Experiences from Supporting Predictive Analytics of Vehicle Traffic

Natalia Andrienko, Gennady Andrienko, and Salvatore Rinzivillo

Abstract— By applying visual analytics techniques to vehicle traffic data, we found a way to visualize and study the relationships between the traffic intensity and movement speed on links of a spatially abstracted transportation network. We observed that the traffic intensities and speeds in an abstracted network are interrelated in the same way as they are at the level of road segments. We developed interactive visual interfaces that support representing these interdependencies by mathematical models, which can be then utilized for forecasting not only the expectable normal traffic situation for a given moment and its development over time but also how the normal conditions may change due to extraordinary mass movements caused by public events or emergencies. We developed further interactive visual tools to support the use of data-derived models for predictive traffic simulation on the basis of an abstracted network. We came to a general conclusion that visualization support to predictive analytics, which consists of three successive tasks (analyze data – develop models – obtain and analyze forecasts), may need to be developed in an evolutionary way. This applies to cases when all tasks cannot be fully defined in advance, but clear definitions for later tasks emerge depending on the results of the preceding ones.

Index Terms— H.2.8.c Data and knowledge visualization, H.2.8.h Interactive data exploration and discovery, I.6.5 Model Development, I.6.4 Model Validation and Analysis, I.6.7 Simulation Support Systems

Introduction

Data concerning vehicle traffic in transportation networks are now collected in great amounts owing to advances in sensing technologies. These data offer new opportunities for improving the understanding of traffic properties and enhancing the accuracy of the models describing and forecasting traffic situations and their evolution. However, the potential of real traffic data remains largely underexploited. By means of visual analytics methods, we performed a systematic study of the opportunities hidden in historical traffic data. We found out that traffic data covering a sufficiently long time period to capture the regular daily and weekly variations allow deriving formal models, which can be utilized for predicting not only regular traffic flows at different times but also extraordinary flow in abnormal situations, such as road closures or mass movements caused by public events or emergencies. Predicting unusual traffic behaviors on the basis of historical data reflecting only normal patterns becomes possible due to reconstruction of interdependencies between the traffic intensity (a.k.a. traffic flow or flux) and the mean movement speed for different links of the transportation network.

We developed visual analytics tools that support building of formal models capturing the speed – intensity relationships from historical data characterizing network-constrained movement of physical objects, such as vehicles. Furthermore, we developed visual analytics tools that support utilizing the derived models for simulation of traffic under various conditions and prediction of normal and abnormal traffic behaviors.

A distinctive feature of our approach to traffic analysis, modeling, and simulation is the use of data abstraction and generalization for modeling transportation networks and traffic properties at different levels of spatial scale.

1 Abstraction and Generalization

Traffic data may be available in the form of trajectories of moving objects. A trajectory consists of records reporting the positions (e.g., geographic coordinates) of moving objects at different times. Given a large set of trajectories of objects, we apply an existing method [1] that derives an abstracted network consisting of cells (territory compartments) and links between them. Smaller or larger cells can be generated by varying method parameters, thus allowing traffic analysis and modeling at a chosen spatial scale. Fig. 1 gives an example of an abstracted traffic network for Milan (Italy) reconstructed from GPS tracks of 17,241 cars collected over a period of one week from Sunday, April 1, to Saturday, April 7, 2007.

The nodes of an abstracted traffic network are polygonal cells. Neighboring cells are connected by pairs of directed links. After constructing a network, the original trajectory data are aggregated spatially by the nodes and links of the network and temporally by time intervals. In our studies, we aggregated vehicle trajectories by hourly intervals. The result of the aggregation includes, among others, two sets of time series for the links: traffic intensities and mean speeds. Traffic intensity on a link, also called traffic flow or flux, is the number of objects traversing the link per time unit.

For the further analysis, we apply data generalization: we cluster the links by similarity of the time series of traffic intensity and speed (Fig. 2) and then derive general models for the clusters instead of trying to consider each link separately. This approach not only reduces the workload but also precludes model overfitting.

