Traffic state estimation using hierarchical clustering and principal components analysis: a practical approach

Georg Foerster
Fraunhofer Institute for Transportation and Infrastructure Systems (IVI)
Zeunerstr. 38, D-01069 Dresden, Germany
phone: +49 351 4640 682, fax: +49 351 4640 803, e-mail: georg.foerster@ivi.fraunhofer.de

Abstract: Traffic state estimation and prediction are fundamental requirements for automatic control of urban road traffic with both adaptive traffic lights and variable message signs. For that, collecting of actual traffic data is necessary. This paper deals with the combined application of principal components analysis (PCA) and hierarchical cluster analysis (HCA) for the specification of the needed number of stationary road traffic sensors and their preferable locations within a given road network. Both methods are introduced briefly. A practicable procedure for using these methods is derived and it is shown that their combination is effective. First tests based on microscopic simulation data and on real volumes of inductive loops lead to plausible and promising results in application of the proposed procedure.

Keywords: traffic state estimation, road traffic control systems, multivariate data analysis, cluster analysis, principal components analysis, floating car data, loop detectors, data fusion

Foundation note: Activities described in this paper are part of the project ORINOKO, which is sponsored by the German’s Federal Ministry of Education and Research (ID: 19B4036B, term of project: Dec 2004 – Feb 2008).

1. Introduction

Various sensor types and methods are available for traffic data detection, such as inductive loops, video sensors, and floating car data (FCD). Each approach for urban traffic control assumes that a specific quality level of traffic state estimation is known. Then the basic questions are how many sensors are needed and where should they be placed. In particular it is interesting to determine the minimum quantity and allocation of additional stationary detection with a given density of FCD probes to achieve this quality level under budget constraints.

Depending on structure and traffic demand in every urban road network links or locations can be found that provide redundant information. However, other links or locations hold significant new information. When there is a known quantity and quality of FCD, of course, additional loop or video detectors are interesting only, if they provide essential new information. It can be a very long and fault-prone way to find relevant links or locations with non-redundant information, because you have to handle and evaluate a lot of data.

In this paper two approaches are described which could help to reduce the amount of data and the number of variables, to find redundancies and to generate an adequate traffic state estimation: first the principal components analysis and second hierarchical clustering. Both approaches are introduced and joined from the practical point of view.

The underlying study is conducted in the ORINOKO research project in Nuremberg, Germany. Urban road traffic management is the main focus of this project. In Nuremberg two areas with different allocation of road traffic control systems exist. On the one hand the event and exhibition area is equipped with an advanced traffic control system containing many inductive loop detectors, dynamic message signs, dynamic direction signs, and dynamic choice of signal programs. Otherwise, the rest of the City area does not have such a system,
but traffic problems like congested road links also exist in this less controlled part of Nuremberg. To close this lack of detection a taxi-based FCD system has been implemented some years ago. One main concern of the ORINOKO project is how many loop or video detectors have to be placed in the road network, so that the fusion of FCD and sensor data can provide authentic information about traffic state (figure 1).

![Diagram](image)

**Figure 1:** For a given density of detection and a needed quality of traffic state estimation the number and allocation of additional sensors have to be found.

Traffic state estimation normally means to deal with time series of a very big amount of traffic data (e.g. volumes, speeds) for many links or locations, respectively. Of course, all the data as a whole represent causal incidents and the corresponding effects, which can be stochastic (e.g. accidents) or deterministic (e.g. known bottlenecks), and they interfere with each other. To decide which road network location or link contains essential information and which location or link only provides redundant information all interdependencies have to be analysed. For that, various methods of multivariate data analysis are available.

2. **A short review of multivariate data analysis**

Multivariate data analysis is a method that is often used by many researchers, e.g. in sociology, in psychology, in meteorology, and in engineering. Many textbooks and manuals on that topic are available, [1] [5]. The principal components analysis (PCA) and the hierarchical cluster analysis (HCA) are sections of the multivariate data analysis.

Generally, it can be said that with respect to road traffic PCA in many cases is applied for exploration of incidents and accidents and of environmental issues. Beyond this it is used in video-based traffic sensors to identify moving objects. HCA in a remarkable number of papers is used for road type classification and short-term prediction. It would be too much to reference all sources on that topic here in this paper.

PCA and HCS were chosen for finding redundancies in road network detection, because they seem to fit best, and because they are comparatively easy to handle and to understand. Even
though there are some powerful computer tools for data analysis [3], it is necessary to get an idea of the basic procedures.

