condition to a predetermined thresholds [10]. The California algorithms use the simple fact that when an incident happens at a particular sensor, it will increase occupancy levels upstream and decrease occupancy levels downstream. These algorithms compute the relative difference in occupancy values and throw an alarm if the difference is above a threshold. A drawback of these algorithms is that the thresholds

*We thank the UIC Urban Transportation Center for guidance in transportation-related issues

[†]e-mail:rsulo1@uic.edu

[‡]e-mail:aanand2@uic.edu

[§]e-mail:leland@spss.com

[¶]e-mail:grossman@cs.uic.edu

^{||}e-mail:eick@cs.uic.edu

used to detect alerts are not updated automatically and therefore they don't always reflect the behavior of the current traffic flow.

2.2 Statistical Algorithms

Statistical algorithms are another important group of algorithms used to model the flow of traffic and occurrence of incidents. The main statistical algorithms are the Standard Normal Deviation algorithm (SND) [3], the Bayesian algorithm [9] and time series algorithms (ARIMA and low-pass filtering). The SND algorithm assumes a normal distribution of traffic incidents and throws an alert when the standard normal deviation of a new value exceeds a critical value. The Bayesian Algorithm uses past data to compute the probability that a new change in occupancy is caused by an incident. Finally, time-series algorithms provide short-term traffic forecasts based on the recent history of one of the traffic variables. ARIMA, one of the most popular time-series models, has been used to handle the variations of traffic data over time and space. Low-pass filtering models smooth the data to minimize the effect of sharp frequency data [13]. The main drawback of these particular statistical models is that they are not highly adaptive to empirical data densities. Models such as ARIMA are sometimes appropriate for seasonal economic data but are likely to be less appropriate for the fairly irregular series found in traffic data.

2.3 Artificial-Intelligence Algorithms

These algorithms use neural networks or other AI models [2] [8] to implement a trainable system. These algorithms often have exponential complexity and cannot easily be parameterized for associated statistical analysis.

2.4 Catastrophe Theory Algorithms

These algorithms focus on catastrophic behavior of individual sensors. The McMaster algorithm uses catastrophe theory [12] [11] to model a sudden change in sensor output. The main idea behind this algorithm is based on the fact that while the speed values change drastically in the event of a traffic incident, volume and occupancy change gradually. This algorithm has good accuracy but doesn't take into account the spatial-temporal nature of traffic data.

3 THE CHICAGO HIGHWAY SYSTEM

Chicago's highways were among the first in the country to be equipped with traffic sensors (in 1961). These sensors consist of single inductive loops buried in the roadway pavement less than one mile apart. Each sensor is assigned a unique signal frequency and the information is transferred for further processing through telephone lines to a main location in Oak Park, Illinois.

The inductive loops work by detecting changes in inductance. The sensor loops turn on and off as cars pass over them. The number of on signals within a time interval (usually 30 seconds) determines one of the most commonly used traffic variables, called *occupancy*. Occupancy is defined as the percentage of time a point on the road is occupied by vehicles [6]. Another variable measured is *volume*, defined as the number of vehicles flowing past a point during a time interval. A third variable, *speed*, is used for traffic flow analysis but is not directly measured by this equipment. Instead, speed is calculated from measurements of occupancy and volume using a formula first introduced by [1]:

$$s = cv/o$$

where c is a constant proportional to the average length of a car, v is volume, and o is occupancy. More recent detectors, such as double loop detectors, offer direct measurements of speed.

The Chicago highway sensor data are collected by a Gateway System that covers the three state, fifteen county Gary-Chicago-Milwaukee (GCM) corridor. The Gateway System uses 830 fixed traffic sensors, in addition to other data sources, to compute real-time traffic congestion conditions and to display these data to the public at two websites (<http://www.gcmtravel.com> and <http://www.travelinfo.org>). The data for the baselines computed in this paper were archived in the Gateway Archive Testbed at the National Center for Data Mining (NCDM).

4 THE STATISTICAL MODEL

Figure 1 shows a five-day time series from one of the 830 Chicago traffic sensors. The vertical scale on each panel measures occupancy from zero to 60 percent. The horizontal scale runs from midnight Friday to midnight Friday. This scale allows us to view weekend traffic on the left end of the graph and weekly commuter traffic toward the right.