2 Deriving Models for Link Clusters

For model derivation, we apply a methodology [2] in which an interactive visual interface to a modeling library is utilized to build a model of the temporal variation of the traffic intensity for each cluster of links. Triple exponential smoothing (Holt-Winters method) is used to model the periodic variation according to the daily and weekly cycles.

To study and quantify the relationships between the traffic intensity and the mean speed, the time series of the intensity and speed values are transformed in the following way. Let A and B be two time-dependent attributes associated with the same object (in particular, link) and defined for the same time steps.

1. Divide the value range of attribute A into intervals.
2. For each value interval of A:
   a. Find all time steps in which the values of A fit in this interval.
   b. Collect all values of B occurring in these time steps.
   c. From the collected values of B, find the minimum, maximum, mean, quartiles, and the 9th decile (i.e., 90th percentile).

- *Natalia Andrienko and Gennady Andrienko are with Fraunhofer Institute IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany, and with City University London, UK.*
  E-mail: {natalia|gennady}.andrienko@iais.fraunhofer.de.
- *Salvatore Rinzivillo is with the Knowledge Discovery Laboratory, Istituto di Scienza e Tecnologie dell'Informazione, Area della Ricerca CNR, via G. Moruzzi 1, 56124 Pisa, Italy. E-mail: rinzivillo@isti.cnr.it.*

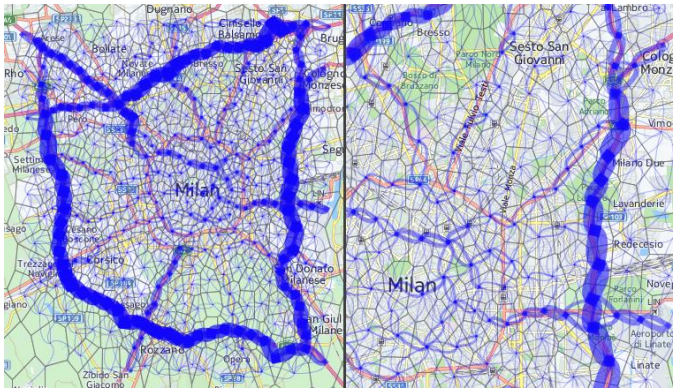

Fig. 1. An abstraction of the street network of Milan, Italy.

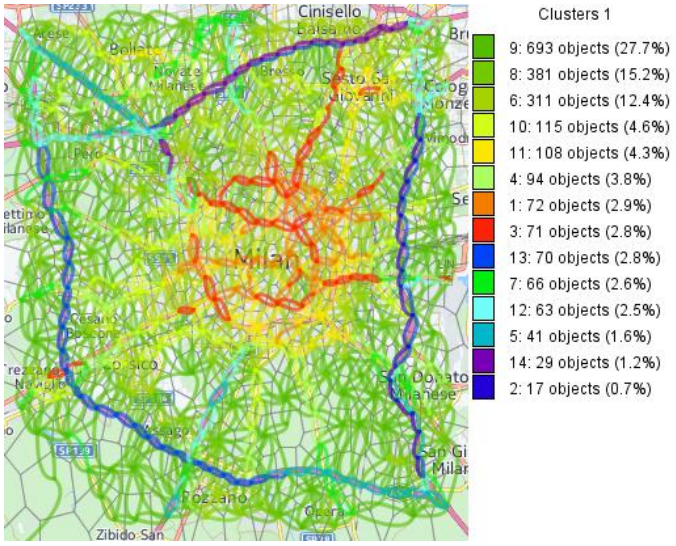

Fig. 2. Links of the abstracted network are clustered by the similarity of the variation of the traffic intensity and mean movement speed.

