Workshops at 18th International Conference on Intelligent Environments (IE2022)
H.H. Alvarez Valera and M. Luštrek (Eds.)

doi:10.3233/AISE220037

On the Application of Unsupervised Clustering to Sound Pressure Data from an Acoustic Sensors Network

Antonio PITA [a], Francisco J. RODRIGUEZ [a] Juan M. NAVARRO [a]

[a] *Research Group in Advanced Telecommunications (GRITA), Universidad Católica de Murcia (UCAM), 30107 Guadalupe, Spain*

Abstract. Many cities around the world are deploying wireless sensor networks to capture information on different environmental parameters. Noise, as one of the main pollutants with negative effects on health and economy, is monitored through sound pressure level. In this work, the application of unsupervised clustering to sound pressure level data from a wireless acoustic sensors network (WASN) is proposed. Data from a sensor network deployed in the city of Madrid are used to show the usefulness of performing a clustering process with the aim of detecting different patterns of behavior of noise levels. The preliminary results obtained have allowed us to divide the city into several acoustic zones, which help city managers to propose improvement plans.

Keywords. Environmental Noise Assessment; Machine Learning; Urban Acoustic Environment; Artificial Intelligence of Things; Unsupervised Identification; Internet of Things

Introduction

The European Directive 2002/49/EC aims to establish a common approach aimed at avoiding, preventing or reducing, as a priority, the harmful effects, including annoyances, of exposure to environmental noise [1].

To this end, it urges the rulers of cities to determine exposure to environmental noise, make this information available to the population and adopt action plans. The objective of these action plans is to prevent and reduce environmental noise when it has harmful effects on human health and to maintain quality of the acoustic environment when this is satisfactory.

To reach this objective, many cities are deploying a wireless acoustic sensors network (WASN) to ensure the gathering of noise data that will be analysed and used to design an action plan. Also, this data could be available in open data portals to the citizens.

As different nodes should have different noise behaviors, the same strategy should not works in all the nodes. Unsupervised learning techniques could help grouping the nodes with the same behavior in clusters to allow cities to identify these behaviors and establishing personalized strategies for each cluster.

Unsupervised machine learning techniques including clustering and dimensionality reduction have been used to optimize the choice and the number of monitoring sites [2]. Using hourly averaged $L_{\mathrm{Aeq}1h}$ acoustic data of a 24 h measurement campaign in the city of Milan, Italy, a methodology for a more efficient way to estimate the mean L_{d} and L_{n} levels in urban roads compared with the legislative road classification [3] was presented. Moreover, in order to associate each of the streets of the pilot zone with one of the two noise profiles detected in the clustering and then calculate the dynamic map, different non-acoustic parameters were evaluated [3]. Recently, the intermittency ratio indicator was combined with the $L_{\mathrm{Aeq}1h}$ data to improve the classification of different types of road in two identified clusters [4].

In this research, clustering techniques are proposed for the analysis of urban noise pollution of the city of Madrid in order to identify and classify different urban acoustic behaviors, rather than only road traffic zones. These behavioral groups will help municipalities to monitor the sound contamination, estabish personalized action plans for each behavior and evaluate the noise pollution actions plans in each behavioral group allowing city council to manage the noise pollution depending of the behavior of the noise not only the sound level pressure or the source of the noise, that are the usual ways to control the noise pollution.

Materials and Methods

In this section, first, the dataset is presented and described. Second, the data analysis and transformations are explained and summarized in some graphs. Third, the clustering techniques are presented and the selection technique evaluation is developed. Finally, the software and technology used in this research is enumerate.

Dataset

The Acoustic Pollution Monitoring Network of the city of Madrid has 31 permanent stations in charge of the control and continuous monitoring of the existing noise levels [5].

The sound pressure level measurement dataset of these stations was retrieved from the acoustic pollution section in the Madrid council's open data portal [6]. The available data provides long-term analysis, from January 2014 until December 2021. The location of these fixed stations is shown in Figure 1.

