Visualisation and Analysis of Urban Traffic Data

Livia Della Ragione*, Veniero Giglio*, Vesna Luzar**, and Mario Rapone*

*Istituto Motori, CNR, Napoli, Italy
**University Computing Centre, Zagreb, Croatia

Abstract:
Data on vehicle operating conditions have been recorded by a specially equipped car, while circulating through Naples on several typical routes, during rush hours. Traffic was defined in terms of parts of movements between two consecutive stops ("kinematic sequences"), characterized by a number of kinematic variables, such as mean velocity, driving and idling time, gear ratio, number of peaks, etc. Cluster analysis was used to identify typical groups of sequences, and was followed by canonical discriminative analysis yielding canonical scores used in plots. Series of successive sequences were then grouped into cycles, based on observed transition probabilities among clusters. Cycles were then clustered, yielding seven groups of typical cycles that could be used to characterize traffic intensity and to visualise traffic conditions. Furthermore, the cycles could be used to simulate typical driving conditions and to measure exhaust emissions and gas consumption in the laboratory. Several different visualisation techniques are proposed, such as time and distance plots of the sequences for visualising traffic along different routes, taken at various time of the day, and for identifying highly congested traffic, or highly polluted zones. Other plots of interest include plots of consecutive sequences by their canonical scores, that distinguish between groups of sequences corresponding to different traffic levels, two and three dimensional representation of the groups of sequences and cycles, etc. Also is described how the video images recorded by video camera can be integrated with other traffic visualisation tools.

1. Introduction

During the last decade several studies have been performed to measure the emission effects of the automobiles on the move. A large experimental study conducted as part of the DRIVE - modern European project, was aimed at modelling vehicle emissions as a function of traffic speed parameters (Andre M., et al., (1991), Maurin M. and J.-P. Crauser (1990)). It has been shown that in order to predict exhaust emissions and gas consumption, it is essential to characterise traffic conditions and car kinematics in terms of parts of movements between two successive stops. In the previous studies kinematic parameters were recorded using different classes of cars, driven in various countries and zones (urban, road, motor way, etc.), and operated by different drivers.

In the study described here, the primary interest was to characterise urban traffic along the most commuted routes of Naples, during rush hours. A planned experiment was performed on one car, using different drivers, on several carefully designed routes, at different times of the day. Therefore, the main objective of the first phase of the project was to identify, describe and visualise different traffic conditions as characterized by kinematic sequences, and series of sequences (cycles). In the second phase, typical driving cycles obtained in this first phase will be repeated on the dynamometer chassis, to measure emissions and ultimately to build emission models.

During the first phase, the test car was driven for 26 hours and 23 minutes. In total, there were 1962 kinematic sequences obtained from 29 trips. All trips were performed during rush hours (8:30 - 10:00a.m., 4:00 - 6:00p.m.), and by five different drivers. Total distance covered was 363 km, with total idling time amounting to 7 hours 58 minutes (30% of total duration). In addition to car operating parameters (velocity, gear ratio, temperatures, consumption, etc.), recorded at one second intervals, a video camera was recording images in front of the car, during each trip.

In the following paragraphs we show how kinematic sequences were grouped into homogenous classes, and the type of graphical tools that were used for their visualisation. Next, grouping of series of consecutive sequences into cycles and their classification will be described. It will then be demonstrated how the traffic during trips can be characterised in terms of kinematic sequences and cycles. Finally, the concept of simultaneous visualisation will be outlined, based on the linking of traffic condition visualisation, video imaging and map positioning tools.
Throughout the paper, the emphasis is on the variety of techniques for graphical presentation, provided by SAS/GRAPH™, and on the advantages of certain types of visualisation tools.

2. Kinematic Sequences: Definition and Grouping

The pieces of movements (stop to stop), denoted as kinematic sequences can be described in terms of a number of variables, instead of just speed and acceleration, as has been the case in some previous studies. In this study sequences are characterised by the following variables:

- **VBAR** (average speed)
- **VMAX** (maximum speed)
- **PEAKS** (number of peaks)
- **V20** (percentage of time velocity is less than 20 km/h)
- **V30** (percentage of time velocity is between 20 and 30 km/h)
- **V40** (percentage of time velocity is between 30 and 40 km/h)
- **V100** (percentage of time velocity is greater than 40 km/h)
- **TSEQ** (sequence duration)
- **DIST** (distance covered)
- **IDLT** (idling time)
- **ACC0V** (percentage of time with stable velocity)
- **GEAR0** (percentage of time in neutral)
- **GEAR1** (percentage of time in first gear)
- **GEAR2** (percentage of time in second gear)
- **GEAR3** (percentage of time in third gear)
- **GEAR4** (percentage of time in fourth gear)
- **GEAR5** (percentage of time in fifth gear)

In the DRIVE project all the above variables except the gear ratios (GEAR0 to GEAR5) have been used to characterize and group sequences. We found that by including the information on gear ratio, better classification was obtained. On the other hand, it is well known that besides velocity and acceleration, a critical factor in emission and consumption prediction is also gear ratio utilisation. This is of particular importance when analysing crowded urban traffic.