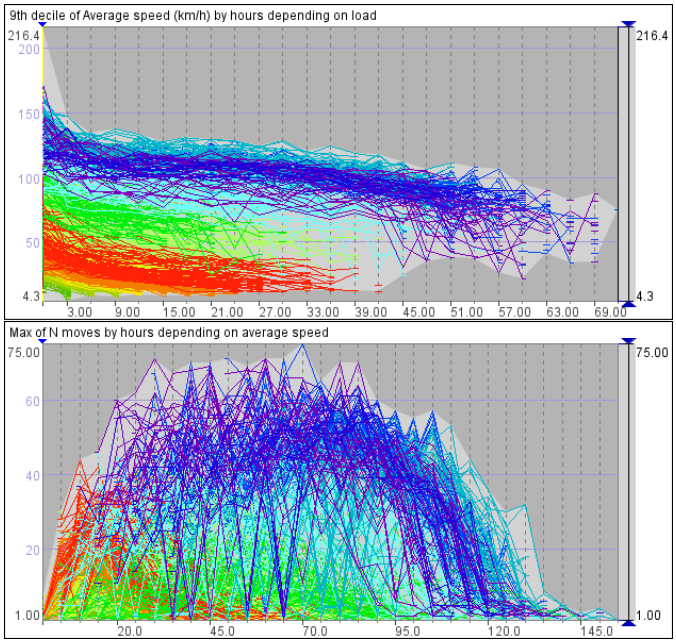

Fig. 3. The dependency series computed for the links of the abstracted Milan traffic network are represented graphically. Top: 9th decile of the mean speed depending on the traffic intensity. Bottom: Maximal traffic intensity depending on the mean speed.

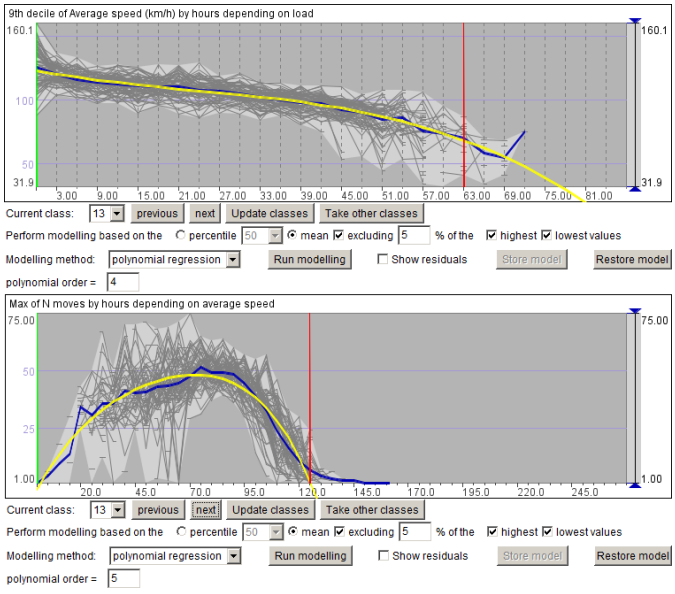

Fig. 4. Polynomial regression models represent the interdependencies between the traffic intensity and the mean speed.

3. For each statistical measure (i.e., minimum, maximum, etc.), construct a sequence of values of B corresponding to the value intervals of A.

In this way, a family of attributes is derived: minimum of B, mean of B, and so on. For each of the derived attributes, there is a sequence of values corresponding to the chosen value intervals of attribute A. This sequence is similar to a time series except that the steps are based not on time but on values of attribute A. We call such sequences dependency series (DS) since they express the dependency between attributes A and B. Attribute A is treated as the independent variable and B as the dependent variable.

To model the interdependencies between the mean speed and the traffic intensity, we perform two transformations. First, we treat the traffic intensity as the independent variable and derive a family of attributes expressing the dependency of the mean speed on the traffic intensity. Second, we treat the mean speed as the independent variable and derive a family of attributes expressing the dependency of the traffic intensity on the mean speed. As an example, two graphs in Fig. 3 represent the 9th deciles of the mean speeds for different traffic intensities (top) and the maximal traffic intensity for different mean speeds. The lines in the graphs correspond to the links of the abstracted transportation network of Milan. The lines are colored according to the cluster membership of the links, as in Fig. 2.

