Project Number 288094

eCOMPASS

eCO-friendly urban Multi-modal route PlANning Services for mobile uSers

STREP
Funded by EC, INFSO-G4(ICT for Transport) under FP7

eCOMPASS – TR – 041

Traffic Congestion Prediction

T. Diamantopoulos

June 2013

Project Number 288094

eCOMPASS

eCO-friendly urban Multi-modal route PlANning Services for mobile uSers

STREP
Funded by EC, INFSO-G4(ICT for Transport) under FP7

eCOMPASS – TR – 041

Traffic Congestion Prediction

T. Diamantopoulos

June 2013

Traffic Congestion Prediction

Internal Research Review*

Themistoklis Diamantopoulos†

June 11, 2013

1 Overview

The purpose of this review is to cite the current state-of-the art approaches in the problem of traffic congestion prediction. Certain interesting approaches are summarized in the following sections.

In general, the traffic congestion prediction techniques have to handle the following problems:

- Data Acquisition: acquiring the data in a particular form, usually either from speed probes or loop detectors.
- Data Manipulation: creating certain metrics, usually traffic flow/volume, occupancy or even mean speed or travel time per link.
- Congestion Modeling: using the aforementioned metrics to define the jam and non-jam states of a road, done in a heuristic manner.
- Congestion Prediction: predicting jams in a short-term future, generally accomplished using classification algorithms, although regression and heuristics are also present in current literature.

Apart from reviewing some interesting approaches in section 2, the pros and cons of each of them are explored and their applicability to a speed-probe dataset is discussed. In addition, any approaches in traffic congestion prediction that clearly deviate from the purpose of this review are mentioned in section 3. Finally, section 4 concludes this review.

*Limited circulation. May be useful for deliverables and/or publications.

†Centre for Research & Technology Hellas, Information Technologies Institute (CERTH/ITI), 6th Km. Charilaou - Thermi Road P.O. BOX 60361 GR - 57001, Thermi, Thessaloniki - Greece

2 Traffic Congestion Prediction Approaches

2.1 On Feature Selection for Traffic Congestion Prediction

2.1.1 Description

S. Yang in [1] formulates traffic congestion prediction as a binary classification problem (jam - no jam). The data is drawn from 4000 loop detectors. Each sensor/detector provides with the *traffic volume*, i.e. the number of vehicles passing through the detector per time unit.

The first step of the analysis is data preprocessing so that traffic congestion is modeled. Assuming v_t^j is the traffic volume of sensor j at time t, the presence or not of jam in the respective road is determined using the following equation:

$$RoadState_t^j = \begin{cases} Jam & \text{if } v_t^j > ThreHigh \\ NonJam & \text{if } v_t^j < RatioNotJam \cdot ThreHigh \end{cases} \tag{1}$$

where:

$$ThreHigh = RatioHigh \cdot \max\{v_t^j\} \tag{2}$$

so that the ratio variables ($RatioHigh$, $RatioNotJam$) are adjusted to create appropriately skewed, yet realistic datasets.

Upon creating the two sets (T and $\bar{T}$), the author discusses the dimensionality problem. Since using the data from all sensors to predict the class of jam j is not possible in terms of dimensionality, the author applies a p-test to identify which features actually affect the jam state of j. For any sensor i, the p-test score with respect to j is defined as:

$$S_{ij} = \frac{|\mu_{ij} - \bar{\mu}_{ij}|}{\sigma_{ij} + \bar{\sigma}_{ij}} \tag{3}$$

where μ_{ij} and σ_{ij} denote the mean and standard deviation of jam training samples, and $\bar{\mu}_{ij}$ and $\bar{\sigma}_{ij}$ denote the mean and standard deviation of non-jam training samples.

Thus, for each sensor, the most important features (i.e. the ones with the highest score) are used as the input of the binary classifier. The classification for each sensor is performed assuming Gaussian distributions over the two datasets, so that the final probability for time t is given by the following equation:

$$S_{t,\tau}^j = \prod_{i=1}^{I} \frac{\Pr\{v_{t-\tau}^i \in Gaussian\{\mu_{ij}, \sigma_{ij}\}\}}{\Pr\{v_{t-\tau}^i \in Gaussian\{\bar{\mu}_{ij}, \bar{\sigma}_{ij}\}\}} \tag{4}$$

where the volume values for time $t - \tau$ are of course known and I is the total number of sensors. Finally, the author uses mean precision to evaluate the results and performs analysis to determine the optimal number of features required.

2.1.2 Review - Ideas

This subsection summarizes certain thoughts about the paper:

- Congestion modeling is based on heuristics that can be adjusted for the dataset at hand. It is rather simplistic, yet solid.
- Gaussian models are generally strong for low number of features, however they are less effective when the features are more. Thus, applying the algorithm to a scenario with speed probes would be both ineffective (in terms of results) and inefficient (in terms of performance).
- The p-test is effective for the scenario studied in this paper. However, performing a PCA over different dimensions of the data (i.e. means, standard deviations) should result in better feature selection.

