applied sciences

MDPI

Article

Analysis and Evaluation of Clustering Techniques Applied to Wireless Acoustics Sensor Network Data

Antonio Pita, Francisco J. Rodriguez and Juan M. Navarro *

Research Group in Advanced Telecommunications (GRITA), Universidad Católica de Murcia (UCAM), 30107 Guadalupe, Spain
* Correspondence: jmnavarro@ucam.edu

Abstract: Exposure to environmental noise is related to negative health effects. To prevent it, the city councils develop noise maps and action plans to identify, quantify, and decrease noise pollution. Smart cities are deploying wireless acoustic sensor networks that continuously gather the sound pressure level from many locations using acoustics nodes. These nodes provide very relevant updated information, both temporally and spatially, over the acoustic zones of the city. In this paper, the performance of several data clustering techniques is evaluated for discovering and analyzing different behavior patterns of the sound pressure level. A comparison of clustering techniques is carried out using noise data from two large cities, considering isolated and federated data. Experiments support that Hierarchical Agglomeration Clustering and K-means are the algorithms more appropriate to fit acoustics sound pressure level data.

Keywords: unsupervised learning; environmental noise assessment; urban acoustic environment; wireless sensor network data; knowledge discovery; clustering algorithms; data clustering

Citation: Pita, A.; Rodriguez, F.J.; Navarro, J.M. Analysis and Evaluation of Clustering Techniques Applied to Wireless Acoustics Sensor Network Data. *Appl. Sci.* **2022**, *12*, 8550. https://doi.org/10.3390/app12178550

Academic Editors: Małgorzata Charytanowicz and Piotr A. Kowalski

Received: 8 July 2022
Accepted: 21 August 2022
Published: 26 August 2022

1. Introduction

The European directive 2002/49/EC [1] encouraged agglomerations of people, namely, cities or groups of cities nearby, to create their strategic noise mapping (SNM) sharing the results with citizens. Moreover, the results of these noise maps led to the establishment of noise-reduction action plans where noise exposure protection zones are defined. To create performance reports with the data obtained in the strategic noise map and to define special noise protection areas within the city, data are usually analyzed by descriptive analysis, with basic statistics, such as the average or median of the defined noise indicator obtained for the overall assessment period. In general, using these statistics, two main types of areas are proposed relying on the places where values are higher than a certain recommended sound level, known as special regime areas, and others where their noise exposure is lower than the average, known as quiet areas. However, the acoustic environment of an area is a complex phenomenon that needs to be characterised not only by the noise levels in the area, but also by other properties such as its behavior in different time periods of the day and its long-term variation. Therefore, it would be interesting to explore the application of clustering techniques for the identification of areas with different behavior in relation to the noise environment.

Murphy et al. [2] analyzed the methodological issues concerning the implementation of the directive across different countries of the European Union (EU), and dealing specifically with noise calculation and noise mapping, highlighting the implications of these issues for cross-country sharing of results. Moreover, a recent research [3] also summarizes the challenges to be faced by the EU Members and concludes that the opportunity to set up a common database of noise exposure based on common methods should be seized on time, encouraging local administrations to establish common frameworks. In the period 2021–2027, the European Commission will invest in a High Impact Project on European

data spaces and federated cloud infrastructures to encourage the establishment of EU-wide common, inter-operable data spaces in strategic sectors, such as mobility and health, and public administrations with data spaces initiatives, such as Gaia-X [4] and Federated European Infrastructure for Genomics data and Cancer Images data [5]. In the future, these data spaces can be used to join noise pollution data owned by public administrations to improve the health of citizens by the creation of more accurate predictive models, or obtaining better insights due to the more available data. In line with this trend, the application of unsupervised learning algorithms using federated data are proposed in this work to identify different acoustic environments that can help city managers to define personalized action plans for each behavior and share data in a common framework.

In recent years, large cities are deploying Wireless Acoustic Sensor Networks (WASN), based on Internet of Things (IoT) technologies [6], to perform continuous monitoring of environmental acoustic parameters at many locations [7]. The acoustic nodes that compose these networks continuously capture information regarding the sound environment over long periods, generating a large amount of data. These acoustic data, together with further environmental data, such as water quality [8] or air pollution [9], are being used by city managers to make decisions and propose improvement actions. Moreover, this smart city system has given rise to the creation of the so-called dynamic noise maps where SNM is more often updated, each day for instance, by integrating data obtained from acoustic sensors and the application of predictive models of sound propagation in cities [10]. The improvement of SNM has been the goal of researchers, such as Puyana-Romero et al. [11] who concludes that colors add supplementary and more intuitive information on soundscape to those provided by the acoustic parameters. In this work, acoustic datasets from WASN deployed in Barcelona and Madrid cities, Spain, are used for comparison of several machine learning clustering techniques.