The sound pressure $p(t)$ is usually measured continuously over a given time period $T = [t_1, t_2]$ for all $t \in T$, to quantify the sound level on a single value using the equivalent sound pressure level in dB, denoted as $L_{\mathrm{eq}T}$ [7],

$$L_{\mathrm{eq}T} = 10 \cdot \log \left[\frac{1}{T} \int_{t_1}^{t_2} \frac{p^2(t)}{p_0^2} \mathrm{d}t \right] \text{ where } T = t_2 - t_1, \tag{1}$$

and p_0 is the sound pressure reference value equal to 20 μPa. In particular, deployed nodes compute the A frequency-weighting equivalent sound pressure level of one minute period, denoted as $L_{\mathrm{Aeq}1m}$ in dBA unit, applying Equation (1).

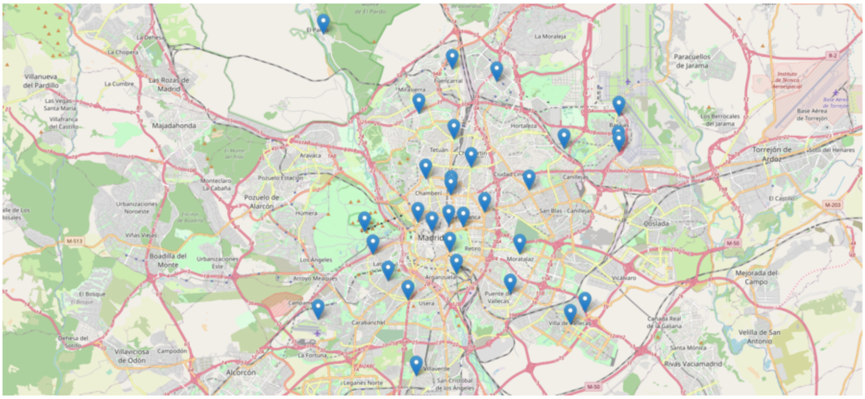


Figure 1. Map showing the location of the 31 acoustic nodes deployed in the city of Madrid, Spain.

In this work, sound pressure level results are presented applying a long-term average of $L_{\mathrm{Aeq1}m}$. Different time periods T can be defined, for instance it is denoted as $L_{\mathrm{Aeq1}d}$ for a 24-h day period and $L_{\mathrm{Aeq1}y}$ for a generic year period. Moreover, the equivalent sound pressure level in a specific year Y is denoted as $L_{\mathrm{Aeq}Y}$, for instance L_{Aeq2020} represents the equivalent sound pressure level for 2020. These values are calculated using an energetic average with the following equation [7],

$$L_{\mathrm{Aeq}T} = 10 \cdot \log \left[\frac{1}{n} \sum_{i=1}^{n} 10^{\frac{L_{\mathrm{Aeq}_i}}{10}} \right], \tag{2}$$

where n is the total number of 1-unit time intervals in period T and L_{Aeq_i} is the equivalent sound pressure level in the interval i obtained by the sensor applying Equation (1). For instance, to calculate $L_{\mathrm{Aeq1}h}$, 60 values of $L_{\mathrm{Aeq1}m}$ are averaged.

The data provided by the Madrid city council contains acquired data from January 2014 until December 2021, downloaded from the Acoustic Pollution repository in the Madrid's open data platform [6]. This data is spread in annual flat files in a semicolon tabulated format with a total storage size of 16,7 MBytes which contains several daily or a period of the day acoustics indicators calculated regarding Directive 2002/49/EC [1]. This Directive [1] establishes that member states must calculate the acoustic parameters L_{den} and L_{night} for the preparation and revision of the Strategic Noise Map (SNM).

L_{den}, defined in Equation (3), refers to the day-evening-night noise indicator obtained for an overall assessment period, usually one year period.

$$L_{\mathrm{den}} = 10 \cdot \log \left[\frac{1}{24} \left(12 \cdot 10^{\frac{L_{\mathrm{day}}}{10}} + 4 \cdot 10^{\frac{L_{\mathrm{evening}}+5}{10}} + 8 \cdot 10^{\frac{L_{\mathrm{night}}+10}{10}} \right) \right], \tag{3}$$

where L_{day}, L_{evening} and L_{night}, also denoted as L_{d}, L_{e} and L_{n}, respectively, are the A-weighted long-term average sound level. In this paper, L_{d}, L_{e} and L_{n} are calculated using Equation (2), determined over all the day periods (07:00–19:00), evening periods (19:00–23:00) and night periods (23:00–07:00), respectively, over the assessment period.