Several different clustering techniques have been applied in order to classify the observed sequences into homogeneous groups, according to the measured variables. After examining alternative classifications, we decided on the six cluster solution yielded by two-stage density linkage, a modification of the single linkage hierarchical clustering method. The number of clusters was determined by inspecting the values of pseudo $t^2$ statistics at different levels of cluster joining, and by viewing different cluster solutions in the 2-dimensional space of the first two canonical variables. Namely, in order to summarize the between-cluster variation, and to visualise the clusters in a lower dimensional space, canonical discriminant analysis was applied. A simple description of the clusters in terms of relative frequencies, percentage of total duration, and of total distance, as well as mean values of some of the variables are given in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>CLUSTERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>STATISTICS</td>
<td>1</td>
</tr>
<tr>
<td>%</td>
<td>53.31</td>
</tr>
<tr>
<td>% duration</td>
<td>29.90</td>
</tr>
<tr>
<td>% distance</td>
<td>6.32</td>
</tr>
<tr>
<td>mean VBAR (km/h)</td>
<td>5.46</td>
</tr>
<tr>
<td>mean TSEQ (s)</td>
<td>27.15</td>
</tr>
<tr>
<td>mean DIST (m)</td>
<td>21.95</td>
</tr>
<tr>
<td>mean IDLT (s)</td>
<td>16.21</td>
</tr>
<tr>
<td>mean PEAKS</td>
<td>3.29</td>
</tr>
<tr>
<td>mean ACC0V (%)</td>
<td>18.65</td>
</tr>
</tbody>
</table>

Table 1. Clusters Summarised by Mean Values and Percentages
3. Visualisation of Kinematic Sequences

There are usually many different ways to present the same data, ranging from the simple univariate charts to various multidimensional graphical tools. Each one can add to understanding and explaining the underlying data structure. In the case of kinematic sequences, we have found the following representation specially useful:

- For a global view and relative positioning of clusters the best tools are either the 2- or 3-dimensional plots of the first canonical variables (or, alternatively, principal components), with different plotting symbols showing different clusters. In Figure 1, the six clusters are presented in the space of the first two canonical variables. Due to the fact that the six gear ratio variables (GEAR0-GEARS), and the four velocity variables (v20-v100) each add to 100%, and that the velocity (VBAR) is the ratio of distance (DIST) and duration (TSEQ), the points shown fall within a hyperbolic paraboloid of a hyperplane of dimension 15 (=17-2). The six black dots are the cluster barycenters, the typical representatives of each cluster.

![Figure 1. Canonical Variables Identified by Cluster](image-url)
The velocity and time diagrams of the "typical sequences" from clusters 1, and 6, corresponding to fluent and congested traffic, respectively, are presented in Figures 2a and 2b. The use of gear ratio is denoted by different plotting symbols. This is the simplest and most understandable representation of the clusters, but is deficient with regard to giving information on the variability within the clusters. The alternative would be to plot an overlay of a random sample of sequences from each cluster. Unfortunately, more than 5 sequences can not be clearly distinguishable when plotted together. (Note: principal curves might be employed here.)

Figure 2a. Typical fluent kinematic sequence

Figure 2b. Typical congested kinematic sequence

Standard stacked bar charts, with cluster as stacking variable can be used for portraying empirical densities of each variable separately, and the separation of clusters. These charts are very useful for explaining the cluster in terms of
original variables, such as mean velocity, mean distance, percentage of time with stable velocity, etc.

- For displaying the ordering of clusters according to frequency, duration, distance covered, and number of gear ratio changes, Pareto-like charts, as shown on Figure 3 are effective. Immediately it can be seen that although cluster one (i.e. congested) is highest in frequency (53% of all sequences), it accounts to only 6% of total distance, 30% of duration, and 18% of all gear ratio changes. The two fast clusters, 5 and 6 account for 93% of all mileage, and 67% of total duration.

- Alternately, for presenting ratio variables such as V20, V30, V40, V100, and GEAR0-GEAR5, and their distribution across the clusters, we used pie charts and stacked bar charts.