Machine Learning has been used with acoustic data, both audio signal and sound level indexes, to help cities to manage noise in recent literature. On one hand, supervised learning techniques were applied to identify the main noise source of the acoustic environment using Mel-frequency cepstral coefficients as features with Gaussian Mixture Model and Artificial Neural Networks as algorithms [12]. In reference [13], Convolutional Neural Networks (CNN) were evaluated to classify urban sound events using local features of short-term sound recording features and with long-term descriptive statistics. Additionally, CNN was implemented to detect anomalous noise source detection to remove unrelated road traffic noise events and then generate a noise map [14]. Another study using CNN over acoustic signal recordings developed a system to detect the presence of an unmanned aerial vehicle in a complex urban acoustic scenario focusing on cities security [15]. On the other hand, the unsupervised learning technique Hierarchical Agglomeration was trained to optimize the choice and the number of monitoring sites [16] for defining a methodology to estimate the mean L_d and L_n levels in urban roads with the noise profiles detected in the clustering [10]. Additionally, the K-means method was trained in reference [17] to identify sound pressure level patterns.

In this paper, the performance of several data clustering techniques is evaluated for discovering and analyzing different behavior patterns of the sound pressure level. A comparison of clustering techniques is carried out using noise data from two large cities, considering isolated and federated data. After this introduction, datasets, applied techniques, and evaluation metrics are described in the next Section 2. Then, the results of the comparison together with a discussion and an analysis of these results are presented in Section 3. Finally, the main conclusions of this work are summarized in Section 4.

2. Materials and Methods

This section presents materials and methods applied during this research. Two datasets, described in Section 2.1, containing sound pressure level indicators for fixed locations during a long period were used. Additionally, a third federated dataset has been created, joining the previous one involving the nodes of both cities together. The list and

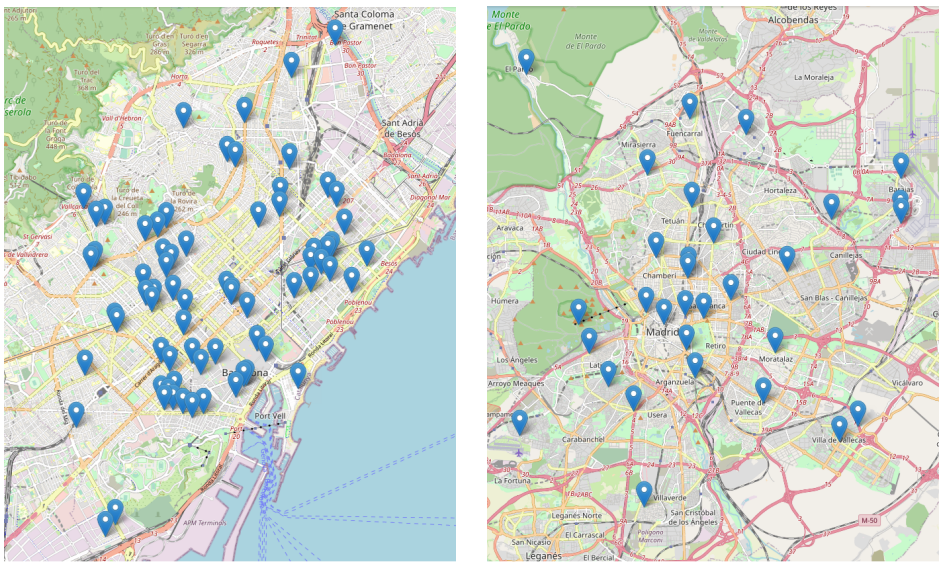
references for the clustering techniques used in this work can be found in Section 2.2. Once the models are trained, an evaluation of their performance allows for comparing the different algorithms using three different metrics that analyze the internal structure of the clusters. The definition of the metrics is presented in Section 2.3. Last, the software and hardware used to perform all the processing and analysis can be found in Section 2.4.

2.1. Data Sources

This research has considered datasets from two different WASNs deployed in big cities, Barcelona and Madrid, Spain, and collected sound pressure level values.

On one hand, the network of acoustics nodes deployed in Barcelona, denoted in this work by BCN_X, by the city council during the last years consists of 86 sound sensors [18,19]. The dataset used in this research was collected from 70 of the 86 sound sensors that were chosen for reasons of stability of the data over time and homogeneity in the spatial distribution of the nodes. The data were provided by the Barcelona City Council after a request from the authors. In the Acknowledgments section, the names of the data managers are indicated. As a summary, the data captured using Cesva TA120 [20] remote sonometers, considering international standards [21,22], is aggregated and sent to a data platform called *Plataforma de Sensors i Actuadors de Barcelona* [23]. A detailed explanation about the technological structure of the WASN and the data pipeline process involved can be found in Camps et al. [18]. A description of the data source, the transformations carried out, the variables created, along with the distribution of the nodes is provided in a previous article of the authors [17].