Each instance of the Madrid's files contains the indicators $L_{\mathrm{Aeq}T}$ corresponding for a day, evening and night periods and 24h for a specific station on every day. Therefore, in this work, $L_{\mathrm{d1}y}$, $L_{\mathrm{e1}y}$ and $L_{\mathrm{n1}y}$ indicators will be used as inputs to model the behavior of the nodes in different periods of the day, so the temporal variability during a day is taken into account. Moreover, yearly standard deviation of $L_{\mathrm{den1}d}$ to identify the variability of the nodes during a year, denoted as $sd_{1y}(L_{\mathrm{den1}d})$. The selection of these variables as inputs is also based on Directive 2002/49/EC [1].

Data Preprocessing and Exploratory Analysis

Firstly, a data quality analysis was conducted, identifying nulls and the completeness of the data. Due to some technical mistakes, such as connections errors, maintenance and breaks, all the information is not usually available. Therefore, an analysis of the completeness of the data must be carried out to identify the amount of available and missing data. In the calculation of the KPIs used in this research, the instances with missing or erroneous data has been dropped out of the calculations.

Once the KPIs are calculated, some basic exploratory analysis can be performed to look for some important features of the data. Although this is not the main objective of this paper, some results are shown to illustrate the experiment. For instance, the sound pressure level time series can be analyzed for each node independently. Figure 2 shows the $L_{\mathrm{den1}d}$ statistics along the available dates. The red vertical lines delimit the period of national state of alarm decreed by the country, with a lockdown from 15 March 2020 to 21 June 2020. Through these graphs, a discussion could arise regarding the effects of the COVID-19 disease in noise pollution. Although this analysis is out of the scope of the current work, readers should note that the impact of the COVID-19 lockdown period in noise levels and soundscapes has been analyzed in different cities, such as Madrid [8] and Milan [9].

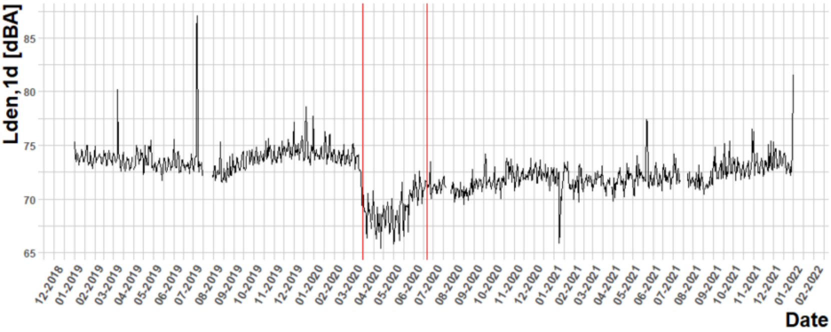


Figure 2. $L_{\mathrm{den1}d}$ time series for node MAD_2. Note that Spanish lockdown corresponds with the period between red lines.

Model Description

In this research, several clustering algorithms were trained, including k-means clustering [10], hierarchical agglomeration [11], expectation maximization algorithm [12], partitioning around medoids [13], the divisive hierarchical algorithm DIANA [14],the fuzzy clustering FANNY [14], the sampling-based algorithm CLARA [14], Kohonen self-organizing maps [15] and the self-organizing tree algorithm (SOTA) [16].

Using the following yearly acoustic indexes, L_{day}, L_{evening}, L_{night} and standard deviation of L_{den}. A comparison of the results using Dunn Index [17], Connectivity [18] and Silhouette Width [19] concludes that hierarchical agglomeration has the best performance for these data. The evaluation of the algorithm are presented in Section Results

The method considered in the following, called hierarchical agglomeration [11], is an unsupervised learning algorithm which groups the unlabeled dataset into different clusters. Initially, each observation or instance (in our case each node) is placed in its own cluster. The clusters are then sequentially (in steps) combined into larger clusters until all elements end up being in the same cluster that contains all the observations. At each step, the nearest two clusters are combined. To identify the nearest two cluster, a distance D between clusters must to be calculated. As a cluster can have one or more observations, the distance D between two clusters (also called height) is the maximum of the distance d of pairs observations, each from each cluster as shows Equation (4) where d is the euclidean distance between observations.

$$D(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \tag{4}$$

Once all the observations are together in the same clusters, you must go back $n - 1$ steps to divide the observations in n clusters.

