CONVENTION ON LONG-RANGE TRANSBOUNDARY AIR POLLUTION
8th JOINT TASK FORCE & EIONET MEETING ON EMISSION INVENTORIES AND PROJECTIONS
23RD OCTOBER – 24TH OCTOBER 2007, DUBLIN (IRELAND)

A statistical methodology for traffic emission temporal profiles evaluation

Cristina Lavecchia[1], Samantha Pilati[1], Elisabetta Angelino[2], Giuseppe Fossati[2]

[1]*Galileo Ambiente s.n.c. – viale Davanzati 5, 20158 Milano* **e-mail: c.lavecchia@galileo-ambiente.it**
[2]*ARPA Lombardia U.O. Modellistica – Viale F.Restelli 1, 20124 Milano* **e-mail: e.angelino@arpalombardia.it**

ABSTRACT

A statistical methodology based on Cluster Analysis has been developed for the evaluation of traffic features in the framework of a study sponsored by Province of Milan – Environment Offices (Direzione Centrale Risorse Ambientali - Settore Affari Generali Ambiente).
The methodology has been applied to data collected over the main roads of the province of Milan to determine:

✓ Types of average yearly vehicle composition in classes at 08:00 - 09:00 AM of a working day (motorcycle, car, light-duty vehicle, heavy-duty vehicle, heavy-duty vehicle > 16 ton)

✓ Types of hourly traffic flow for each vehicle class in any season and day typology (working day, day before holiday, holiday)

The statistical representativeness of the temporal profiles has been evaluated by the analysis of parameters such as internal homogeneity of groups and dendrogram cut linkage distance. The results of this study has been applied to improve the road traffic emission inventory for the province of Milan. Besides the study has given a chance of implementing a methodological proposal that can be shared by other researchers.

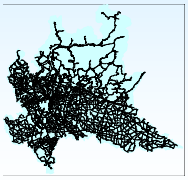
PROLEM DOMAIN AND DATA UNDESTANDING

The following databases have been looked up:

- Traffic detection campaigns worked out by Province of Milan - Traffic Offices
  The data refer to the period from april 2005 to march 2006 and concern 54 sites of the provincial road network, 34 of which operating during every season (figure on the right side).
  For every lane and for every hour the following data have been collected:
  → number of vehicles
  → class length of the vehicle among 8 length categories
  → average speed of each length category

  Critical points are:
  - short representativeness of the detected speeds as regards the traffic congestion typical of the province of Milan. In fact the location of monitoring sites has to be away from traffic light and area subjected to queue formation to prevent malfunctioning in the counting system. It prevents the possibility of inferring the velocity as a function of the traffic flux
  - underestimates of motorcycles as they often travel away from the detection area

- INEMAR (Air Emission Inventory of Lombardy Region) by ARPA Lombardia
  This dataset collects data on 18000 oriented links (sections of road) of the road network about:
  - type of road (highways, statal roads, primary/secondary regional roads, primary/secondary provincial roads),
  - Maximum speed
  - street capacity
  - flux of equivalent vehicles at 8:00 – 9:00 AM



The following steps are necessary to estimate traffic emissions in INEMAR:

1. to evaluate the fluxes of each vehicle class as a function of equivalent vehicles at 08:00 AM in a working day
2. to evaluate hourly flow trend for each vehicle class in a type of day (working day, day before holiday and holiday) of every season

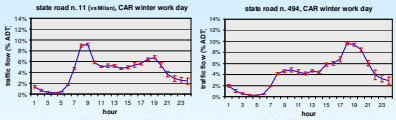
METADATA

The following metadata have been calculated on 63 oriented sections of road in relation to each flow direction, day type, season and vehicle class:

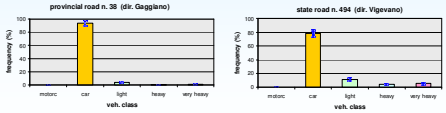
✓ **the normalized trend of hourly vehicular flow**, that is a set of 24 values computed according to the following equation:

$$x_{nmjk} = \frac{F_{nmjk}}{TGM_{mk}} \cdot 100$$

Fnmjk = number of hourly vehicle transits at hour *n* (n=1,24) for vehicle class *m* (m=1,5), daytype *j* (j=1,3) and season *k* (k=1,4)
TGM*mk* = average daily traffic flow on working days for vehicle calss *m* and season *k*



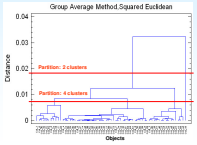
✓ **the yearly average vehicle composition** in 5 classes (motorcycle, car, light-duty vehicle, heavy-duty vehicle, heavy-duty vehicle > 16 ton) at 8:00 – 9:00 AM on working days



They turned out to be a good representation of vehicular dynamic for the respective average period (15 days campaign at season in regard to traffic flow, the whole year for vehicle composition). The standard deviations are really small.