2.1.3 Applicability to a Speed-Probe Dataset

The modeling part cannot be applied since there is no way of approximating traffic volume using speed probes. The p-test and the Gaussian Model classifier, however, may be applied with few adjustments to their input.

2.2 Congestion Prediction on Motorways: A Comparative Analysis

2.2.1 Description

Upon prior research [2], G. Huisken and M. V. Maarseveen in [3] collect data using 35 induction loops on the motorway A10 of Amsterdam. The detectors are "on" when a vehicle passes by them and they are "off" when no vehicle passes. The number of vehicles passing the detector per time unit (*volume*) as well as the percentage that the detector is "on" (*occupancy*) are known. In addition, the average and standard deviation of speed in a road segment is calculated using the respective series of loop detectors. Finally, the authors claim having an oracle indicating the presence of congestion.

The volume, mean speed, occupancy and standard deviation of speed can be given as features to any classifier, whereas the output class comprises of the binary congestion indicators for the ring road, which were totally 6. The authors test different classifiers, including multi-linear regression, an ARMA model, a heuristic Fuzzy Logic classifier, and three neural network implementations.

2.2.2 Review - Ideas

This subsection summarizes certain thoughts about the paper:

- The number of features is clearly not indicative of an urban scenario. Although the method may be effective for ring roads, using it in a large road network is prohibitive.

- The analysis performed in this paper is actually univariate. Transforming it to multivariate by using information from a large number of detectors would overload the method, especially since the feature space has 4 dimensions (volume, mean speed, occupancy and standard deviation of speed). An local or global analysis could be used to isolate useful features.

- The 6-class label since rather interesting. E.g. if a road has 3 points, A, B, C, it could be considered congested if 2 out of them indicate congestion. This might be useful.

2.2.3 Applicability to a Speed-Probe Dataset

A speed probe dataset generally offers only mean and standard deviation of speed. Although applying the methods might be feasible by using only these two features, there are some problems. At first, the absence of an oracle, indicating the need for modeling. In addition, the methods described in this paper are not noise tolerance. Finally, certain scalability issues may arise if multivariate analysis is required.

2.3 Prediction of Congested Traffic on the Critical Density Point Using Machine Learning and Decentralised Collaborating Sensors

W. Labeeuw et al. in [4] perform quite interesting analysis, despite their dataset being generated. The generated dataset consists of speeds in different points of a ring road, (supposedly) taken using cameras.

The authors initially address the problem as a regression problem, attempting to predict future velocity values using Bayesian regression over Gaussian processes. Due to performance issues, the problem is reduced to binary classification. Assuming u_t^r is the velocity for a road segment r at time t, three different cases are determined for the road state using the following equation:

$$RoadState_t^r = \begin{cases} Slowdown & \text{if } u_t^r < 7\,m/s \\ Congested & \text{if } 7\,m/s < u_t^r \leq 14\,m/s \\ Normal & \text{if } u_t^r \geq 14\,m/s \end{cases} \tag{5}$$

The authors use velocity data from every sensor and its local sensors to form the features. Two classifiers are tested for each sensor: an SVM and a C4.5 tree classifier. The results indicate that the C4.5 classifier is quite faster while it has better precision concerning the *Congested* and *Slowdown* sets. Finally, the paper hints applying the algorithms in a distributed multi-agent manner.

2.3.1 Review - Ideas

This subsection summarizes certain thoughts about the paper:

- The paper indicates that classification is actually a viable option when considering jams. However, the dataset is small and the algorithms do not seem to scale. The features considered are few.
- Since the dataset is generated for a ring road, the approach is rather problem-specific. Noise and sparse data are not taken into account.

2.3.2 Applicability to a Speed-Probe Dataset

The approach of this paper models the problem using speed values, thus the methods are generally applicable when it comes to a speed-probe dataset. However, the methods may not be noise tolerant enough for a real dataset. In addition, the analysis performed lacks of a generative classifier (e.g. Gaussian Bayes or Gaussian Mixtures Model) that could perform better in cases of few features.

2.4 IEEE ICDM 2010 Contest: TomTom Traffic Prediction for Intelligent GPS Navigation

The 2nd task of the contest, described in [5], refers to jams. The data used is generated. The dataset consisted of road segments such as:

```
32049370_32049364 32597785_32599710 251856122_224814449 ...
```

where the numbers are node ids and a road is defined between two nodes as `Node1_Node2`. In the training set, the first 5 road segments reflect roads that are closed due e.g. to roadwork, while the others are road segments that where jammed during a 20-minute interval.