Software and Technology

The preparation, transformation, analysis and modelling of the data have been performed using the Statistical Programming Language R [20], combining a local environment using R version 4.1.0 with a cloud environment provided by RStudio Cloud using R version 4.1.2. The cloud environment has been used to parallelize some tasks. The following libraries have been involved in the tasks: stringr (Version 1.4.0), dplyr (Version 1.0.5), tidyr (Version 1.1.3), cluster (Version 2.1.1), ggplot2 (Version 3.3.3), hrbrthemes (Version 0.8.0), imputeTS (Version 3.2) and zoo (Version 1.8-9).

To ensure the reproducibility of the research, in every task that includes a random step, the seed using the R function set.seed() has been fixed. Due to changes in random numbers generation in R version 4.0.0, the way to generate them to be sure that the analysis will be reproducible in every R version has also been defined.

Results

In this section, the evaluation of the different algorithms is presented and compared. Also the results obtained from applying the optimal clustering technique, see subsection Model Description for details, to the collected data, see subsection Data Preprocessing and Exploratory Analysis for details, are presented.

In order to show the relevance and the relation between these indicators, $L_{\mathrm{d}2019}$, $L_{\mathrm{e}2019}$, $L_{\mathrm{n}2019}$ and $sd_{2019}(L_{\mathrm{den}1d})$, a smoothed color density scatterplot representing all the nodes can be seen in Figure 3. The smoothed color density helps to identify dense zones that groups nodes with similar behavior. The first row of plots compares the sound pressure level statistics pairwise. The black line is the so-called identity line meaning that both statistics are equal. The nodes in the upper right part of each plot show high sound pressure level values that affects citizen well being. Almost all nodes has similar sound pressure level at daily and evening periods and lower nightly, except MAD_{18} that has higher nightly sound pressure level than daily and evening. The second row of plots compares each sound pressure level statistic with the standard deviation of $L_{\mathrm{den}1d}$.