CLUSTER ANALYSIS

The Cluster Analysis is a set of multivariate statistical techniques that can be applied to observations to find out "natural" groupings. The aim is to sort the observations into clusters so that the degree of association is strong between members of the same cluster and weak between members of different clusters, according to some defined distance measure. If the observations or objects are described by means of data, the distance function can be derived by statistics.
Data clustering algorithms can be hierarchical or non hierarchical.
The **hierarchical methods** start by treating each object as a separate cluster, then group them into bigger and bigger clusters. At each grouping between two clusters, the new distances among the emerging groups are computed. Hierarchical algorithms differ in the definition of the distance measure between clusters (i.e. either maximum or minimum or average distance between cluster objects).
Hierarchical methods give a series of possible partitions or groupings graphically depicted by a **dendrogram**.
The distance level to cut the dendrogram, e.g. the clusters number, is a choice of the analyst and implies a certain degree of subjectivity .



METHODS

The application of Cluster Analysis has being founded on the assumption that traffic flow is repetitive with respect to time and space.
We have used a hierarchical clustering method. Different combination between grouping method (7 methods) and linkage distance (3 functions) were tested. The comparison among the results has suggested the implementation of ***Group Average method*** and ***Squared Euclidean Distance***: the distance between two clusters is computed as average squared euclidean distance among all the objects of the two clusters. The function of squred euclidean distance is:

$$\left(\sum (x_i - y_i)^2\right)$$

The method was applied separately to the two dataset of objects reported in the box on the right side ("Vehicle composition" and "Hourly traffic flow").

With respect to the "Hourly traffic flow" dataset, the aim is to find typologies that stand for objects having similar trend for each vehicle class given the season and the day type. The Cluster Analysis was applied to the CAR class, that is the class prevailing as regards the vehicles number. Then it was verified if the clusters for car were significant also for the other vehicle classes: it was truth for all the classes, with the exception of heavy-duty vehicle>16t. Therefore it was performed a further partition of the car types by applying again the Cluster Analysis to the heavy-duty vehicle class (>16t).
It was characterized each cluster by means of the prevailing technical features of the road members and the respective ***centroid***, that is the mean value (variable by variable) for all the objects in the cluster.
The centroids represent the types to assign to each links in the road network of INEMAR database.

The objects for clustering are N-dimensional vectors

Tables format of data preparation for clustering

Average yearly vehicle composition in classes (N=5):

road \ variable	motorcycle (%)	car (%)	light-duty veh. (%)	heavy-duty veh. (%)	heavy-duty veh. >16t (%)
road 01	0.4	91.0	5.4	1.3	2.0
road 02	0.8	85.5	8.4	2.1	3.2
…	…	…	…	…	…
road 63	0.0	78.2	11.6	4.6	5.6

Hourly traffic flow (N=24):

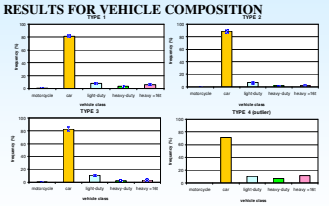
road \ variable	flux 01:00 (%ADT)	flux 02:00 (%ADT)	…	flux 23:00 (%ADT)	flux 24:00 (%ADT)
road 01	1.4	0.7	…	2.6	2.4
road 02	0.6	0.2	…	1.6	1.2
…	…	…	…	…	…
road 63	1.1	0.5	…	2.6	2.1

In the box above there is an example of the physical meaning of the objects. It's shown also, by means of standard deviations, the representativeness of the objects with respect to the average period (i.e. season).

RESULTS

The number of clusters resulting from analysis is small compared to the dimension of the data set and the types identified have physical meaning, that is they represent traffic load trends and features unlike each other, from both temporal and spatial point of you.
Every centroid presents standard deviations of each variable smaller than the value of the variable itself. It ensures there are not overlapping among clusters, that is the groups are separated each other and homogeneous inside (the degree of association between objects is maximal if they belong to the same group and minimal otherwise).

RESULTS FOR VEHICLE COMPOSITION



	centroids (%)			
	type 1	type 2	type 3	type 4
motorcycle	0.4	0.2	0.3	0.1
car	81	89	82	71
light-duty veh.	8	7	11	11
heavy-duty veh.	4	2	3	7
heavy-duty>16t	6	2	4	11
	CLUSTER COHESION (st.dv. %)			
	type 1	type 2	type 3	type 4
motorcycle	0.5	0.2	0.4	0.0
car	2.2	2.2	2.7	0.0
light-duty veh.	0.2	1.4	0.8	0.0
heavy-duty veh.	0.8	0.6	1.0	0.0
heavy-duty>16t	0.9	1.0	1.3	0.0
total	5.4	6.4	6.2	0.0

Main features:

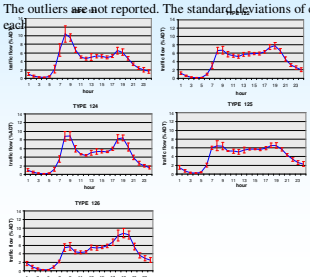
CLUSTER 1: main roads, average speed 50-60 km/h and average traffic flow 1000 veh./h
CLUSTER 2: minor provincial roads, average speed 50-60 km/h and average traffic flow <1000 veh./h
CLUSTER 3: main and state roads, average speed >70 km/h and average traffic flow 1200 veic/h
CLUSTER 4: outlier (one main road), average speed 70 km/h and average traffic flow 1000 veic/h

RESULTS FOR TEMPORAL PATTERN

Working days: 6 – 7 types according to season. Some clusters are outliers, that is they represent only one o two items.
Days before a holiday: 5 – 6 types according to season, but only 10 clusters in the year have enough members to be considered representative.
Holidays: 2 – 3 types a season representatives because of the large number of members.

Example for CARS in winter working day (Cluster Analysis)

The outliers are not reported. The standard deviations of centroids are displayed as index of the homogeneity of each.

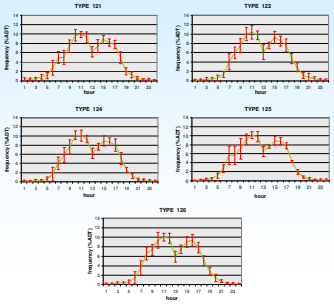


	centroids					cluster cohesion				
hour	121	122	124	125	126	121	122	124	125	126
1	1	1	1	1	2	0.4	0.3	0.3	0.4	0.5
2	1	1	1	1	1	0.3	0.2	0.2	0.3	0.4
3	0	0	0	0	0	0.1	0.1	0.1	0.1	0.2
4	0	0	0	0	0	0.1	0.1	0.1	0.1	0.1
5	0	0	0	0	0	0.2	0.1	0.1	0.1	0.1
6	2	1	1	2	1	0.6	0.2	0.3	0.3	0.3
7	3	3	4	5	2	1.5	0.7	0.7	0.4	0.3
8	10	7	9	7	5	1.9	0.8	1.1	1.1	0.9
9	10	7	9	6	6	0.8	0.9	1.0	0.8	0.9
10	7	6	6	5	4	0.6	0.7	0.7	0.1	0.5
11	5	5	5	5	4	0.3	0.4	0.2	0.5	0.3
12	5	5	4	5	4	0.3	0.4	0.2	0.6	0.3
13	5	6	5	6	5	0.8	0.5	0.6	0.7	0.6
14	5	6	5	6	5	0.4	0.5	0.6	0.2	0.5
15	5	5	5	5	5	0.4	0.3	0.4	0.2	0.3
16	5	6	5	6	6	0.3	0.4	0.2	0.3	0.5
17	5	6	6	6	7	0.4	0.3	0.4	0.4	0.7
18	7	7	8	7	8	0.8	0.6	1.0	0.5	1.3
19	6	6	8	7	9	0.7	0.7	0.7	0.6	1.1
20	5	7	6	6	8	0.5	0.6	0.9	0.4	0.9
21	3	5	4	4	6	0.3	0.5	0.5	0.3	0.8
22	2	3	3	3	4	0.3	0.3	0.4	0.4	0.6
23	2	3	2	3	3	0.5	0.4	0.3	0.5	0.6
24	2	2	2	2	3	0.4	0.4	0.3	0.5	0.6
day						13.0	10.4	11.5	9.7	13.3

RESULTS FOR TEMPORAL PATTERN

Example for HEAVY DUTY VEHICLES in winter working day

The clusters come from Cluster Analysis applied to cars (see box "Methodology"). The centroid standard deviations are greater than the car ones.



ADDITIONAL COMMENTS

Use:
- bottom-up estimates of traffic atmospheric emissions
- temporal top-down disaggregation of road transport emissions
- data input of air dispersion models to improve the estimates of the concentration of pollutants at local urban scale
- implementation of temporal pattern according to a "standard" and "shared" methodology.

Developments:
- hourly resolution of traffic flow – application of Cluster Analysis to each vehicle class;
- vehicle composition – analysis of seasonal trends;
- application of methodology to other years databases: the estimates can be compared to each other to answer the question "How many years the patterns are representative for ?";
- better characterization of uncertainty and propagation of errors (the amount of input activity data uncertainties passed on the uncertainty of traffic emissions estimates);
- speed variation with traffic load: study the problems related to speed measurements in more depth (to infer representative velocity).

Verification procedures:
- air quality modelling tests to infer the influence of each activity data (temporal patterns, vehicle composition) on model outputs:
- air quality modelling to evaluate the benefits from detailed activity data on a better estimation of air pollution concentrations (temporal variations).

Workshop on Uncertainties in emission inventories and atmospheric models held jointly with the Task Force on Measurements and Modelling (TFMM) – 22nd October