DEVELOPMENT OF OPTIMIZED ALGORITHMS FOR THE CLASSIFICATION OF NETWORKS OF ROAD STRETCHES INTO HOMOGENEOUS CLUSTERS IN URBAN AREAS

Giovanni Zambon, Roberto Benocci, Alessandro Bisceglie

Dipartimento di Scienze dell'Ambiente e del Territorio e di Scienze della Terra-Università degli Studi di Milano–Bicocca, p.zza della Scienza 1, 20126 Milano, Italy.
giovanni.zambon@unimib.it

Noise maps are considered as a powerful tool to determine the population exposure to environmental noise. A statistical approach to real-time noise mapping will be developed in DYNAMAP (Dynamic Acoustic Mapping), a co-founded project in the framework of LIFE 2013 program. The main preliminary action of the project is to define a statistically-based method to optimize the choice and the number of monitoring sites, which will provide the information to update the dynamic mapping process. In this work, preliminary results referring to a sample of roads of the city of Milan are presented. The sample database is made of 24 hour continuous acoustic monitoring of the hourly equivalent levels L_{Aeqh} in different sites, corresponding to 8 road functional classifications (from A to F and sub-classes). Once normalized, such trend profile provides a tool to group roads by their vehicular dynamics. Acoustic trend profiles will be also studied on a shorter time basis, with the aim of identifying road clusters that allow an updating of the map with an higher time frequency. Linking a non-acoustical parameter (hourly traffic flow) to the elements in each cluster represents the key-issue which allows each road segment of the urban network to be univocally assigned to the obtained clusters.

1. Introduction

Urban traffic noise has been the object of several studies aimed at investigating different aspects of its impact [1-6]. Initially, the environmental noise has been studied by using systematic sampling, that is selecting measurements points by the use of grids over a map [2]. However, this approach showed to be time and cost consuming for road administrations and local or central authorities, as well as to give more weight to noisier streets [7], thus providing biased maps. In fact, the noise on a street generally depends on its activity, the use in the urban context, width, presence of reflecting surfaces, presence of obstacles, type of paving, etc.. Acoustic simulation algorithms, implemented by software, allow to reproduce noise emission and propagation on a wide area, starting from some static information about sound sources and environment. Dynamap project has the aim to develop a dynamic approach to noise mapping, able to update environmental noise levels through a direct link with a limited number of noise monitoring terminals. Hence, the need to group road network stretches in homogeneous clusters represent a possible method to size the network of monitoring terminals. Roads sharing the same characteristics for some parameters such as vehicles' flow capacity, number of lanes, etc., are grouped together. Such parameters are usually included in the functional classification of roads and linked to the role played in the urban mobility. However, this classification generally does not reflect the actual use of roads and, therefore, the actual noise emission. For a better description of the real behavior of noise in complex scenarios such as the road network of the city of Milan, we approached the problem considering an agglomeration method based upon similarities among the 24-h continuous acoustic monitoring of the hourly equivalent L_{Aeqh} levels. Once normalized, such trend profile provides a tool to group together roads according to their vehicular dynamics, therefore allowing a more real description of such road networks.

2. Acoustic level profiles

The dataset considered in the present work refers to the city of Milan, Italy, and is made of 138 24-h continuous acoustic monitoring of the hourly equivalent levels L_{Aeqh} in 58 different sites corresponding to 8 functional road classes (from A to F and sub-groups). Sub-groups belonging to classes E and F were merged. Data were recorded on weekdays and in absence of rain as prescribed by D.M. Ambiente 16/3/1998 [8]. Because of the non-homogeneity of L_{Aeqh} level dataset, due to different monitoring conditions such as different distances from the road but also to the condition of the street itself (its geometry, the presence of reflecting surfaces and obstacles in sound propagation and types of paving), we referred each i^{th} hourly L_{Aeqhij} level of the j^{th} temporal series to the daytime reference level, L_{Aeqdj}:

$$(1)\ \delta_{ij} = L_{Aeqh_{ij}} - L_{Aeqdj}\ \text{[dB]} \qquad (i = 1\text{ h}, \ldots, 24\text{ h};\ j = 1, \ldots, 58)$$

The normalization referred to the daytime L_{Aeqd} level was chosen because this descriptor is, in general, more often available than the nighttime L_{Aeqn} value. For all 58 sites, the rush-hour (time interval 7:30 a.m.-8:30 a.m.) and the night minimum (time interval 2:30 a.m.-3:30 a.m.) vehicle flow rate was available too. In 32 sites, monitoring periods extended over more days. In such cases the median of δ_{ij} hourly values was calculated. The median was chosen as this index is less influenced by the presence of outliers. Figure 1 illustrates the 24-hour mean profiles $\bar{\delta}_{im}$ (green line) and the corresponding ± the standard error of the mean for each road functional class (light green area). Due to the poor sample size (3 profiles), category A roads present higher standard error.