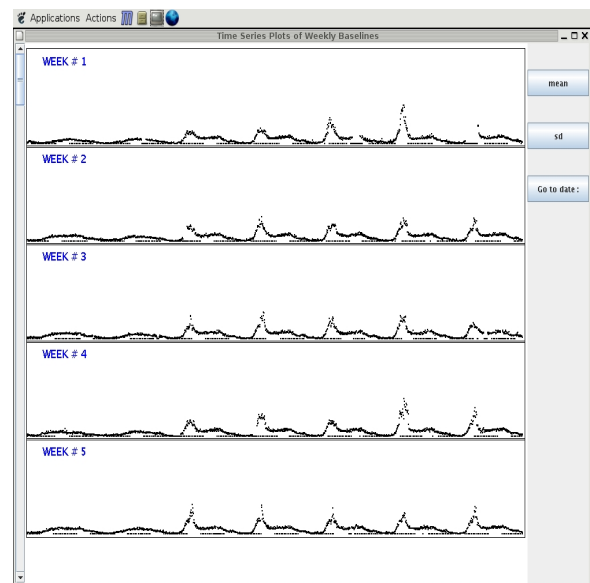


Figure 1: Single sensor reading plotted over five weeks of series

The annual series for this sensor consists of measurements at 6 minute intervals through the year, comprising altogether 87,600 time points ($10 \times 24 \times 365$). The full dataset contains 830 of these sensor series, so there are almost 73 million datapoints in the baseline dataset.

The series in Figure 1 reveals a pattern found in almost all the series. Namely, the daily commuter periods show a higher occupancy rate (and, presumably, lower speeds). Weekend occupancy rates are, of course, lower.

Our Chicago Alert System (CAS) requires a baseline model for representing normal traffic flow together with an error distribution for determining alerts. After closely examining the sensor series for a variety of sensors, we decided that simple parametric time series models were a poor choice for representing the data. These models have some difficulty dealing with missing values (malfunctioning sensors), irregular seasonal variation (holidays and weather), and outliers (late night 160 mph flights of an exotic car along a stretch of suburban highway – which we observed in several series). Consequently, we decided to construct *empirical* baseline time series.

We began the CAS by grouping sensors geographically. In a pilot project [5], we developed a collection of baseline models. This application used over 42,000 separate baseline models one for each

hour, one for each day, and one for approximately every 3 sensors. This approach was unnecessarily unwieldy and impeded the performance of the real-time system, but it proved the feasibility of a real-time monitoring system on this data scale.

In order to reduce the pilot system’s storage and response times, we decided to cluster sensor time series without regard to geography. The motivation for this approach assumes that *topographically* similar sensors (sensors located near merging lanes, for example) are more likely to share traffic characteristics than *geographically* similar sensors (sensors near each other but located on a topographically heterogeneous highway segment, for example),

The topographical approach began with the 830 by 87,600 matrix of sensor data over one calendar year. We used *k*-means clustering [7] to aggregate the time series across sensors. The multipass algorithm we used provides better estimates than single-pass *k*-means algorithms because it allows iterative refinement of the centroids and it derives initial estimates of cluster centroids from preliminary passes through the data. We used *normalized* Euclidean distance (Euclidean distance over non-missing data points divided by the number of non-missing points) in order to adapt to the presence of missing data. Since all measurements were on the same percentage scale and in a roughly similar range, we did not pre-normalize the data.

Sequential chi-square tests on reduction in residual sum-of-squares indicated we needed only 4 clusters to fit the data reasonably well. We thus reduced the 830 series to 4 time-series clusters. Figure 2 shows the standard deviations for the clusters over 5 weeks of the time series. Note that the standard deviations tend to be larger for the higher occupancy values. This led us to suspect that the distribution of within-cluster raw scores is not normal. In fact, we should expect in general that distributions on percentages tend to be positively skewed for mean values near zero and negatively skewed for mean values near 100 [4].