On the other hand, the acoustic pollution monitoring network of the city of Madrid has 31 permanent stations, denoted in this work by MAD_X, in charge of the control and continuous monitoring of the existing noise levels. Garrido et al. [24] described Madrid's WASN in detail showing how sound pressure level measurement dataset of these stations was retrieved from the acoustic pollution sensors and stored in a database management system platform that allows data analysts to work with the data in a structured way. The data are available on the Madrid council's open data portal [25]. In particular, data from recent years can be downloaded in the acoustic pollution data repository [26]. In the current research, only data from 2019 from both cities have been selected to explore a regular year period and avoid the pandemic period. More details regarding descriptive analysis and data processing can be found in previous authors' studies [17,27] for both cities. Figure 1 shows the location of the chosen nodes in both cities.



(**a**) (**b**)

Figure 1. (**a**) Location of the of the 70 acoustic nodes in the city of Barcelona, Spain and (**b**) Location of the 31 acoustic nodes in the city of Madrid, Spain.

These datasets are transformed into a normalized common structure that allows the comparison. The common structure is an structure table where rows represent each node with the following features: L_{d2019}, L_{e2019}, L_{n2019} and $sd_{2019}(L_{\text{den1}d})$. Three first sound pressure level features have been selected considering the recommendations established in Directive 2002/49/EC [1] and to take into account levels during different time periods of the day. The last feature has been chosen to take into account long-term variation of the main parameter L_{den} in Directive 2002/49/EC [1]. These acoustic parameters are defined below.

ISO 1996-2: 2017 [22] developed by the technical committee ISO/TC 43/SC 1 Noise describes how sound pressure levels intended as a basis for assessing environmental noise limits or comparison of scenarios in spatial studies can be determined. Determination can be performed by direct measurement and by extrapolation of measurement results through calculation. In this research, the definition, notations, and calculations performed over acoustic data follow the referred ISO [22]. As the sound pressure $p(t)$ is 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 $L_{\text{eq}T}$, Equation (1) is used.

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

where 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 $L_{\text{Aeq1}m}$ in dBA unit, applying Equation (1).

From these one minute period data, L_{d}, L_{e} and L_{n}, defined as the A-weighted long-term average sound pressure level for day, evening and night periods respectively, are calculated using Equation (2). These features are determined over all the day periods (07:00–19:00 h), evening periods (19:00–23:00 h), and night periods (23:00–07:00 h), respectively, across all the assessment periods.

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

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 $L_{\text{Aeq1}h}$, 60 values of $L_{\text{Aeq1}m}$ are averaged.

Finally, the daily standard deviation $sd_{2019}(L_{\text{den1}d})$ is computed. L_{den}, defined in Equation (3), refers to the day–evening–night noise indicator obtained for an overall annoyance in the assessment period [1] for one year.

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

2.2. Unsupervised Learning Algorithms

There are a large number of algorithms in the literature dedicated to data clustering. In this research, several representative algorithms from three unsupervised learning approaches, in particular, hierarchical, partitional, and model-based techniques have been considered to evaluate which one performs better over acoustic data. As it is mentioned in Section 1, Hierarchical Agglomeration and K-means have been previously applied to acoustic data. In this paper, other clustering algorithms, together with the mentioned above, were trained to fit the data:

1. HC: Hierarchical Agglomeration [28];
2. DIANA: a divisive hierarchical algorithm [29];
3. KM: K-means [30];
4. PAM: Partitioning Around Medoids [31];

5. CLARA: the sampling-based algorithm [29];
6. SOM: Kohonen Self-Organizing Maps [32];
7. SOTA: the Self-Organizing Tree Algorithm [33];
8. GAUSS: Expectation Maximization (EM) algorithm over a finite mixture of Gaussian distributions [34].

Hierarchical Agglomeration [28] and DIANA [29] methods, belonging to hierarchical clustering methods, create the clusters grouping the elements in hierarchical steps. K-means [30], PAM [31] and CLARA [29] methods, belonging to partitional clustering methods, are based on centroids and they iterative the algorithm until convergence. Moreover, SOM [32] technique applies an unsupervised neural network, and SOTA [33] is an evolution of the SOM algorithm which included a binary tree topology, both belonging to model-based methods, in this case in machine learning algorithms. Finally, GAUSS [34] technique is based on the maximization of the likelihood for a statistical distribution, belonging to model-based methods, in this case in statistical normal distributions.

In reference [35], a revision of different approaches for grouping similar objects into different groups is presented with an analysis of the advantages and disadvantages of every algorithm family. The features for the clustering algorithms chosen for this work are summarized in Table 1.

Table 1. Advantages and disadvantages of clustering algorithms used in this work.