![Sequence Clusters Ordered by Frequency](image1)

![Sequence Clusters Ordered by Duration](image2)

![Sequence Clusters Ordered by Number of Gear Changes](image3)

![Sequence Clusters Ordered by Distance](image4)

*Figure 3. Pareto-like diagrams*
• Since the two variables which have the highest effect on emissions are velocity and acceleration, their empirical bivariate density representations, such as contours or 3-D plots, for each cluster separately, were produced. Incidentally, the differences among the joint densities for different clusters indicate that the clusters are also well separated on emission levels and consumption.
• SAS/INSIGHT™ was very useful for the initial visualisation of sequences and clusters, in particular the rotating 3D plots.
• Although there are many other ways to represent multivariate data, such as star or faces plots, but we haven't found them to be applicable to this study because of too many sequences. Which is to say, the velocity diagram seems more useful in this particular case.

4. Kinematic Cycles

Once the kinematic sequences are defined and classified, the next step in the analysis of the car kinematics was to find a rule for joining successive sequences into homogenous series. From Figure 1 it is apparent that there are two large groups of sequences, according to the values on the first canonical variable, with the sequences on the positive (fast) and negative (slow) pole corresponding to fluent and congested traffic, respectively. For this reason it is more convenient now to consider only 2 large groups, denoted as 1 and 2, consisting of clusters 1-4, and 5-6, respectively. Grouping of consecutive sequences into homogenous series is then simply based on transitions between the two large sequence classes. The beginning of a new cycle is always defined by the transition from one group to the other. Thus, each trip can now be represented (see Figure 4) as a series of sequences shown by their value on the first canonical variable. The number next to each sequence point represents the large class (1 or 2) to which the sequence belongs. It can be seen that the cycles defined in the above way distinguish between different traffic conditions: series of sequences with high CAN1 values corresponding to fluent traffic, on one hand, and congested traffic, shown by low CAN1 values, on the other hand. The length of a cycle ranges from 1 to as much as 44 sequences, covering distances from only .04 to 8311 meters. Cycle duration vary from 2 seconds to as much as 2835 seconds

Figure 4. A Trip Represented as a Series of Sequences

Once the cycles are defined, the next task was to classify them into groups of movements, corresponding to different traffic
situations, in a way similar to how the sequences were classified into 6 clusters. While a sequence is strictly bounded by two consecutive stops, a cycle extends over a section with homogenous traffic conditions (e.g., see Figure 4, sequences 62-110). Therefore, the grouping of cycles into clusters was based on CAN1, which distinguishes well between the two basic traffic levels (slow vs. fluent). In addition, the number of sequences in a cycle and percentage of driving time were used for grouping cycles into clusters. Ward's hierarchical clustering technique and pseudo $t^2$ statistics yielded seven clusters, corresponding to different levels of traffic. Clusters of cycles have been presented in a way analogous to the clusters of sequences, displayed in Figure 1, but are not shown here.

5. Characterization of Trips in Terms of Kinematic Sequences

Besides their use in measuring and building models for emission prediction, kinematic sequences were found convenient for graphical presentation of different trips patterns. Kinematic sequences, plotted in two dimensional space of cumulative time and distance from the route starting point, were particularly useful in their ability to identify bottleneck areas, and for viewing traffic problems/patterns in general. Two of such plots are shown in Figures 5a and 5b. Each sequence is composed of two parts: idling and driving period. Therefore, a sequence is represented in Figures 5a and 5b by two points, corresponding to end of idling and driving period, respectively. Thus, a slope of a curve showing an individual trip is directly related to mean velocity (VBAR). Steep slope relates to slow sequences (group 1), plotted as triangles, while fast sequences (group 2) are presented by square symbols. Figure 5a displays 5 trips, all performed relatively early in the morning (before 9a.m.). They all show almost the same fluent pattern, except for slight congestion at the 3rd and around 8th km from the starting point. More congestion can be viewed on Figure 5b, where 4 trips made in the afternoon and on another route are plotted. Bottleneck areas seem to be at the first km of the route, and around the 4th kilometre. Trip performed at later time (18:33) was much more fluent than those driven at rush hours.

Figure 5a. Morning Trips, Fluent Traffic

Alternately, trips can be characterized in the usual way, by mean velocity, total time, percentage of driving time, etc., as shown in Table 2 for trips plotted in Figure 5b. In addition, we present here also observed transition probabilities between two groups of sequences, for each trip separately. It can be seen that the most fluent trip, driven at 20km/h on average had highest percentage of transitions between the fast sequences, 2 to 2 (30.4%), while the slowest one (started at 16:56) had highest percentage of transitions from slow to slow sequence, 1 to 1 (63.8%).