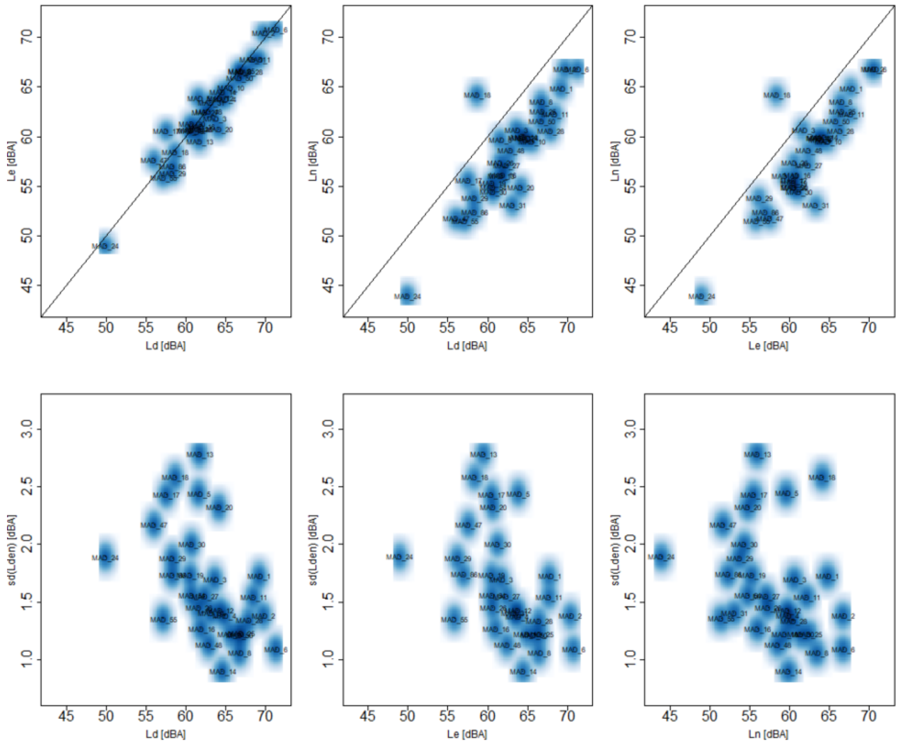
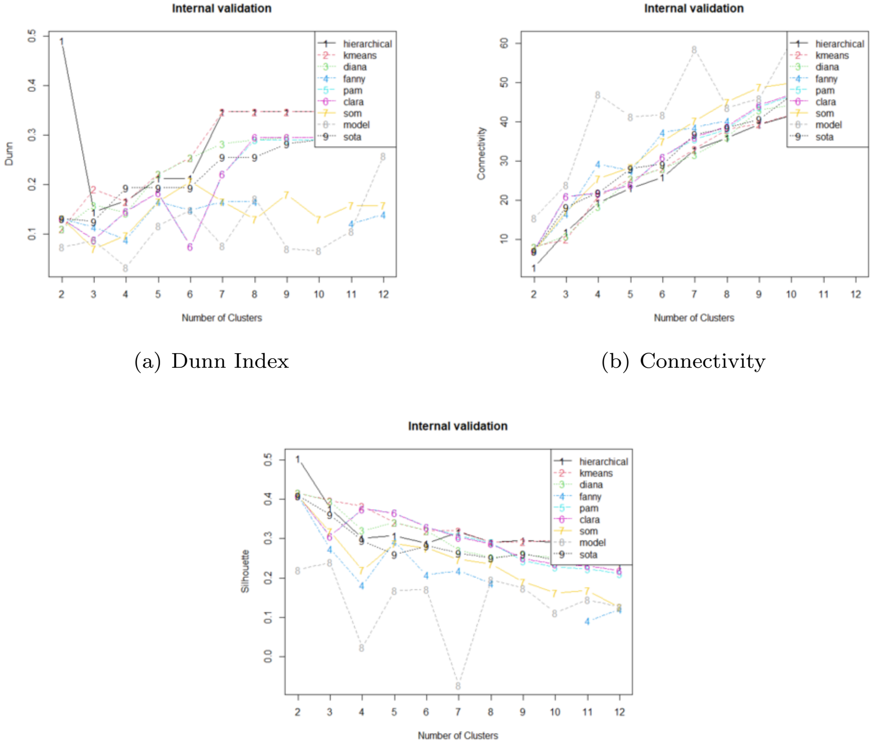


Figure 3. Scatter plot of the $L_{\mathrm{d}2019}$, $L_{\mathrm{e}2019}$, $L_{\mathrm{n}2019}$, and $sd_{2019}(L_{\mathrm{den}1d})$ metrics representing all the nodes. Black line represents identity line, i.e., equal value for both KPIs.

In these plots, it can be identified different types of nodes: nodes with low sound pressure level and low standard deviation related with quite zones, nodes with high sound pressure level and low standard deviation related with a constant high noise pollution and nodes with high standard deviation that have some days with low sound pressure level and other days with high sound pressure level. A dense zone around the point $L_{\mathrm{d}2019} = 63$ dBA, $L_{\mathrm{e}2019} = 63$ dBA, $L_{\mathrm{n}2019} = 55$ dBA and $sd_{2019}(L_{\mathrm{den}1d}) = 1.2$ dBA groups nodes with a constant noise pollution along both the day and the year. Node MAD_{24} is located in the middle of the iconic park in Madrid called *Casa de Campo*, this is the reason why this is the one with lowest sound pressure level in daily, evening and nigthly periods.

Figure 4 shows Dunn Index, Silhouette and Connectivity for every algorithm considering 2 to 12 clusters. Hierarchical agglomeration with 2 clusters is the optimal one because maximizes Dunn Index and Silhouette Width and minimizes Connectivity.



(a) Dunn Index

(b) Connectivity

(c) Silhouette Width

Figure 4. Validation measures for a given set of clustering algorithms and number of clusters.