The goal was to create an algorithm that could identify the jams in the next 40-minute interval (of course given for the training set). The output road sequence should be ordered according to jam appearance (from earlier to later). The evaluation metric used was such that earlier appearing jams were more important. Finally, note that the length of the output road sequence (number of jams) was unknown and had to be predicted. The winner and the first runner-up of the competition are analyzed in the following subsections.

2.4.1 kNN solution by L. Romaszko

The winning algorithm of the contest by L. Romaszko was a modified version of the k-Nearest Neighbors (kNN) algorithm. The algorithm compares sequences of the training set with the respective ones in the test set. A set named $Similarity$ is created so that each value of the set ($Similarity(D)$) denotes how similar is the respective training sequence (D) to every testing sequence. In other words, assuming the most similar training sequences to a specific testing sequence, the streets jammed in training sequences are probably jammed also for the testing sequence. The algorithm also outputs jam probability for each road, considering its position in the training sequences.

The second part of the algorithm concerns the *LengthSimilarity* set, which is used for approximating the length of each testing sequence. The average length of the k nearest training sequence is used to determine the length of the testing sequence.

Finally, an optimization is performed evaluating the results of the training set. A weight was adjusted for each street when its position in the sequence was not correct. In addition, using the fact that jams spread along neighboring streets, the minimum distance to any other street is used as a criterion.

2.4.2 Ensemble-Based Method for Task 2: Predicting Traffic Jam

J. He et al. [6] created an ensemble-based method which finished second in the 2nd task of the contest. The method combines the scores of different base predictors. Two types of predictors were created: the geographic propagation predictors and the nearest neighbor predictors.

The geographic propagation predictors track the flow of a jam based on the connectivity of the road segments. The predictor is based on the adjacency matrix of the road network C and a vector v_0 containing 1 only for the roads that are jammed at the first 20-minute interval. Multiplying v_0 with C gives a new vector denoting possible jam flow for future intervals.

The nearest neighbor predictors are based on comparing the training sequences with the testing sequence at hand. The authors describe 5 different predictors based on 5 different metrics. Each predictor assigns every jammed road with a score, which denotes its ranking. The aforementioned ranking is given using the distance of the kNN classifier as well as several heuristic parameters.

The scores of all predictors are combined in a linear fashion to form the final sequences. Concerning the length of each testing sequence, the authors use a simple average over the respective nearest training sequences. Finally, the different parameters of the method are adjusted using 10-fold cross validation over the training set.

2.4.3 Review - Ideas

This subsection summarizes certain thoughts about the two approaches:

- Both solutions are specifically adjusted to the task of the competition, thus their applicability in different scenarios is rather impossible.
- Nearest neighbor approaches seem powerful enough, however they are relatively slow when it comes to real-time applications since the training data has to be saved in place of a model that is not constructed.
- Using geospatial information for detecting the flow of traffic jams may prove as a quite useful idea.

2.4.4 Applicability to a Speed-Probe Dataset

The methods of this section cannot be applied to the speed-probe dataset since the problem is quite different. Obviously, performing a classification using kNN is feasible, however fully exploiting a speed-probe dataset includes also using various metrics (e.g. mean, variance). Furthermore, in terms of modeling jams, the contest's dataset is very specific, using external information about jammed roads and roads that are under construction for each scenario.

2.5 Summary

The different approaches are summarized in the following table:

Table 1: Approaches on Traffic Jam Congestion

Approach	**Year**	**Descriptor**	**Predictor**	**Applicability**
Yang [1]	2013	Loop detectors, traffic volume, formed binary classification problem	Gaussian models	With appropriate modifications
Huisken [2, 3]	2000	Loop detectors, volume - occupancy - speed, classification	ARMA, Neural Networks, Fuzzy Logic	Scalability issues
Labeeuw [4]	2009	Speed - simulated on ring road	SVMs, C4.5	Noise tolerance and scalability issues
ICDM [5, 6]	2010	Jam descriptor - simulated	kNN, spatial	Highly different descriptor

3 Other approaches

This section summarizes some approaches that are certainly interesting, yet quite divergent with respect to the scope of this review. Obviously, these approaches may also give rise to new ideas. However, the main topic of these papers is usually not congestion prediction but the construction of a system to effectively distribute congestion information and facilitate traffic.

As one might observe, several researchers prefer using heuristics to define congestion and different Data Mining algorithms to predict it. A slightly different approach is the one proposed by the line of work by G. Marfia et al. [7, 8, 9, 10]. The authors suggest a distributed Advanced Traveler Information System (ATIS), where cars send in their travel times upon traversing each road. Two heuristic thresholds are defined per road segment, one for having high congestion and one for leaving congestion, both based on the number of cars that traverse the segment in more than a travel time threshold. Since the main scope of that line of work is distributed ATIS integration, the procedure of

receiving and using probe data is explained in detail. However, the prediction approach is based only on univariate (with respect to each road segment) heuristics.