We found that the trips are best characterized by joining visual and analytical information, such as provided from Figure
5a and Table 2, respectively.

![Graph](image.png)

Figure 5b. Afternoon Trips, Partially Congested Traffic

<table>
<thead>
<tr>
<th>TRIP START TIME</th>
<th>MEAN VELOCITY</th>
<th>MEAN DRIVING VELOCITY</th>
<th>TOTAL TIME (s)</th>
<th>% OF DRIVING TIME</th>
<th>NO. OF CYCLES/KM</th>
<th>NO. OF SEQ/KM</th>
</tr>
</thead>
<tbody>
<tr>
<td>18:33</td>
<td>15.9</td>
<td>20.0</td>
<td>2282</td>
<td>79</td>
<td>1.4</td>
<td>4.7</td>
</tr>
<tr>
<td>17:17</td>
<td>12.7</td>
<td>17.0</td>
<td>3363</td>
<td>75</td>
<td>3.5</td>
<td>6.2</td>
</tr>
<tr>
<td>17:08</td>
<td>10.9</td>
<td>16.7</td>
<td>3336</td>
<td>65</td>
<td>2.9</td>
<td>6.5</td>
</tr>
<tr>
<td>16:56</td>
<td>10.9</td>
<td>15.0</td>
<td>2855</td>
<td>73</td>
<td>3.0</td>
<td>9.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TRIP START TIME</th>
<th>TRANSITION PROBABILITIES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 to 1</td>
</tr>
<tr>
<td>18:33</td>
<td>.41</td>
</tr>
<tr>
<td>17:17</td>
<td>.23</td>
</tr>
<tr>
<td>17:08</td>
<td>.45</td>
</tr>
<tr>
<td>16:56</td>
<td>.64</td>
</tr>
</tbody>
</table>

Table 2. Trips Characterisation

6. Simultaneous Visualisation

Finally, we introduce a concept of simultaneous visualisation, that is planned to be developed in future. It is based on joining visual presentation of traffic, urban map data, video images, and pollution prediction. Thus, simultaneous visualisation is based on the following graphics windows:
Cumulative time and distance plots, for different routes, such as shown in Figures 5a, 5b and 6a,

![Cumulative time and distance plots](image)

*Figure 6a. Afternoon Trips, Unusually Congested Traffic*

Velocity diagrams of individual sequences and cycles, as presented in Figures 2a, 2b and 6b,

![Velocity diagrams](image)

*Figure 6b. A Cycle Corresponding to Unusually Congested Area*
• Urban map, showing the routes performed,
• Video images of the trips, as shown in Figure 7,
• Emission and pollution prediction plots and tables.

The application that will allow for linking of presentation tools listed above, still has to be developed. Linking will be done by distance (or co-ordinate) and time data. Simultaneous visualisation could then be used for viewing:
• on the map, the locations of congested areas, detected by a time and distance plot;
• velocity diagrams of cycles or sequences, corresponding to areas indicated in time and distance plot, such as cycle diagram, shown in Figure 6a, corresponding to the congested area, indicated in Figure 6b;
• video images, as shown in Figure 7, relative to the position on the map, at different times, or to position on time and distance plot, etc.

Figure 7. Video Image of Congested Traffic

All data manipulation, analysis and graphics was performed on a UNIX based SUN Workstation, under OpenWindows (X interface) in Istituto Motori, CNR, Naples, using SAS™ base, with intensive macro facility utilisation, STAT™, GRAPH™ and INSIGHT™ modules. In the future, some of the new SAS™ visualisation facilities will be applied in the integration of recorded traffic video images with the vehicle condition data.

ACKNOWLEDGEMENTS

This project was carried out under the contract within the research project "Urban areas and environment - life quality
in large urban systems", funded by National Research Council.

REFERENCES:


*SAS*, *SAS/GRAPH*, *SAS/STAT* and *SAS/INSIGHT* are registered trademarks of SAS Institute Inc., Cary, NC, USA.

Correspondence Addresses:

**Project:**
Ing. Mario Rapone  
CNR, Istituto Motori  
Viale Marconi 8  
80126 Napoli, Italy  
FAX: (3981) 2396097  
Phone: (3981) 7177114

**Data Analysis, SAS Code:**  
Vesna Luzar, Ph.D. (speaker)  
University Computing Centre  
J. Marohnica bb  
41000 Zagreb, Croatia  
FAX: (38541) 278908  
Phone: (3981) 8041644