According with the analysis, it has been consider the hierarchical agglomeration clustering model to groups the nodes in the 2 optimal clusters because is the optimal one since maximizes Dunn Index and Silhouette Width and minimizes Connectivity.

The first cluster (red color) is made up of the 21 nodes that have the highest sound pressure and therefore entail a greater risk to health, while the second (blue color) is made up of the 10 nodes that have the lowest sound pressure as can be seen in Table 1. It is interesting to note that although the sound pressure values of the first cluster are higher, their variability is much lower while it is higher in those nodes that belong to the second cluster.

Table 1. Size and centroid of clusters using data collected during 2019.

Cluster	$L_{\mathbf{d2019}}$	$L_{\mathbf{e2019}}$	$L_{\mathbf{n2019}}$	$sd_{2019}(L_{\mathbf{den1}d})$	Size	Color
1	65.0	64.4	59.8	1.39	21	blue
2	58.4	57.9	54.4	2.16	10	red

Discussion

Table 1 shows that clusters centroids created by hierarchical clustering are different but it's necessary to analyse the distribution of the node's values and their volatility to help city managers to understand the underlined patterns to take decisition in order to manage noise pollution.

Figure 5 represents the variable distributions for both clusters showing that the behavior of the clusters are different where the 75th percentile of the cluster distribution for the low value variable is lower than the 25th percentile of the high value variable. So cluster 1 and cluster 2 need different action to manage noise pollution. Also, there is an outlier node corresponding with MAD_{24} that should be managed independently.

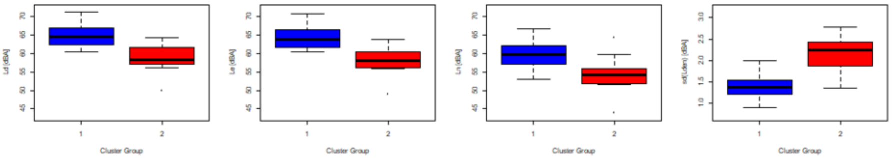


Figure 5. Clustering Boxplot

Dendogram allows city managers to select smaller groups than clusters with similar characteristics related with sound pressure level. It is possible to do a deep dive to understand the relationships between nodes showing a dendogram as Figure 6, where each square represents each cluster and the nodes belonging to. A dendrogram is a tree based diagram showing hierarchical clustering relationships between instances in a dataset, in particular, it's a summary of the distance matrix between instances or grouped instances used in the hierarchical agglomeration clustering algorithm. The height represent the distances between the two joined groups. Normally it is used to understand which intances or group of instances are more similar to others and to show the clusters. In this case, we can identify nodes with similar behavior, as MAD_{25} and MAD_{50} witch are very similar because their link has a small height.

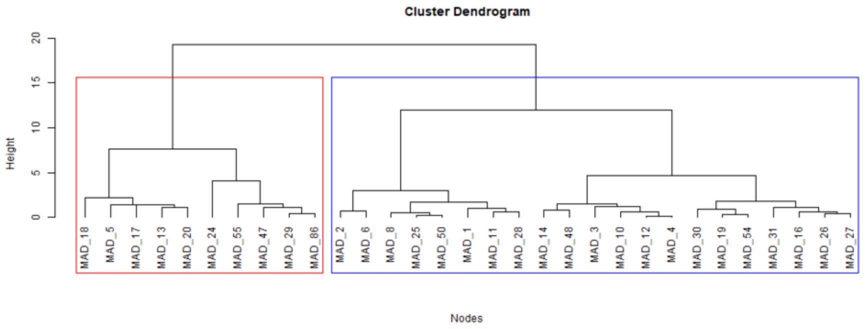


Figure 6. Hierarchical agglomeration

In this preliminary work, it is shown that unsupervised learning technique is a promising technique that allows city managers to identify different sound pressure level behaviors to help them proposing personalized strategy for each group. The next step of designing plans in order to reduce the noise pollution in the city and measure the impact of each action plan in each behavioral group, matching behaviours groups with optimized actions in order to minimize noise pollution and improve healthy. Unsupervised learning provides meaningful insight for urban-designing and healthcare sector to control and prevent noise pollution.