3. Statistical Analysis

The functional classification of roads generally does not reflect their actual use, that is the 24-h hourly L_{Aeqh} level profiles might be extremely different for roads belonging to the same category. In fact, such difference mostly depends on the activity of each road in its urban context. For this reason and as suggested in [9-10], we chose to explore our dataset by means of a cluster analysis.

For this purpose, unsupervised clustering algorithms were employed to group together level profiles found to be "close" to one another. Various algorithms (hierarchical agglomeration using Ward algorithm [11], K-means algorithm [12], Partitioning Around Medoids [13], Expectation Maximization algorithm by "mclust" module [14]) were considered, and their results compared. The number of clusters was chosen as a compromise between satisfactory discrimination and the need to limit the number of groups. The range of solutions for clustering was set from four groups (for a straightforward comparison with the number of road categories considered) to two (corresponding to the minimal discrimination). Euclidean distance was chosen as the metric of the distance among observations.

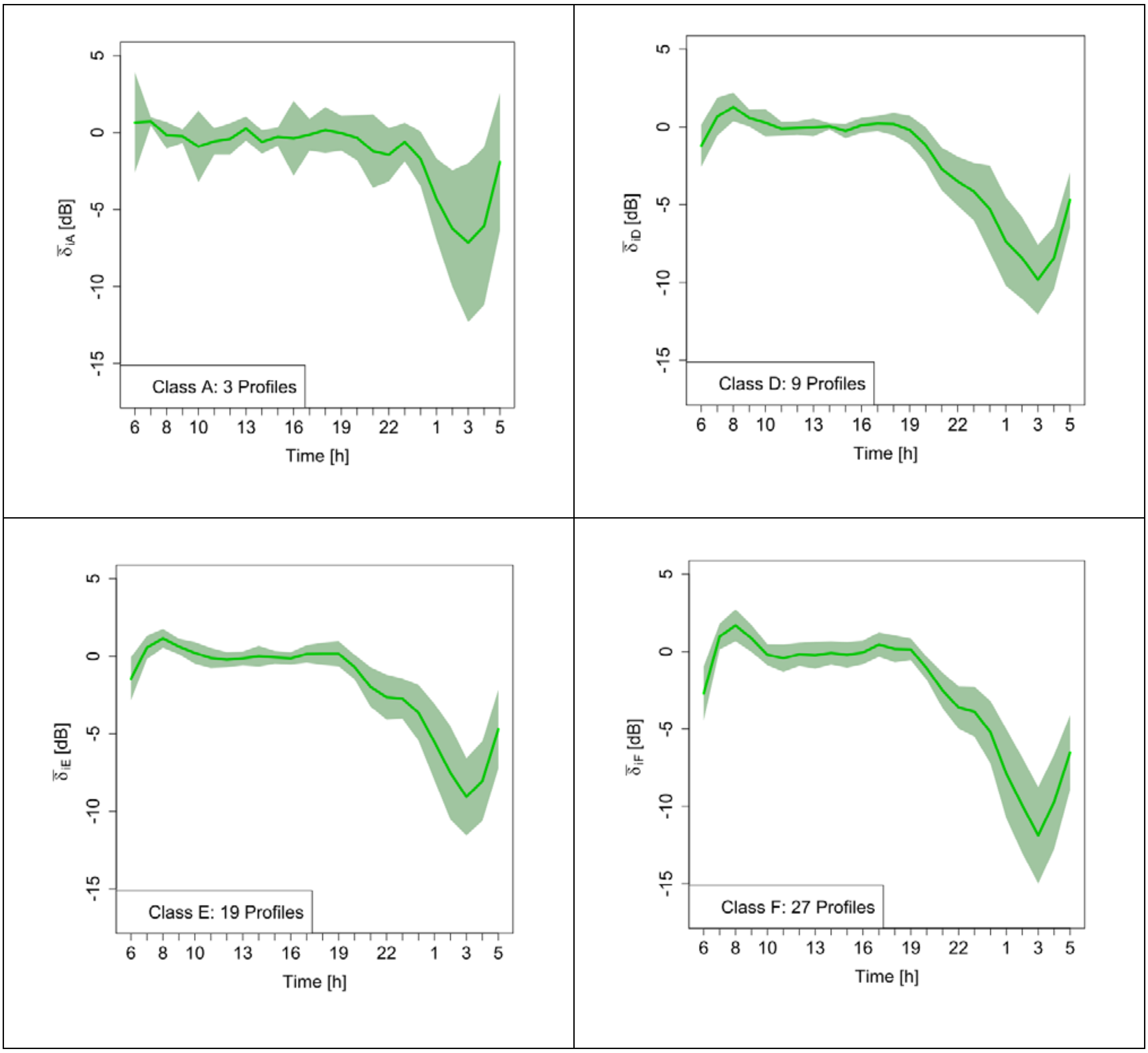

Figure 1. 24-hour mean profiles $\bar{\delta}_{im}$ (green line) and the corresponding ± standard error of the mean for each road functional class (light green area).

The statistical software R, a free software environment for statistical computing and graphics, was applied for the clustering. The package "clValid" [15-16] was used for validating the results and assess the quality of the clustering. All the clustering algorithms were ranked based on their performance as determined simultaneously by all the validation measures [17]. Thus, the optimal list, obtained through a score assigned by each validation index, gives a two-cluster K-means