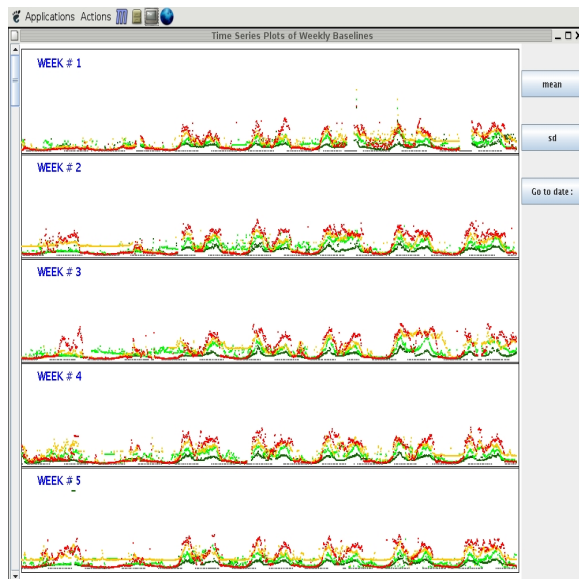


Figure 2: Cluster standard deviations plotted over five weeks of series

Figure 3 shows histograms for a sample of values from each cluster. The histograms validate our suspicion. Cluster1 has the smallest mean and variance. Cluster4 has the largest mean and variance.

To adjust for this heterogeneity of variance, we applied the arcsine transformation:

$$t(p) = \sin^{-1} \sqrt{p}$$

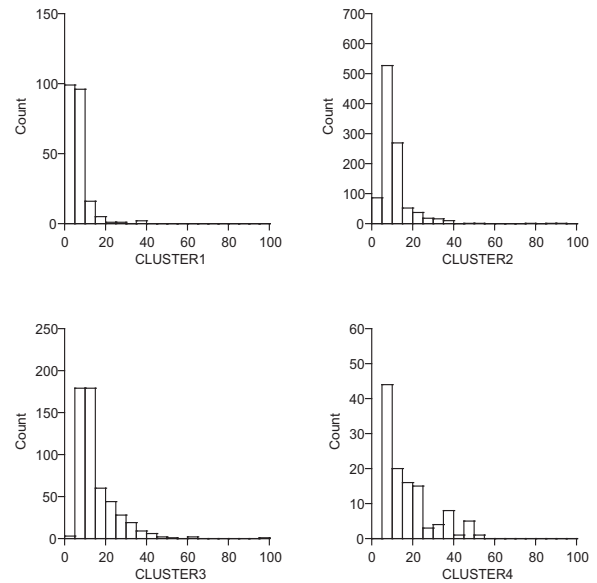


Figure 3: Histograms of sampled within-cluster occupancy values

to the occupancy values, where *p* in the formula is occupancy divided by 100.

Figure 4 shows the transformed histograms. The transformation has reduced the skewness seen in Figure 3.

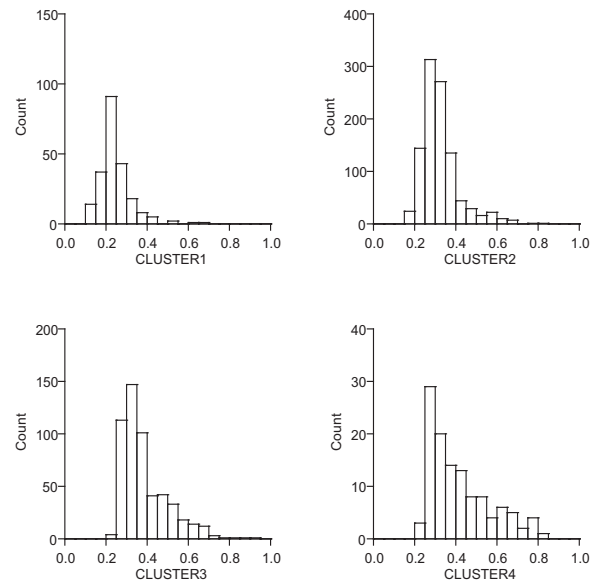


Figure 4: Histograms of transformed within-cluster occupancy values

Our final cluster model was computed on the arcsine transformed values. Figure 5 shows the result of this analysis. The clustered series reveal more clearly the periodicity in weekday commuting patterns. Morning and evening commute times are clearly visible in each series. In addition, we see the highest occupancy series (colored red) at the top of each plot. The sensors belonging to this cluster tend to output the slowest highway traffic speeds in the metropolitan region.

Figure 6 shows the location of the sensors on the Chicago area highway map. The sensor points are colored according to cluster membership. There are several interesting features to note in this

choosing a specific date instead of scrolling Third, it is possible to examine raw data series using the same display format found in the cluster series plots.

6 DISCUSSION

There are several improvements we plan for our system. First, we have not included two exogenous variables that can affect the performance of our model. One of these variables is the presence of holidays. While our baselines include holiday behavior (because of holiday traffic in the real series), these periods are not synchronized for subsequent years. We plan to condition our baselines on holiday effects and then use the estimated parameters in later-year adjustments.