Fig. 4 demonstrates the derivation of formal models (specifically, polynomial regression models) representing the two-way dependencies between the mean speed and traffic intensity. Analogously to models of the temporal variation, dependency models are built for link clusters. The upper and lower parts of Fig. 4 correspond to the same cluster of links from Milan. In the upper part, the dependency of the mean speed on the traffic intensity is modeled. In the lower part, a model of the dependency of the traffic intensity on the mean speed is built. The models capture the following dependencies: (1) As the traffic intensity of a link increases, the possible mean speed of movement decreases. (2) When the possible speed of movement is low, the number of vehicles capable to traverse a link in a time unit is also low. With increasing the possible speed up to a certain optimal value, the possible number of traversing vehicles per time unit also increases. However, for higher speed values, the possible number of vehicles per time unit decreases.

The shapes of the curves in Fig. 4 resemble the fundamental diagram of traffic flow [3] describing the relationship between the flow velocity and traffic flux (i.e., intensity). The fundamental traffic relationships are traditionally defined for street segments, but we

found in our research that similar relationships exist also on a higher level of spatial abstraction.

The process of modeling the two-way dependencies between the traffic intensity and mean speed is described in more detail in a recently published book [4].

3 Use of Models for Traffic Prediction

The models of the temporal variation of the traffic intensity can be used for prediction of the regular traffic for chosen time intervals in the future, assuming that the properties of the temporal variation do not change. When real traffic data are collected on a regular basis, it is reasonable to periodically check the models against the real data. If the prediction quality degrades, the models need to be updated.

In our approach, we build models for clusters of links. Each model makes a common prediction for all cluster members, but this prediction is individually adjusted for each cluster member based on the statistics of the distribution of its original values [2].

The models of the dependencies between the traffic intensity and the mean speed can be used to simulate and predict unusual traffic behaviors. The main idea is following:

1. For each link, determine how many vehicles need to move through it in the current minute.
2. Using the dependency model traffic intensity → mean speed, determine the mean speed that is possible for this link load.
3. Using the dependency model mean speed → traffic intensity, determine how many vehicles will actually be able to move through the link in this minute.
4. Promote this number of vehicles to the end place of the link and suspend the remaining vehicles in the start place of the link.

To perform a simulation, the analyst needs to define the scenario to be simulated. This includes defining a set of extra objects that will be moving in the network in addition to the regular traffic, the origins and destinations of their trips, the routes they will follow, and the time when each vehicle starts moving. The process of scenario definition is supported by a wizard guiding the analyst through the required steps and providing visual feedback at each step.

For Milan, we have performed experiments on simulating the movement of a large number of personal cars from the area around the San Siro stadium after a soccer game. To be able to simulate this scenario, we need to solve the problem of data scaling. The data that we used for model building represent not all vehicles that moved in Milan but only about 2% of the private cars. We apply the following approach. If we need to simulate movements of N private cars, we downscale this number to 2% of N, to make it compatible with the models. Figs. 5 and 6 present simulated trajectories of 250 cars, which correspond to about 12,500 cars in the real scale.

In Fig. 5, the trajectories are shown as lines in a space-time cube, which allows us to see the followed routes and the progress of the movement over time. We can spot the places where many cars will be suspended, waiting for the possibility to move. The suspensions appear in the cube as vertical trajectory segments, which mean that the spatial positions do not change as the time passes.

In Fig. 6, the trajectory lines are drawn on a map, ignoring the temporal component. In this view, the routes can be easier related to the physical street network of Milan and to the spatially abstracted network of linked cells. The red circles on the map are drawn in four cells around the San Siro stadium, which we chose as the trip origins for the simulated cars. The green circles mark the trip destinations. For choosing the destinations, we used the following reasoning. After the game, most of the spectators would drive to their home places. Hence, the probability of a cell to be a trip destination is proportional to the number of people living there. We have no data about the spatial distribution of the resident population of Milan; however, we have hourly counts of trip ends in the aggregated historical data. The number of trip ends in the evening and night hours can be expected to correlate with the number of homes in a cell, since in the evenings people typically go home. This commonsense expectation is consistent with results of empirical studies. Hence, the distribution of the trip ends in the evening and night can serve as a proxy for the resident population distribution. Based on this reasoning, we let the tool distribute the trip destinations randomly throughout the territory, so that the probability of choosing a cell is proportional to the cell weight, which is the sum of the hourly counts of trip ends in the hours from 18:00 to 24:00.