2.1 **Principal components analysis (PCA)**

Principal components analysis (PCA), sometimes seen as a specific kind of factor analysis, aims on reducing the amount of data by varying the “point of view” on the variables. The basic idea is that all measured variables are results of not yet known factors or so called principal components. From mathematical focus the dimensions of the data set are reduced from the number of variables to a less number of principal components by rotating the data set within the multidimensional data space.

One result of the PCA will be a matrix of coefficients, which represent the original data within the new data space of principal components. Second, there will be a matrix of so called loadings for components and variables. The loadings describe an extent of the variance of original data declared by the principal components. Their sums for each principal component represent the appropriate eigenvalues.

A useful interpretation of the loadings can be seen in the relevance of specific variables for the components. Components that explain much variety of the original data set are essential, whereas components that explain little variety are not essential. The main task in PCA is to understand the meaning of these calculated components.

2.2 **Hierarchical cluster analysis (HCA)**

Hierarchical cluster analysis (HCA) aims at finding data sets or variables that “belong together” and at separating them from the other data. The result is a number of clusters of variables. Within one cluster the variables are homogeneous in a specific manner. Between two different clusters the variables are more heterogeneous.

One way of identifying clusters is to use paired distances between variables. These distances are calculated by analysing sorted pairs of observations of the accordant two variables. For instance, for clustering of variables that show linear coherences linear correlation coefficients would be used to assign a distance. For clustering of variables whose observations are near to each other by their absolute values the Euclidean distance would be used. Other distance or similarity dimensions are described in [1], [3], and [5].

The following example describes this method. Six variables (A-F) should be clustered by absolute (“Euclidean”) distances. The upper table in *figure 2* shows a set of fictive distances between the variables. These values can be visualised in a graph, where the nodes represent the variables and the edges represent the distances (*figure 2, down left*). Hierarchical (single linkage) clustering first finds the minimal spanning tree within that graph (*figure 2, down right*), i.e. it links all the nodes by their shortest paths. That graph can be rearranged as a dendrogram (*figure 3*).

In this dendrogram clusters can be realised, so A and B seem to belong together, or C and D as well. E and F have bigger distances to AB or CD, respectively, so they seem to be more unique. There are various methods to separate clusters from each other. For example a maximum number of clusters can be set, or they may be separated in consideration of distance ratios within and between clusters. See [1], [3] or [5] for details.

One interpretation of the resulting clusters can be used to find redundant and essential information. The less the distance between two variables is the less the gain of information is when using both data sets instead of one only. In the example in *figure 3* redundancies between the variables decrease from left to right.
3. An approach to join both methods for the detector allocation problem

Several studies that use HCA or PCA for prediction of traffic data have been made in the past, [2], [4], [7]. Even statistical tool boxes for traffic data analysis were implemented [8]. All these papers show in principal the practicability of these methods for traffic state estimation. The main problem in the ORINOKO project is: How many additional inductive loops or video detectors have to be placed in the road network, so that the fusion of existing FCD and
sensor data can provide authentic information about traffic state, and where should they be taken to? To solve it both methods are joined.

3.1 Needed accuracy and quality of traffic state estimation

Having a 100-percent-estimation of the traffic state in a road network means to be really omniscient, because traffic flow and speed in the network would be known everywhere at every time. Simplifying the network links can be divided into segments of the same length, e.g. 25 m, 50 m or 100 m. For further simplification the time can be split up into intervals of 5 min, 15 min or maybe 60 min.

The accuracy of the traffic state estimation would be 100 %, if all the temporal aggregated and spatial segmented data for the segments are known. Certainly, segment length and time interval will depend on the specific temporal and spatial accuracy needed for traffic state monitoring or operative traffic management. So the planning engineer must force it regarding to the purpose of the data.

Of course, from economic issues it is not effective to equip 100 % of road segments with detection. Furthermore, FCD are not available for all links and all time at the same sufficient level. Various methods for traffic state estimation and prediction are known and in use to handle this lack of data. To rate the quality of these methods the modelled values can be compared with the real values. The less the (relative) difference between model and real-world data is, the higher the quality level is. The presetting of this needed quality level is a concern of the planning engineer, too.

In the approach presented in this paper it is assumed that needed accuracy and quality of traffic state estimation are known. These are the objective items for the optimisation of sensor allocation in the road network.