Family	Algorithms	Advantages	Disadvantages
Hierarchical	HC, DIANA	suitable for the data set with arbitrary shape and attribute of arbitrary type, the hierarchical relationship among clusters easily detected, and relatively high scalability in general	relatively high in time complexity in general.
Partitional	KM, AM, CLARA	relatively low time complexity and high computing efficiency in general	not suitable for non-convex data, relatively sensitive to the outliers, easily drawn into local optimal, the number of clusters needed to be preset, and the clustering result sensitive to the number of clusters.
Model-Based	SOM, SOTA, GAUSS	diverse and well developed models providing means to describe data adequately and each model having its own special characters that may bring about some significant advantages in some specific areas	relatively high time complexity in general, the premise not completely correct, and the clustering result sensitive to the parameters of selected models.

2.3. Evaluation Metrics

For the evaluation and comparison of the clustering algorithms, Berry et al. [36] proposed two criteria for clustering evaluation and selection of an optimal clustering scheme: compactness and separation. Later, Hand et al. [37] introduce a new criteria: connectedness. In this article, these three internal characteristics are chosen to be calculated and analyzed.

Connectedness is related to what extent observations are placed in the same cluster as their nearest neighbors in the data space. To measure that connectivity [38], Equation (4) is

applied. For each element i, the n_{ij} represents the j-th nearest neighbor of i using a distance (often euclidean distance) and $I_{i,n_{ij}}$ is a boolean function that takes value $\frac{1}{j}$ when i and n_{ij} are not in the same cluster and zero otherwise. This metric is called the Connectivity metric.

$$Connectivity = \sum_{i=1}^{N} \sum_{j=1}^{M} I_{i,n_{ij}} \tag{4}$$

where N is the number of elements to group into K clusters and M is a parameter that determines the number of neighbors that contribute to the Connectivity measure, fixed to ten in this research as established in [39]. The Connectivity metric is equal to or higher than zero and the lower the value the better the clustering trained so must be minimized.

Compactness is related to cluster cohesion or homogeneity, measuring how close are the objects within the same cluster, usually by looking at the intra-cluster or within-cluster variance. A lower within-cluster variation is an indicator of good compactness, and, hence, a good clustering. So compactness must be minimized. The different indices for evaluating the compactness of clusters are based on distance measures, such as the cluster-wise within average/median distances between observations.

Separation measures how well-separated a cluster is from other clusters quantifying the degree of separation between clusters, usually by measuring the minimum distance between cluster centroids or the pairwise minimum distances between objects in different clusters. Therefore, separation must be maximized.

When the number of clusters increases, by definition compactness and separation used decrease. To manage this trade-off, some methods combine the two measures into a single score. The Dunn index [40] and Silhouette width [41] are both examples of non-linear combinations of compactness and separation.

The Dunn index aims to identify dense and well-separated clusters. It is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. Equation (5) shows how to calculate the Dunn index for clustering with K partitions.

$$Dunn = \frac{\min_{1 \le i \le j \le K} d(i,j)}{\max_{1 \le k \le K} \hat{d}(k)} \tag{5}$$

where $d(i,j)$ is the distance between cluster i and j (measuring separation) and $\hat{d}(k)$ is the intra-cluster distance of cluster k (measuring compactness). As separation should be maximized and compactness minimized, it results in that Dunn index must be maximized.

Silhouette width estimates the average distance between clusters considering how well an observation is clustered, in particular, how close each element in one cluster is to elements in the neighboring clusters. To calculate Silhouette width, it is necessary to first calculate the average dissimilarity a_i between the element i and all other elements of the cluster k to which i belongs (C_k) using Equation (6).

$$a_i = \frac{1}{|C_k| - 1} \sum_{j \in C_k, i \neq j} d(i,j), \tag{6}$$

representing the compactness of an element to the cluster to which belongs.

Secondly, for each element i, the average dissimilarity $d(i,C)$ of i to all elements of C are calculated and the minimum is computed, as enunciated in the following Equation (7).

$$b_i = min_{1 \le l \le K, i \notin C_l} \frac{1}{|C_l|} \sum_{j \in C_l} d(i,j) \tag{7}$$

where K in the number of clusters. This metric represents the separation of an element from the rest of the clusters.

Lastly, using results from Equations (6) and (7), the Silhouette width for an element i is calculated applying Equation (8).

$$S_i = \begin{cases} \frac{b_i - a_1}{\max(a_i, b_i)} & if \quad |C_k| > 1 \\ 0 & if \quad |C_k| = 1 \end{cases} \tag{8}$$

It is important to note that, the Silhouette coefficient of clustering is the mean of the Silhouette width of all the elements. Therefore, the objective is to maximize this index.

There are other internal validation metrics available to be used in the validation of an unsupervised learning algorithm [42–47] that could be alternatives to the selected ones. However, the chosen measures cover the three clustering criteria in order to evaluate and compare the trained clustering models [37].