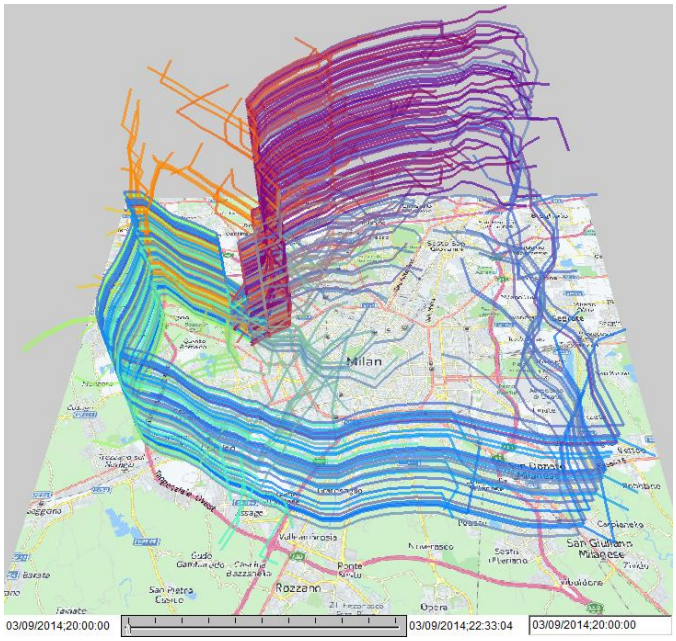

Fig. 5. Simulated trajectories of cars moving from the vicinity of the San Siro stadium to supposed home places after a soccer game are shown in a space-time cube.

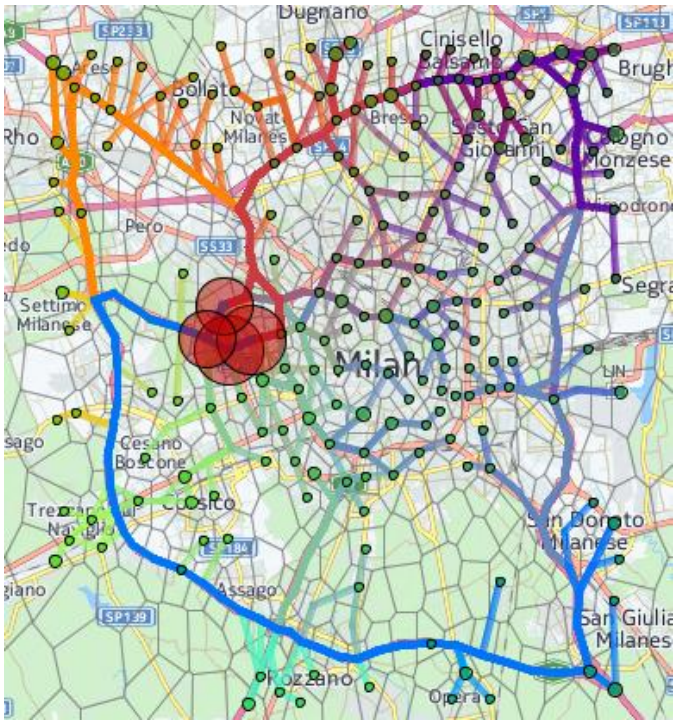

Fig. 6. The simulated trajectories are shown on a map. The red and green circles represent the trip origins and destinations, respectively.

Besides viewing the simulated trajectories in a space-time cube and on a map, which may be animated for showing the car movements over time, there are further opportunities for analysis. The tool aggregates the simulation results for the cells and links by time intervals of user-chosen length. Using time graph displays, we

can analyze the link loads, attained mean speeds, and numbers of suspended cars in the cells. Bottlenecks in the transportation infrastructure can be revealed.