3.2 Data needed for running the method

For running the proposed method time series of data for as much segments as possible, best is 100 % of all segments, are needed. To generate these time series microscopic traffic simulation tools like VISSIM [6] can be used. Of course, an accurate validation of the simulation model of the network is necessary. It can be done by using data from existing sources (loop detectors, FCD). Or FCD only can be used if data for a sufficient long time period is available. The raw data table would consist of columns representing the single segments $d_j$, i.e. the potential sensor locations, and rows representing the continuous time intervals $t_i$. In the single cells the (simulated) reference data $x_{ij}$, e.g. flow or speed, can be found. The data should be arranged as follows in table 1.

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_j$</th>
<th>$d_-$</th>
<th>$d_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>$x_{11}$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_2$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_i$</td>
<td>...</td>
<td>...</td>
<td>$x_{ij}$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_-$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$t_D$</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>$x_{JD}$</td>
</tr>
</tbody>
</table>

*Table 1: Raw data matrix.*
For practical use the time interval should be set long enough that under average traffic conditions cars could pass the network within. This assumption implicates that the network dimensions depend on the temporal sharpness of the analysis. For example, if the time is set up to 15 min and the average travel speed along the longest route in your network is at 25 km/h or more, then the network could have an dimension of round about 4.0 km in diameter or less. If the network exceeds the appropriate dimension then the method presented here could result in failings for mathematical reasons.

3.3 Application of principal components analysis

The PCA is used to investigate the essential facts (components) that are inherent in the road network and traffic conditions and that are represented by the raw data matrix (table 1). These factors could be seen in systematic variations in-between the hours of one single day (normal conditions, rush hour etc.), in-between the days of the week (working days, weekend, holiday etc.) or in spatial-temporal interdependencies within the network (e.g. event-related specific origin-destination-matrices). These factors typically are not known completely and that’s why PCA is used.

Figure 4 shows the first step of operations to be done. By processing PCA all data are transformed from the variables space into the components space. The left part of the figure is the raw data set and the right part represents the loadings matrix as one result of the PCA. The number of P principal components has been found. Looking at row \(d_j\), it can be said that for detector \(d_j\) the components \(p_k (k = 1 \ldots P)\) explain 100 % of variation of the original data vector \(d_j\). Looking at column \(p_k\), it can be said that detectors \(d_j (j = 1 \ldots D)\) load 100 % on component \(p_k\). The value of one single coefficient \(c_{jk}\) describes the corresponding matters between the single detector \(d_j\) and the single component \(p_k\).

The number \(P\) of components should be significant smaller than the number \(D\) of detectors. The principal components \(p_k\) are sorted by descending impact on the variance of original data set. If all the loadings in that matrix from \(c_{11}\) to \(c_{DP}\) are added then the resulting sum \(C\) marks the fact that all components \(P\) explain all original data variance of all detectors \(D\).

One basic decision now is to define how many principal components should be considered for further steps. We use the percentage of explained variance for this. When the loadings \(c\) of each column \(p_k (k = 1 \ldots P)\) in the right part of figure 4 are added and for each column this sum is divided by the overall sum \(C\) then the percentage of explained variance for each single principal component with respect to all detectors is determined. If you cumulate all these percentages from columns \(1\) to \(P\) then you get 100 %, naturally. The planning engineer has to decide, which percentage of original data variance shall be explained by the components. A traditional value could be 95 %, for example. The percentages would be cumulated beginning from \(k = 1\) until they reach or just exceed this value. The last added column (component) is \(k = E\). The first \(1 \leq k < E\) principal components are called “essential” for traffic state in the
investigated network. All other components \( E < k \leq P \) are called “not essential” and they will be disregarded.

The described procedures result in a reduced loadings matrix (table 2). Calculating the sum \( C_i \) \( (i = 1 \ldots D) \) for each row leads to the variance explained by the essential principal components for each detector, and \( C_i,\text{rel} \) is the percentage of the loadings of all components on the appropriate detector \( d_i \).

<table>
<thead>
<tr>
<th>detectors</th>
<th>essential principal components</th>
<th>( p_1 )</th>
<th>( p_k )</th>
<th>( p_E )</th>
<th>( \text{sum} )</th>
<th>( \text{rel. sum} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td>( c_{11} ) ... ... ( C_1 )</td>
<td>( C_i,\text{rel} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( d_j )</td>
<td>... ( c_{jk} ) ... ( C_i )</td>
<td>( C_i,\text{rel} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( d_D )</td>
<td>... ... ( c_{DE} ) ( C_D )</td>
<td>( C_D,\text{rel} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{sum} )</td>
<td>( C_E )</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Reduced matrix of loadings containing essential principal components only.