Other approaches include car-to-car transmissions, such as the one by Y. Ando, O. Masutani et al. [11, 12, 13]. The authors formulate congestion using a pheromone model, assuming cars are insects that generate pheromone. The amount of pheromone generated is proportional to the traffic, thus a decentralized system is updated based on the location of the cars and the amount of pheromone they emit. Although interesting, these approaches are mainly directed towards the area of Multi-Agent Systems, thus deviate from the scope of this review.

4 Conclusion

Most approaches presented in this review handle the scenario as a classification problem. Indeed, congestion is generally defined using certain heuristics, if not given by an oracle. The problem, thus, is formed and known classifiers are used to confront it. Concerning scalability, one might observe that it is not generally taken into careful consideration since loop detector datasets tend to be quite smaller than speed-probe ones.

Furthermore, certain techniques omit scalable solutions since they are mainly based on univariate analysis, which is by definition much more easily scalable than multivariate. A remark may also be made about decentralized approaches. The latter indeed are quite interesting, yet the lack of real and dense data tends to threaten their applicability.

Finally, the approaches discussed in this review are quite specific to the respective scenarios, thus cannot be easily generalized to other scenarios. Comparisons with speed-probe datasets are feasible only if the algorithms are stripped of some of their main elements, so that the remaining parts contain only classifiers. However, the ideas drawn from these approaches shall prove quite helpful for further research on the problem of traffic prediction congestion.

References

[1] Su Yang. On feature selection for traffic congestion prediction. *Transportation Research Part C: Emerging Technologies*, 26(0):160 – 169, 2013.

[2] Giovanni Huisken. Short-term congestion forecasting: Time series versus fuzzy sets. In *Proceedings of the 19th Annual South African Transport Conference*, 2000.

[3] Giovanni Huisken and Martin van Maarseveen. Congestion prediction on motorways: A comparative analysis. In *Proceedings of the 7th World Congress on Intelligent Transport Systems*, 2000.

[4] Wouter Labeeuw, Kurt Driessens, Danny Weyns, Tom Holvoet, and Gert Deconinck. Prediction of congested traffic on the critical density point using machine learning and decentralised collaborating cameras. In *New Trends in Artificial Intelligence : 14th Portuguese Conference on Artificial Intelligence pages*, pages 15–26, 2009.

[5] Marcin Wojnarski, Pawel Gora, Marcin Szczuka, Hung Son Nguyen, Joanna Swietlicka, and Demetris Zeinalipour. Ieee icdm 2010 contest: Tomtom traffic prediction for intelligent gps navigation. *2012 IEEE 12th International Conference on Data Mining Workshops*, 0:1372–1376, 2010.

[6] Jingrui He, Qing He, G. Swirszcz, Y. Kamarianakis, R. Lawrence, Wei Shen, and L. Wynter. Ensemble-based method for task 2: Predicting traffic jam. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1363–1365, 2010.

[7] Marco Roccetti and Gustavo Marfia. Modeling and experimenting with vehicular congestion for distributed advanced traveler information systems. In Alessandro Aldini, Marco Bernardo, Luciano Bononi, and Vittorio Cortellessa, editors, *Computer Performance Engineering*, volume 6342 of *Lecture Notes in Computer Science*, pages 1–16. Springer Berlin Heidelberg, 2010.

[8] G. Marfia and M. Roccetti. Vehicular congestion modeling and estimation for advanced traveler information systems. In *Wireless Days (WD), 2010 IFIP*, pages 1–5, 2010.

[9] G. Marfia and M. Roccetti. Vehicular congestion detection and short-term forecasting: A new model with results. *Vehicular Technology, IEEE Transactions on*, 60(7):2936–2948, 2011.

[10] Gustavo Marfia, Marco Roccetti, and Alessandro Amoroso. A new traffic congestion prediction model for advanced traveler information and management systems. *Wireless Communications and Mobile Computing*, 13(3):266–276, 2013.

[11] Osamu Masutani, Hiroshi Sasaki, Hirotoshi Iwasaki, Yasushi Ando, Yoshiaki Fukazawa, and Shinichi Honiden. Pheromone model: application to traffic congestion prediction. In *AAMAS'05*, pages 1171–1172, 2005.

[12] Yasushi Ando, Osamu Masutani, Hiroshi Sasaki, Hirotoshi Iwasaki, Yoshiaki Fukazawa, and Shinichi Honiden. Pheromone model: Application to traffic congestion prediction. In *Engineering Self-Organising Systems'05*, pages 182–196, 2005.

[13] Yasushi Ando, Yoshiaki Fukazawa, Osamu Masutani, Hirotoshi Iwasaki, and Shinichi Honiden. Performance of pheromone model for predicting traffic congestion. In *AAMAS'06*, pages 73–80, 2006.