agglomeration at the first place followed by PAM and hierarchical methods, each one yielding also a two-cluster separation. The two-cluster groups represent a satisfying balance between an adequate differentiation among profiles and the need to get a simple practical solution. Therefore, there exists the possibility of naturally grouping the 24-h average profiles $\bar{\delta}_{ik}$ according to their shape. The obtained clusters were composed of roads belonging to different categories as reported in Tab. 1.

The two clusters appeared to be composed primarily of contributions from different temporal profiles belonging mainly to roads of category D and F for Cluster 1 (made up of 31 temporal profiles corresponding to 53.4% of total) and to roads of category A and E for Cluster 2 (made up of 27 temporal profiles corresponding to 46.6% of total).

Table 1: Composition of clusters.

Cluster	**Road Category**				**Total**
	A	**D**	**E**	**F**	
1	1 (33.3%)	5 (55.6 %)	6 (31.6%)	19 (70.4%)	31
2	2 (66.7%)	4 (44.4%)	13 (68.4%)	8 (29.6%)	27

This confirms that road traffic is primarily linked to the effective urban mobility use rather than its functional classification, as shown by the outcomes of previous studies [9]. Figure 2 shows the profiles of mean values $\overline{\delta}_{ik}$ and the corresponding ± the standard error of the mean for each cluster. Cluster 1 (blue line) presents two peaks: the first in the time interval 8-9 h and the second at 17 h. It fluctuates closely around the L_{Aeqd} until 19 h, afterwards it goes down in the night period till 3 h after which it starts raising again. Cluster 2 has just one lower peak at 8-9 h and higher values at nighttime. In the remaining time period, it shows a similar behavior of Cluster 1.

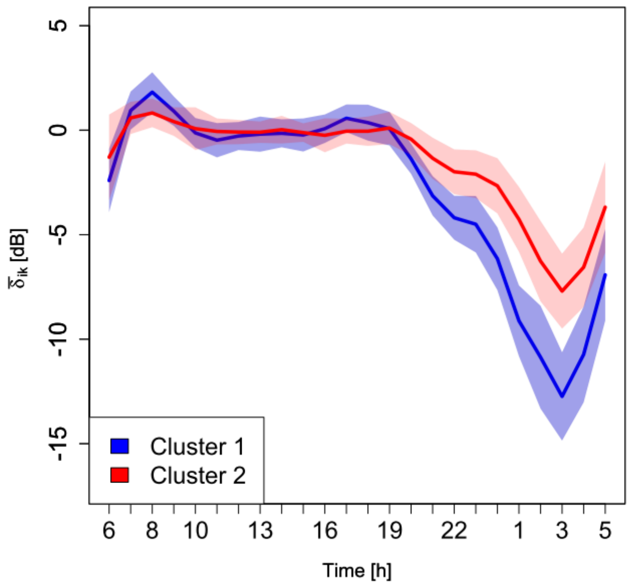

Figure 2: Mean values of $\overline{\delta}_{ik}$ and their standard error for each cluster.