2.4. Software and Hardware

The preparation, transformation, analysis, and modeling of the data have been performed using the Statistical Programming Language R [48] with the configuration presented in Table 2 for two environments, on-premise and cloud. The latter one has been used to parallelize some tasks.

Table 2. Libraries and Software Versions

Software Environment	**Version**
On-Premise	R version 4.1.0 called "Camp Pontanezen"
AMD Ryzen 7 3700X 8-Core Processor 3.60 GHz with 16 GB RAM and a GTX 1660 Super GDDR5 GPU.	
Cloud	R version 4.2.0 called "Vigorous Calisthenics"
RStudio Cloud Server	
Library	**Version**
stringr	1.4.0
dplyr	1.0.5
tidyr	1.1.3
cluster	2.1.1
ggplot2	3.3.3
clValid	0.7
mclust	5.4.8
kohonen	3.0.10

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.

3. Results and Discussion

In this section, the results of the performance of the different unsupervised learning algorithms are shown to evaluate and compare them with the three metrics explained in Section 2.3. Moreover, a selection of the best clustering algorithm to work with acoustic data to identify behavior patterns is completed. Additionally, a more detailed discussion of the resulting clustering is presented. This discussion is carried out by comparing cluster outputs from both federated data, that is, dataset containing data from both cities, and non-federated data, that is, dataset containing data from only one city.

For each normalized dataset, clustering algorithms listed in Section 2.2 are trained several times, increasing the number of clusters from 3 to 12, to fit the three different datasets presented in Section 2. For the interest of the research, the case $k = 2$ is avoided because it uses to separate the nodes in one group of high sound pressure level values

and another of low sound pressure level values, not adding value since that is what city managers usually do. This particular case has been enunciated in previous literature [16,27] that would not help to discover new knowledge.

Firstly, Table 3 shows the results for the Connectivity metric of the different techniques.

Table 3. Evaluation and comparison of clustering techniques over WASN data based on the Connectivity metric.

	Number of Clusters									
Algor. [1]	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
BCN [1]'s WASN data										
HC [1]	9.10	13.89	17.09	23.23	26.30	29.52	37.63	40.19	41.00	45.05
KM [1]	**7.78**	12.56	24.89	22.37	28.27	29.44	38.80	47.25	54.12	53.34
DIANA	26.85	28.33	37.19	40.71	45.27	48.09	51.31	53.45	55.73	58.23
PAM	15.20	20.51	25.78	33.39	38.66	44.63	46.66	47.84	55.92	62.45
CLARA	15.20	19.74	22.84	33.39	36.49	42.51	46.01	50.15	58.31	58.44
SOM	7.78	12.56	24.89	32.98	40.81	53.87	56.66	54.86	70.92	83.47
GAUSS	21.38	39.29	33.10	53.58	43.54	40.99	42.92	NA [2]	98.71	68.33
SOTA	29.91	34.73	36.59	44.04	47.04	48.21	62.99	72.67	74.18	82.76
MAD [1]'s WASN data										
HC [1]	11.98	19.48	23.13	25.89	32.93	35.93	39.33	41.66	44.27	46.08
KM [1]	**10.08**	21.10	25.50	28.10	33.08	38.07	39.48	41.81	44.43	50.40
DIANA	10.95	18.51	25.50	28.10	31.70	36.11	42.94	44.26	46.88	50.81
PAM	21.05	21.87	23.96	31.10	35.63	38.23	43.82	46.57	49.19	50.42
CLARA	21.05	21.87	23.96	31.10	36.15	38.74	44.25	47.00	49.62	50.85
SOM	15.31	21.10	28.35	36.85	40.19	43.76	49.56	49.87	52.88	56.58
GAUSS	23.95	47.00	41.32	41.72	58.44	43.63	45.97	60.70	54.00	55.86
SOTA	18.36	21.97	28.15	29.24	36.80	38.53	40.64	46.66	48.98	NA [2]
Federated MAD [1] and BCN [1] joined WASN data										
HC [1]	**12.26**	16.52	20.38	22.53	27.76	33.53	40.42	48.79	52.38	54.84
KM [1]	30.02	35.89	31.19	33.40	43.14	45.00	50.82	52.97	57.73	64.06
DIANA	18.08	28.51	34.08	41.73	49.59	52.55	56.41	62.90	64.32	66.23
PAM	23.01	28.18	28.79	34.66	49.68	58.94	62.72	64.30	66.58	67.86
CLARA	24.92	28.18	32.40	38.59	44.21	50.22	52.50	61.45	68.14	67.11
SOM	23.03	29.11	31.19	37.76	48.08	57.93	61.27	66.99	77.79	75.84
GAUSS	69.57	63.19	93.73	79.91	81.40	83.81	87.62	89.39	97.70	97.46
SOTA	30.55	41.53	44.74	53.57	60.29	62.13	68.76	75.11	78.06	83.70

[1] Abbreviations: Algor.: Algorithm, HC: Hierarchical Agglomeration and KM: K-means, MAD: Madrid, BCN: Barcelona. [2] No convergence.