The interpretation of this table is as follows: The \( E \) principal components are essential for the explanation of the traffic state in the whole network. The percentage \( C_i,\text{rel} \) says what effect the data of each detector \( d_i \) has to all essential components. The aim is to implement those detectors which contribute most to these essential principal components.

But it is evidently not sufficient to compare the single \( C_i,\text{rel} \) and to order them by their descending values. The reason is easily to understand looking at volumes (vehicles per time period) as original data. In a complex road network, normally some links exist which are good indicators for the network traffic state (e.g. major streets) whereas other links (e.g. in residential areas) are not. It is obvious that detectors on these indicator-links are recommendable. By doing the PCA each detector is provided with loadings. Spatial and directional neighbouring detectors on one link will have similar loadings and they will share the loading of the whole link. So, if there are a very short non-indicator-link containing only few detectors on the one hand and a much longer indicator-link containing more detectors on the other hand, then the \( C_i,\text{rel} \) of the single detectors placed in the non-indicator-link could be at a higher percentage than the \( C_i,\text{rel} \) of the detectors in the indicator-link. To handle this problem hierarchical cluster analysis is used the next step.

### 3.4 Application of hierarchical cluster analysis

The HCA is used to find detectors that have very similar values in the original data time series, i.e. to identify clusters having redundant detector data within. This similarity can result from spatial closeness or from other systematic interdependencies among detectors.

First, paired distances between detector data have to be calculated. Due to the interest in absolute differences, the Euclidean distance is used. From the original data set \( x_{ij} \) (figure 5, left) a quadratic and symmetric matrix of Euclidean distances \( a \) is developed (figure 5, right).
Figure 5: Calculation of Euclidean distances based on original data set.

At this point the planning engineer has to decide how to establish the border between clusters. One option is to preset the absolute number of clusters. This could seem to be practicable if the needed accuracy and quality of traffic state estimation is not at a very high level, or if the number of needed supporting data points in the traffic model is known “from experience”. Another option is to take the distance ratio into consideration. This is suggestive if systematic interdependencies are not known yet. In this study the second path was taken. Processing the HCA leads to the assignment of each detector \( d_j \) to a cluster \( y_j \) (table 3). Each cluster contains one or more detectors, so the number of clusters should be less than the number of detectors.

### Table 3: Detectors are assigned to clusters by HCA.

<table>
<thead>
<tr>
<th>detectors</th>
<th>cluster ID</th>
<th>inner cluster average distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td>( y_1 )</td>
<td>( \bar{a}_1 )</td>
</tr>
<tr>
<td>( d_2 )</td>
<td>( y_2 )</td>
<td>( \bar{a}_2 )</td>
</tr>
<tr>
<td>( d_j )</td>
<td>( y_j )</td>
<td>( \bar{a}_j )</td>
</tr>
<tr>
<td>( d_D )</td>
<td>( y_D )</td>
<td>( \bar{a}_D )</td>
</tr>
</tbody>
</table>

Within every single cluster the average (Euclidean) distances \( \bar{a}_j \) between detector \( d_j \) and all other containing detectors are calculated. The detector which has the smallest average distance to all the other detectors within its cluster is marking a kind of “centre of gravity” for that specific cluster. This central detector is set for each cluster.

#### 3.5 The combination of PCA and HCA

Finally the results of PCA and HCA have to be combined. From PCA every detector’s relevance to the essential principal components is known, and results from HCA show clusters of detectors that carry pretty redundant information.

The idea of the proposed method is to take those clusters into consideration whose impact on the essential components is above average. This postulation is done because of cost efficiency. The maximum number of detectors will be the number \( Y \) of clusters. The total costs of detection are a result of the number of implemented detectors. So the cost of one implemented detector is the rate of \( Y^{-1} \). We only choose those clusters (central detectors) whose relative gain of impact is bigger than its relative gain of costs.

Table 4 shows the number of \( Y \) clusters with the cluster IDs \( y_m \) \( (m = 1 \ldots Y) \). Every cluster contains one or more detectors. Adding all relative sums \( C_{rel} \) (table 2) of the detectors assigned to a specific cluster, the output is the sum \( L_m \) of detector percentages for each cluster. The higher this sum is the more the appropriate cluster loads on the principal components.
The arithmetic average of the impact of one cluster $y_m$ on the essential components is just $Y^{-1}$. If $L_m \geq Y^{-1}$ then the cluster loads as much or more than the average, if $L_m < Y^{-1}$ then the cluster does not.