After analyzing the predicted development of the traffic situation, it is possible to introduce modifications in the scenario (e.g., disable the use of some links and/or modify link weights, to model traffic re-routing) and run a new simulation. Through such "what if" analysis, it may be possible to find suitable measures for decreasing traffic suspensions and congestions.

4 Conclusion

In the recent years, our research was strongly focused on analysis of data concerning movement [4], including network-constrained movement. By developing and applying various visual analytics methods, we strived at comprehensive exploration of the potential opportunities that can be provided by movement data.

For network-constrained movement, we found data transformations that allowed us to visualize the interdependencies between two key aspects of the movement, traffic intensity and speed. Having a vivid picture, as in Fig. 3, we noticed common patterns and got an idea that the interdependencies can be quantified and expressed formally in a uniform way. To implement this idea, we developed new visual analytics tools that enabled us to represent the dependencies by formal models. This shows that visual analytics methods can help analysts not only to gain understanding (i.e., a mental model) of a phenomenon represented by data, but also to transform this mental model into explicit formal models.

Our next idea was that the models capturing the traffic intensity – speed relationships can allow prediction of not only typical movements but also unusual movements that were not represented in the original data. This is possible because the models generalize the data and can do extrapolation beyond the scope of the data. We have developed a traffic simulation tool capable of using the models derived from historical data and a visual analytics infrastructure that supports definition of traffic scenarios to simulate and analysis of simulation results.

Our research showed a principal possibility of using knowledge gained from historical movement data for prediction of development of traffic situations, even under unusual conditions. Moreover, one of our findings was that the dependencies between the traffic intensity and speed existing in a spatially abstracted network are similar to the known dependencies existing in road traffic and observed at the level of road segments. This opens a potential opportunity for performing rapid large-scale simulations of traffic situation developments on large territories when fine details are not required. This opportunity needs to be comprehensively investigated and tested in collaboration with transportation domain specialists.

A general conclusion concerning the use of visualization for predictive analytics that we can draw from our experience is that development of methods and tools to support the advancement from raw data to predictions can be done in an evolutionary way, as explained below.

At the first stage, analysts need tools enabling understanding and discovery of significant patterns. After such tools have been developed and applied, and patterns have been revealed, the character of these patterns can suggest what kind of formal model can be used to capture them.

At the second stage, analysts need tools for putting the observed patterns into formal models. What tools are needed depends on the character of the patterns and the type(s) of the model(s), which may not be known in advance. Hence, a new round of tool development may take place, to enable the second stage to be accomplished. When building formal models, analysts envisage how these models can be used for prediction.

At the third stage, analysts need tools enabling them to obtain, investigate, and compare predictions for various situations and circumstances. This may necessitate a yet new round of tool development driven and directed by the new analysts' goals and requirements.

Hence, predictive analytics consists of three successive tasks (analyze data – develop models – obtain and analyze forecasts). There may be cases when not all tasks can be fully defined in advance; hence, it may be not clear what supporting tools are needed. Clearer definitions for later tasks may emerge when previous ones are performed and results obtained. This sets requirements for further supporting tools. By this evolutionary process involving both tool development and analysis, a methodology and a supporting toolkit for predictive analytics in a specific application domain can be built.

References

[1] N. Andrienko and G. Andrienko. Spatial Generalization and Aggregation of Massive Movement Data, *IEEE Trans. Visualization and Computer Graphics*, vol. 17, no. 2, pp. 205-219, 2011.

[2] N. Andrienko and G. Andrienko. A Visual Analytics Framework for Spatio-temporal Analysis and Modeling, *Data Mining and Knowledge Discovery*, vol. 27, no. 1, pp. 55-83, 2013.

[3] D.C. Gazis, *Traffic Theory*, Kliwer Academic, Boston, USA, 2002.

[4] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer, 2013.

[5] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data, *VLDB Journal*, vol. 20, no. 5, pp. 695-719, 2011.