The other critical variable is precipitation. Although real-time and near real-time weather data is widely available, the data are collected at specific locations, such as airports. In Chicago, for example, weather data are readily available for the Chicago and Midway airports but not for other sites.

Unfortunately, rain and snow are often local (especially in Chicago) and vary widely over the sensor region. For this reason, we will be using real-time Doppler radar data to extract local precipitation information from the corresponding pixel values. As with holidays, we will condition our baselines on precipitation levels and use current precipitation to derive a posterior estimate of expected traffic.

Second, we plan to animate our system to allow retrospective evaluation of change. In our pilot work [5], we explored some of these techniques and found that animation can be used to understand processes leading up to even changes. Because we have reduced a rather large data source to a manageable baseline model, the response time of our system to event changes is in milliseconds.

7 CONCLUSION

We have demonstrated the feasibility of real-time visual anomaly detection over a relatively large system (over 73 million data points). Because we incorporate spatial and temporal components in the system and because we display results spatially and temporally, we facilitate visual analysis in the visual language best understood by professional responders.

While we have not done this type of visual modeling in other environments, we are confident that the procedures developed here could be applied to other transportation data and to network data where topography plays a role. Developing these systems must begin with careful visual inspection and analysis of the raw data. To build appropriate models for visual analytic platforms, we must pay attention to the structure of the data and avoid the temptation to apply canned data mining procedures.

In the end, visual analytics is about *guided* exploration. Our visual exploratory tools must be grounded in appropriate and adaptive models that can help us navigate through enormous fields of sensor data. As Bacon noted, to know the ways of Nature is to notice her deviations. To recognize the deviations, we need a model of Nature.

REFERENCES

- [1] P. Athol. Interdependence of certain operational characteristics within a moving traffic stream. Technical report, Transportation Research Record 72, TRB, National Research Council, Washington, D.C., 1967.
- [2] R.L. Cheu, S.G. Ritchie, W.W. Recker, and B. Bavarian. Investigation of a neural network model for freeway incident detection. In B.H.V Topping (Ed.), editor, *Artificial Intelligence and Civil Structural Engineering*, pages 267–274. Civil-Comp Press, 1991.
- [3] C.L. Dudek, C.J. Messer, and N.B. Nuckles. Incident detection on urban freeways. Technical report, Transportation Research Record 495, TRB, National Research Council, Washington, D.C., 1974.
- [4] J.L. Fleiss. *Statistical Methods for Rates and Proportions*, (2nd Ed.). John Wiley & Sons, 1981.
- [5] Robert L. Grossman, Michal Sabala, Anushka Aanand, Steve Eick, Leland Wilkinson, Pei Zhang, John Chaves, Steve Vejchik, John Dillenburger, Peter Nelson, Doug Rorem, Javid Alimohideen, Jason Leigh, Mike Papka, and Rick Stevens. Real time change detection and alerts from highway traffic data. In *ACM/IEEE SC 2005 Conference (SC '05)*, 2005.
- [6] F. L. Hall. Traffic stream characteristics. In N.H.Gartner, C.J.Messer, and A.K. Rathi, editors, *Traffic Flow Theory*, pages 21–34. US Federal Highway Administration, Washington, D.C., 1996.
- [7] J. A. Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.
- [8] J. Ivan. *Real-time data fusion for arterial street incident detection using neural networks*. PhD thesis, Northwestern University, 1994.
- [9] M. Levin and G.M. Krause. Incident detection: A bayesian approach. Technical report, Transportation Research Record 495, TRB, National Research Council, Washington, D.C., 1978.
- [10] H. Payne and S. Tignor. Freeway incident detection algorithms based on decision trees and states. Technical report, Transportation Research Record 682, TRB, Washington, D.C., 1976.
- [11] B.N. Persaud and F.L. Hall. Catastrophe theory and patterns in 30 second freeway traffic data implication for incident detection. *Transportation Research*, 23A:103–113, 1989.
- [12] B.N. Persaud, F.L. Hall, and L.M. Hall. Congestion identification aspects of the mcmaster incident detection algorithm. *Transportation Research Record*, pages 167–175, 1987.
- [13] Y.J. Stephanedes and A.P. Chassiakos. A low pass for incident detection. In *Applications of Advanced Technologies in Transportation Engineering, Proc. Second International Conference*, pages 378–382, Minneapolis, 1991.