For each chosen cluster the central detector is set as relevant. The number of chosen central detectors and their locations are interpreted as the solution of the above formulated detector problem. Existing detectors or FCD just have to be taken into account by presetting the appropriate segments.

4. First tests of the approach

In this early phase of the study the microscopic simulation system VISSIM [6] was used to generate a set of original data. Figure 6 shows the modelled visionary road network having a dimension of approximately 1.8 km in diameter and a total link length of about 14 km.

The network consists of one central junction and seven intersections where cars can enter or leave the network. The duration of the simulation is at 10 h, traffic demand alternates between the two origin-destination-matrices (OD) which are shown in figure 6, too. The OD matrices here are completely fictive. The only intention of the contained traffic volumes is to generate a synthetic data sample for investigation of the manner of the proposed method. So because of that at this early point only one simulation run has been accomplished. Future work must show how stochastic influences have to be taken into account.
The whole set of results generated by the means of simulation can be found in figures 7-9. For example, if the spatial accuracy is set on segments of a length of 25 m and the temporal agglomeration is set on 15 min, then the number of detectors representing 100 % information of the traffic state is 531 in the fictive simulation network. Running an algorithm that follows the method described in section 3 of this paper results in the definition of only 35 sensor locations. This is corresponding to a sensor density of 6.6 % and their impact on the traffic state estimation is at 73 %. If the level of needed spatial accuracy is reduced from 25 m down to 50 m (maximum number of detectors is 255) the number of resulting sensor locations is 25 (9.8 %) with an impact of 72 %.

![Figure 7: Absolute number of sensors depending on length of segments and length of intervals as a result of simulation.](image)

![Figure 8: Relative density of sensors depending on length of segments and length of intervals as a result of simulation.](image)
The results are plausible. The higher the level of spatial accuracy is the higher the total number of needed detectors has to be. Furthermore, the higher this level is the higher is the relative reduction of sensors. The last issue is due to the effect of clustering of redundant neighbouring sensors.

A second test was performed on the basis of real traffic volumes in the testing area of the above mentioned ORINOKO project in Nuremberg. In this test field 52 detector sites can be found. According to the number of lanes almost each site includes more than only one detector. The data of inductive loops was collected during February 2007 and was put into aggregates of a length of 20 minutes. The proposed analysis method resulted in six detector sites (11.5 %) that have an impact of 58 %. The recommended detector sites seem to be plausible, because four of them are situated at important links near the exhibition and sports area, which is one of the most important origins/destinations in this test field.

5. Conclusion and future work

The previous results seem to be very promising to make a perceptible contribution to the detector allocation problem. Two basic issues for traffic state estimation on a road network have been taken into consideration. First, the systematic interdependencies between network topology and traffic demand are interpreted as principal components. Second, the redundancies within the network are handled with clustering. The number of detectors found by processing the presented method seems to be practicable and realistic.

For the underlying research project ORINOKO the method has to be tested more intensely. First the number of experiments using the simulation model must increase to better understand the “adjusting screws” and to validate the results. Especially, the question whether the proposed sensor locations are optimal from a mathematical point of view has to be answered. The microscopic simulation as means of investigation of the manner of the proposed method can be used to learn more about the stochastic stability of the results. This is necessary because real traffic flow and speed follow stochastic principles. Only if the method leads to robust sets of essential sensors it could be helpful for practical use.

Then real traffic data from the Nuremberg taxi FCD system will be used as input data. Doing that another interesting matter has to be taken into consideration. Depending on various facts the availability and the temporal-spatial density of taxis passing the network strongly varies, so the present spatial-related method has to be enhanced by a temporal dimension.
The very good availability of traffic data generated by inductive loops in the Nuremberg test area shall be used to extend the model. At the moment only spatial interdependencies have been taken into account. Of course inter-temporal interdependencies can be found, too. So it is common to process traffic state prediction using recent time series of measured data. A method for optimisation of sensor sites should consider this.

Using methods of multivariate data analysis for traffic state estimation is an efficient way to investigate road traffic networks. Advanced computer-based tools make it easy to handle the big amount of data and to concentrate on the primary problem.

References