Author Contributions: Conceptualization, A.P., J.M.N.; data curation, A.P.; modeling, A.P.; investigation, A.P.,J.M.N.,F.R.; methodology, A.P.; software, A.P.; supervision, J.M.N. and F.R.; validation, J.M.N. and A.P.; writing—original draft, A.P.; writing—review and editing, A.P., J.M.N and F.R. All authors have read and agreed to the published version of the manuscript.

Funding: Financial support for this research has been provided under grant PID2020-112827GB-I00 funded by MCIN/AEI/10.13039/501100011033.

Data Availability Statement: The data analyzed in this research has been downloaded from the Acoustic Pollution repository in the Madrid's open data platform [6]

References

[1] European Commission. *END, Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 Relating to the Assessment and Management of Environmental Noise*; European Commission: Brussels, Belgium, 2002.

[2] Zambon, G.; Benocci, R.; Brambilla, G. Cluster categorization of urban roads to optimize their noise monitoring. *Environ. Monit. Assess.* **2016**, *188*, 26.

[3] Zambon, G.; Benocci, R.; Bisceglie, A.; Roman, H.E.; Bellucci, P. The LIFE DYNAMAP project: Towards a procedure for dynamic noise mapping in urban areas. *Appl. Acoust.* **2017**, *124*, 52–60.

[4] Brambilla, G.; Benocci, R.; Confalonieri, C.; Roman, H.E.; Zambon, G. Classification of urban road traffic noise based on sound energy and eventfulness indicators. *Appl. Sci.* **2020**, *10*, 2451.

[5] Garrido, J.C.; Mosquera, B.M.; Echarte, J.; Sanz, Roberto. Management Noise Network of Madrid City Council in *InterNoise19, Madrid, Spain.* INTER-NOISE and NOISE-CON Congress and Conference Proceedings, Madrid, Spain, June 16-19, Institute of Noise Control Engineering, InterNoise19, Madrid, Spain, pages 996-1997, pp. 1700-1711(12).

[6] Portal de Datos Abiertos del Ayuntamiento de Madrid. Available online: https://datos.madrid.es/portal/site/egob (accessed on 20 February 2022).

[7] ISO 1996-2:2017. Acoustics—Description, Measurement and Assessment of Environmental Noise—Part 2: Determination of Environmental Noise Levels; *International Organization for Standardization: Geneva* **2017**, Switzerland.

[8] Asensio, César; Pavón, Ignacio; Arcas, G.. (2020). Changes in noise levels in the city of Madrid during COVID-19 lockdown in 2020. The Journal of the Acoustical Society of America. 148. 1748. 10.1121/10.0002008.

[9] Benocci, R.; Roman, H.E.; Confalonieri, C.; Zambon, G. Investigation on clusters stability in DYNAMAP's monitoring network during COVID-19 outbreak. *Noise Mapp.* **2020**, *7*, 276–286.

[10] MacQueen, J.B. Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June – 18 July 1967; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.

[11] Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

[12] Fraley, C.; Raftery, A.E.; Scrucca, L.; Murphy, T.B.; Fop, M. "Mclust" Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. 2012. Available online: http://cran.r-project.org/web/packages/mclust/index.html (accessed on 26 June 2021).

[13] Kaufman, L.; Rousseeuw, PJ. *Clustering by means of medoids.* In: Dodge Y (ed) Statistical Data Analysis Based on the L1 Norm and Related Methods, **1987**, pp 405–416.

[14] Kaufman, L.; Rousseeuw, PJ. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley Series in Probability and Mathematical Statistics; John Wiley & Sons: NJ, USA, 1990; ISBN 9780471878766.

[15] Kohonen, T. *Self-Organizing Maps.* Springer-Verlag, second edition, 1997.

[16] Dopazo, J.; Carazo, JM. Phylogenetic Reconstruction using a Growing Neural Network that Adopts the Topology of a Phylogenetic Tree. *Journal of Molecular Evolution* **1997**, pp. 226–233.

[17] Dunn, J.C. Well separated clusters and fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104.

[18] Handl, J.; Knowles, K.; Kell, D. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **2005**, *21*, 3201–3212.

[19] Peter, J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.

[20] R. Available online: https://www.r-project.org/ (accessed on 1 June 2020).