Unlike the functional classification of roads, the two obtained cluster profiles cannot be applied straightforward to the whole network without any indication, which link them to a specific feature. To overcome such limitation, each mean cluster profile was associated with the corresponding traffic flow rate at rush hour (TFRH) and the night minimum vehicle flow rate (NMVF) for each of the 58 roads under consideration. Figures 3-6 show the probability density and the box plots for these parameters for the two-mean-cluster profiles. In particular, for the TFRH parameter we can observe that it presents quite separate density distributions. In addition, the interquartile range of the two clusters does not overlap. We can, therefore, consider a vehicular flow rate at rush hour of 2000 vehicles/hour as threshold between the two profiles, that is roads featuring lower values (<2000 vehicles/hour) can be associated with cluster 1 whereas higher flow rates (>2000 vehicles/hour) with cluster 2. In the case of NMVF parameter, the density distributions present different behaviors:

cluster 1 shows a sharp profile centered around zero vehicles per hour, whereas cluster 2 shows a flatter distribution though peaked at higher values. The corresponding boxplot gives distinct inter-quartile ranges for the two clusters. In this case, the threshold value between the clusters is around 40 vehicles per hour.

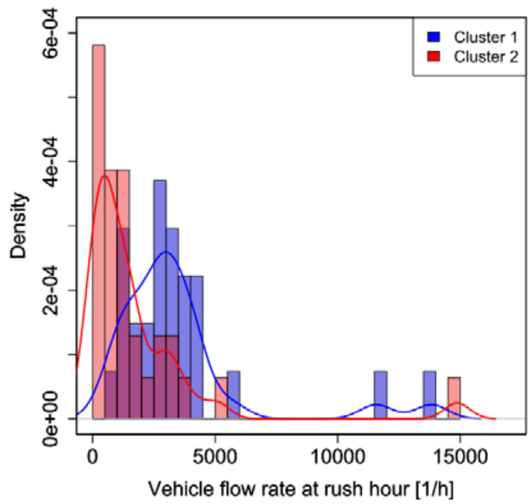

Figure 3: Histogram and probability density vs. vehicle flow rate at rush hour for the K-means two-cluster results.

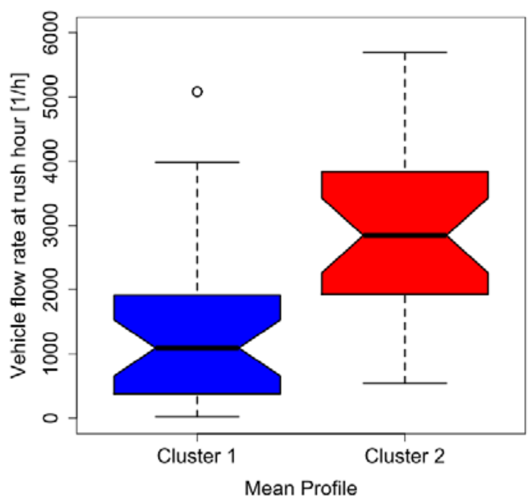

Figure 4: Box plots of the traffic flow rate at rush hour for the two mean cluster profiles.

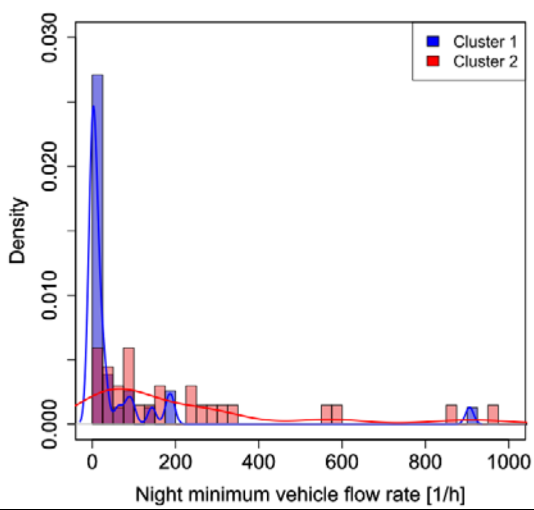

Figure 5: Histogram and probability density vs. night minimum vehicle flow rate for the K-means two-cluster results.