A first comparison of the resulting values offers that the best algorithm for the Connectivity metric, the optimum algorithms are Hierarchical Agglomeration and K-means. Note that values are highlighted in Table 3. From these results an important insight could be extrapolated, that the number of optimal clusters for the Connectivity metric holds in three.

Now, in Table 4 the Dunn index obtained for all the algorithms and 3 to 12 clusters are shown.

It is observed that this metric aims to create a higher amount of clusters, prioritizing separation from compactness. Again, note that the highest values are highlighted. For the Dunn index, Hierarchical Agglomeration and K-means algorithms are also the top performers.

Finally, Table 5 shows the Silhouette Width for all the clustering techniques and for the same number of clusters and the datasets previous indicated.

Hierarchical Agglomeration and K-means algorithms also maximize the Silhouette Width.

For this metric, it is shown in Table 5 that the Barcelona dataset and the federated dataset are recommended to be split into 4 clusters, but for the Madrid dataset, the recom-

mendation is 3 clusters. However, a hypothesis could be that Madrid only has three of the four behaviors identified in the full dataset.

Table 4. Evaluation and comparison of clustering techniques over WASN data based on the Dunn index.

Algor. [1]	**Number of Clusters**									
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
BCN [1]'s WASN data										
HC [1]	0.218	0.236	0.236	0.227	0.227	0.227	0.234	0.234	0.234	0.245
KM [1]	0.161	0.199	0.178	0.255	0.275	0.275	0.248	0.226	0.166	**0.282**
DIANA	0.073	0.074	0.098	0.102	0.102	0.125	0.146	0.157	0.165	0.171
PAM	0.069	0.146	0.172	0.173	0.173	0.173	0.222	0.240	0.240	0.240
CLARA	0.069	0.075	0.163	0.173	0.210	0.210	0.210	0.240	0.210	0.246
SOM	0.161	0.199	0.178	0.179	0.163	0.145	0.131	0.154	0.094	0.100
GAUSS	0.050	0.028	0.125	0.076	0.056	0.133	0.164	NA [2]	0.063	0.099
SOTA	0.091	0.101	0.101	0.101	0.101	0.101	0.059	0.059	0.059	0.059
MAD [1]'s WASN data										
HC [1]	0.144	0.166	0.212	0.212	0.347	0.347	0.347	0.347	0.364	**0.388**
KM [1]	0.191	0.165	0.220	0.253	0.347	0.347	0.347	0.347	0.364	0.358
DIANA	0.158	0.141	0.220	0.253	0.282	0.290	0.295	0.295	0.336	0.352
PAM	0.088	0.145	0.183	0.074	0.220	0.290	0.290	0.290	0.290	0.290
CLARA	0.088	0.145	0.183	0.074	0.220	0.295	0.295	0.295	0.309	0.309
SOM	0.193	0.165	0.165	0.084	0.165	0.129	0.061	0.129	0.181	0.244
GAUSS	0.086	0.032	0.116	0.147	0.075	0.171	0.070	0.066	0.105	0.258
SOTA	0.125	0.193	0.193	0.193	0.255	0.255	0.282	0.290	0.290	NA [2]
Federated MAD [1] and BCN [1] joined WASN data										
HC [1]	0.120	0.131	0.163	0.163	0.163	0.153	0.160	0.193	0.225	0.225
KM [1]	0.071	0.031	0.143	0.169	0.170	0.172	0.167	0.170	0.236	**0.251**
DIANA	0.075	0.082	0.089	0.103	0.110	0.122	0.145	0.146	0.157	0.158
PAM	0.061	0.101	0.100	0.167	0.110	0.083	0.071	0.071	0.093	0.095
CLARA	0.103	0.101	0.123	0.137	0.103	0.143	0.159	0.096	0.092	0.103
SOM	0.027	0.101	0.143	0.126	0.109	0.071	0.110	0.084	0.071	0.078
GAUSS	0.033	0.031	0.019	0.045	0.041	0.040	0.053	0.058	0.072	0.072
SOTA	0.044	0.044	0.044	0.051	0.059	0.059	0.059	0.059	0.079	0.079

[1] Abbreviations: Algor.: Algorithm, HC: Hierarchical Agglomeration and KM: K-means, MAD Madrid, BCN Barcelona. [2] No convergence.

As a summary, regarding the federated dataset, see Table 3 for details, the Connectivity metric is minimized with the Hierarchical Agglomeration algorithm for $k = 3$ clusters. It is important to note that, when the amount of elements to group is small, an increase in the number of clusters will increase Connectivity, thus this metric tends to select low values for the number of clusters. Then, the Dunn index selects K-means and Hierarchical Agglomeration clustering with $k = 12$ clusters as can be seen in Table 4. Finally, the Hierarchical Agglomeration algorithm for $k = 4$ has been selected by the Silhouette Width metric, see Table 5, showing that the Hierarchical Agglomeration method has a good equilibrium between the three clustering characteristics presented in Section 2.3.