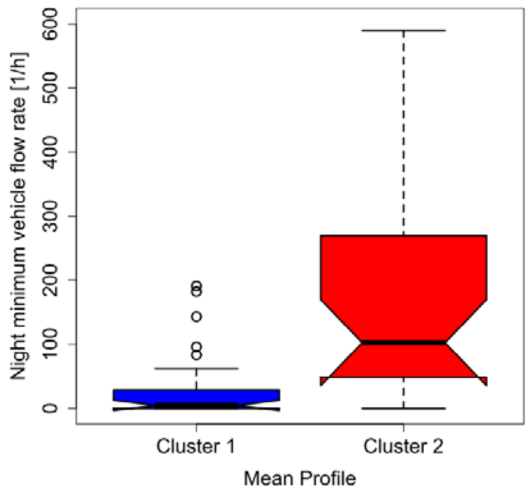

Figure 6: Box plots of the night minimum vehicle flow rate for the two mean cluster profiles.

4. Comparative analysis among profiles of different temporal discretization

Another interesting issue related to noise mapping regards the smallest time interval a noise map can be updated without losing significant information from the original data (hourly levels). To this purpose, we extracted five new level profiles with temporal resolution of 30, 20, 15, 10, 5 minutes. Unfortunately, only a sub-set of the original data was available for this operation and, therefore, each new dataset was made up of 36 sites.

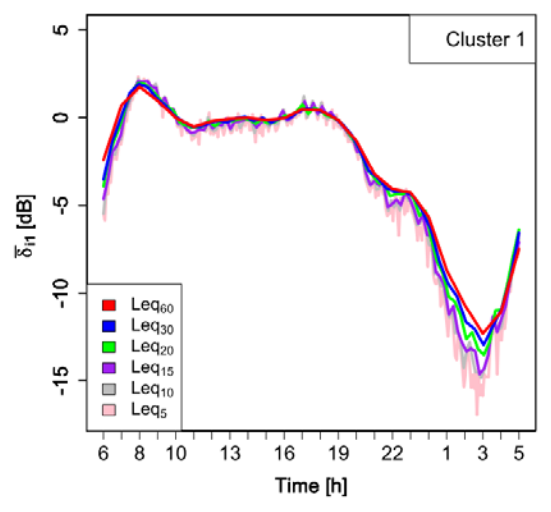

Figure 7: Comparison between the mean profiles $\bar{\delta}_{i1}$ for cluster 1 with different temporal discretization.

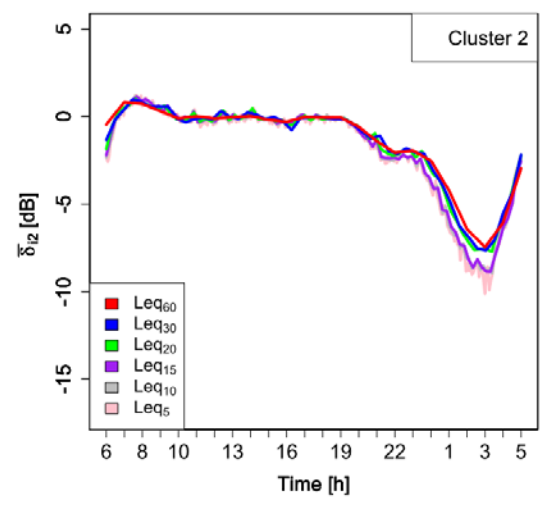

Figure 8: Comparison between the mean profiles $\bar{\delta}_{i2}$ for cluster 2 with different temporal discretization.

Table 2: Composition of clusters for different temporal discretization**.**

Temporal Discretization [min.]	**Cluster**	**Road Category**				**Total**
		A	**D**	**E**	**F**	
60	**1**	1 (33.3%)	3 (75.0%)	3 (27.3%)	12 (66.7%)	19
	2	2 (66.7%)	1 (25.0%)	8 (72.7%)	6 (33.3%)	17
30	**1**	1 (33.3%)	3 (75.0%)	3 (25.0%)	12 (70.6%)	19
	2	2 (66.7%)	1 (25.0%)	9 (75.0%)	5 (29.4%)	17
20	**1**	1 (33.3%)	3 (75.0%)	3 (25.0%)	11 (64.7%)	18
	2	2 (66.7%)	1 (25.0%)	9 (75.0%)	6 (35.3%)	18
15	**1**	0 (0.0%)	0 (0.0%)	3 (25.0%)	10 (64.7%)	13
	2	3 (100.0%)	4 (100.0%)	9 (75.0%)	7 (35.3%)	23
10	**1**	0 (0.0%)	0 (0.0%)	3 (25.00%)	10 (58.8%)	13
	2	3 (100.0%)	4 (100.0%)	9 (75.0%)	7 (41.2%)	23
5	**1**	0 (0.0%)	0 (0.0%)	3 (25.0%)	10 (58.8%)	13
	2	3 (100.0%)	4 (100.0%)	9 (75.0%)	7 (41.2%)	23

Each level profile with different temporal discretization was statistically analyzed and the results of the mean values, $\overline{\delta}_{ik}$, for cluster 1 and 2 are shown in figure 7 and 8. The “high resolution” temporal profiles (5, 10 and 15 min.) present quite different behavior when compared to the 60, 30 and 20 min. profiles especially in the nighttime period. This is due to the statistic process of clustering which gives different composition of the two clusters for different temporal discretization (see Tab. 2). For temporal discretization of 30 and 20 min. the composition of clusters is quite stable. In particular, cluster 1 is mainly made of roads of category belonging to classes D and F, whereas cluster 2 to classes A and E.

5. Conclusions

A completely “blind” approach has been used to analyze the 24-h continuous acoustic monitoring of the hourly equivalent levels L_{Aeqh} of different road categories aiming at searching for a better classification criterion of such profiles that reflected the actual use of roads.

The cluster analysis approach showed that the dataset of measurements can be suitably grouped into two-mean profiles to be applied to roads with vehicular flow rate less (Cluster 1) and greater (Cluster 2) than 2000 vehicles/hour at rush hour and about 40 vehicles/hour at the night minimum.

These profiles show also a different composition of the two clusters proving a loss of stability of the noise profiles for temporal resolution of 15, 10, 5 minutes.

REFERENCES

1 S. Fidel; Nationwide urban noise survey, *J. Acoust. Soc. Am*., **64**, 198-206, 1978.

2 A. L. Brown, K. C. Lam, Urban noise surveys, *Appl. Acoust*., **20**, 23-39, 1987.

3 K. Kumar, V. K. Jain, A study of noise in various modes of transports in Delhi, *Appl. Acoust*., **43**, 57-56, 1994.

4 H. M. E. Miedema, H. Vos, Exposure-response relationships for transportation noise, *J. Acoust. Soc. Am*. **104**, 3432-3445, 1998.

5 A. Garcia, L. J. Faus, Statistical analysis of noise levels in urban areas, *Appl. Acoust*., **34**, 227-247, 1991.

6 J. M. Fields, Effect of personal and situational variables on noise annoyance in residential areas, *J. Acoust. Soc. Am*., **93**, 2753-2763, 1993.

7 J. M. Barrigon, V. Gomez, J. Mendez, R. Vilchez, J. Trujillo, A categorization method applied to the study of urban road traffic noise, *J. Acoust. Soc. Am*., **117**, 2844-2852, 2005.

8 Decreto Ministero dell’Ambiente 16 marzo 1998, Tecniche di rilevamento e di misurazione dell’inquinamento acustico, *Gazzetta Ufficiale serie generale* n. 76, 1.4.1998.

9 G. Brambilla, V. Gallo, Andamenti dei livelli L_{Aeq} orari nelle 24 ore del rumore urbano e indicazioni per il campionamento spaziale stratificato, *Atti AIA 2010*, Siracusa, 2010.

10 F. Angelini, A. Bisceglie, G. Brambilla, V. Gallo, G. Zambon, Campionamento spaziale stratificato per il rumore da traffico stradale: un’applicazione alla rete viaria di Milano, *Atti AIA 2012*, Roma, 2012.

11 J. H. Ward, Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, **58**, 236–244, 1963.

12 J.A. Hartigan, M.A. Wong, A K-means clustering algorithm, *Applied Statistics* **28**, 100–108, 1979.

13 L. Kaufman, P. Rousseeuw, *Finding Groups in Data*, Wiley Series in Probability and Mathematical Statistics, 1990.

14 C. Fraley et al., "mclust" version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation, [Online.] available: http://cran.r-project.org/web/packages/mclust/index.html

15 G. Brock, V. Pihur, S. Datta, S. Datta, clValid: An R Package for Cluster Validation, *Journal of Statistical Software* **25**(4), 1-22, 2008.

16 Package "clValid" version 0.6-4, 2013.

17 V. Pihur, S. Datta, S. Datta, Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach, *Bioinformatics*, **23**(13), 1607-1615, 2007.