After this first discussion, more details for $k = 3$ and $k = 4$ clusters cases are explained below. Applying the Hierarchical Agglomeration algorithm using $k = 3$, the data are divided into different groups, as it is graphed in a Dendogram in Figure 2. This Figure 2 shows the three main patterns that the algorithm has identified. To study the behavior of these three clusters, four box-plots graphs are shown in Figure 3, corresponding to the parameters used for the training phase, L_{d2019}, L_{e2019}, L_{n2019} and $sd_{2019}(L_{den1d})$. It can be observed in Figure 3 that, the first cluster is related to the nodes with high sound pressure levels during the day and evening period, medium sound pressure levels during the night, and the lowest standard deviation of the three clusters, to sum up, there are 42 nodes with

a stable and high noise level. The second cluster includes 29 nodes, and presents high sound pressure levels during the three periods reaching maximum noise level values, in addition to the highest standard deviation, in other words, the variation over the mean is high. Finally, the third cluster includes 30 nodes with the lowest sound pressure level during all periods. Moreover, its standard deviation is at an intermediate value between the two other clusters.

Table 5. Evaluation and comparison of clustering techniques over WASN data based on the Silhouette Width.

Algor. [1]	**Number of Clusters**									
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
BCN [1]'s WASN data										
HC [1]	0.294	0.353	0.317	0.386	0.368	0.358	0.327	0.308	0.318	0.312
KM [1]	0.376	**0.431**	0.356	0.404	0.395	0.392	0.362	0.376	0.369	0.377
DIANA	0.138	0.308	0.299	0.310	0.324	0.319	0.318	0.321	0.338	0.318
PAM	0.368	0.415	0.358	0.350	0.334	0.344	0.350	0.362	0.335	0.345
CLARA	0.368	0.418	0.361	0.350	0.339	0.343	0.346	0.353	0.338	0.338
SOM	0.376	0.431	0.356	0.348	0.323	0.297	0.293	0.342	0.257	0.197
GAUSS	0.147	0.041	0.268	0.255	0.315	0.338	0.361	NA [2]	0.022	0.240
SOTA	0.336	0.245	0.284	0.264	0.280	0.277	0.243	0.238	0.257	0.226
MAD [1]'s WASN data										
HC [1]	0.376	0.301	0.307	0.287	0.317	0.291	0.294	0.291	0.272	0.230
KM [1]	**0.396**	0.382	0.341	0.320	0.319	0.290	0.292	0.294	0.275	0.261
DIANA	0.395	0.320	0.341	0.320	0.270	0.252	0.258	0.251	0.245	0.261
PAM	0.306	0.375	0.365	0.329	0.307	0.290	0.243	0.228	0.223	0.211
CLARA	0.306	0.375	0.365	0.329	0.303	0.287	0.250	0.235	0.231	0.218
SOM	0.354	0.382	0.287	0.192	0.246	0.222	0.187	0.162	0.182	0.156
GAUSS	0.239	0.023	0.167	0.171	0.072	0.195	0.174	0.112	0.144	0.127
SOTA	0.361	0.294	0.259	0.282	0.263	0.250	0.261	0.246	0.275	NA [2]
Federated MAD [1] and BCN [1] joined WASN data										
HC [1]	0.399	**0.415**	0.359	0.309	0.297	0.347	0.355	0.388	0.396	0.385
KM [1]	0.303	0.371	0.380	0.389	0.371	0.378	0.394	0.391	0.406	0.385
DIANA	0.382	0.395	0.335	0.326	0.349	0.335	0.345	0.341	0.341	0.331
PAM	0.411	0.379	0.379	0.383	0.366	0.339	0.315	0.330	0.337	0.341
CLARA	0.310	0.379	0.366	0.374	0.356	0.379	0.382	0.346	0.280	0.359
SOM	0.411	0.380	0.380	0.383	0.366	0.330	0.341	0.317	0.305	0.293
GAUSS	0.192	0.187	0.086	0.028	0.077	0.099	0.205	0.209	0.232	0.213
SOTA	0.230	0.298	0.279	0.282	0.251	0.249	0.241	0.239	0.286	0.280

[1] Abbreviations: Algor.: Algorithm, HC: Hierarchical Agglomeration and KM: K-means, MAD Madrid, BCN Barcelona. [2] No convergence.

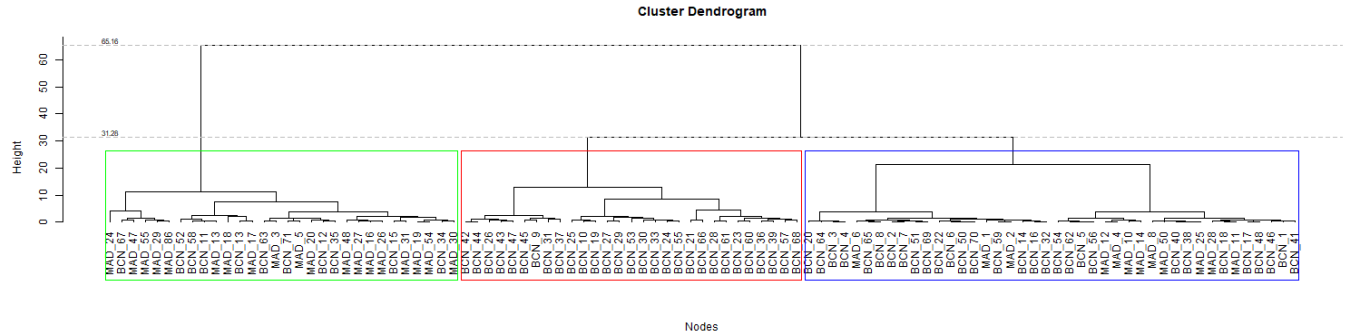
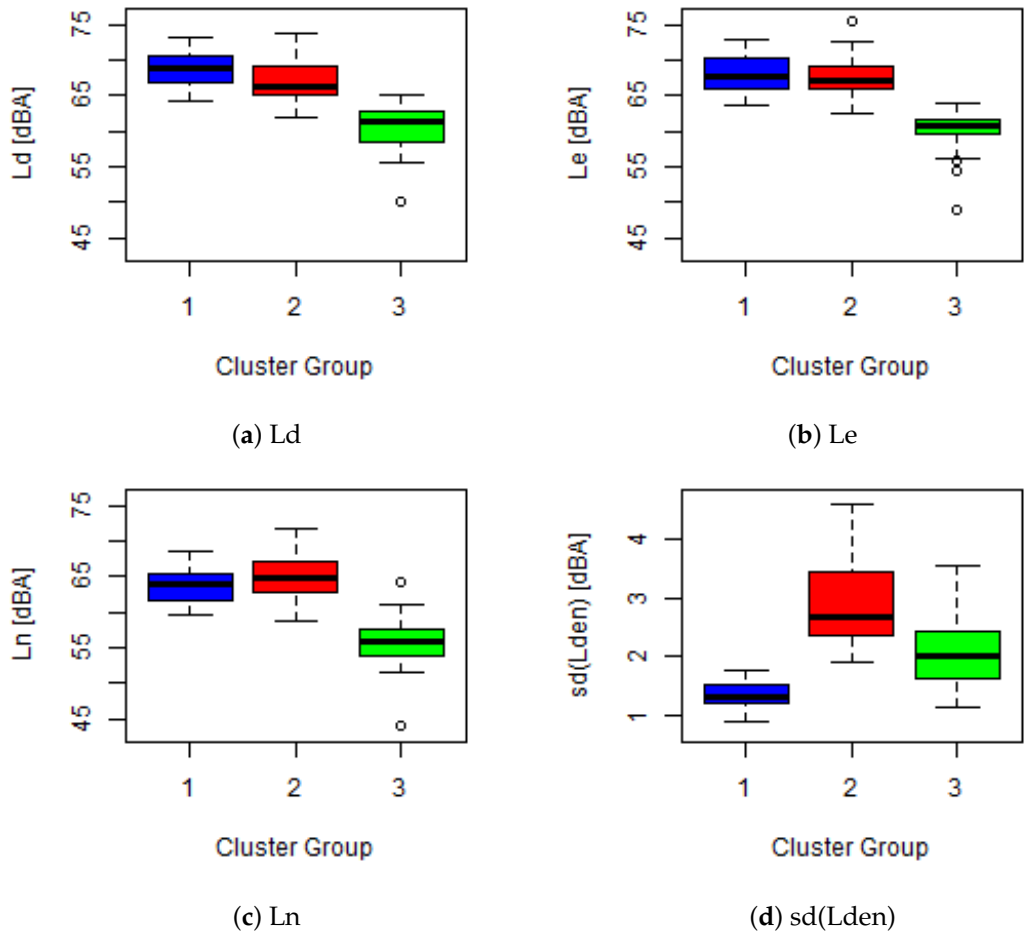


Figure 2. Nodes distribution within clusters by distance for k = 3 Hierarchical Agglomeration clustering. abbreviations: MAD_X Madrid Stations, BCN_X Barcelona Stations. Clusters groups are colored as presented in Table 6.



(**a**) Ld (**b**) Le

(**c**) Ln (**d**) sd(Lden)

Figure 3. Boxplot representation of the statistical distributions of the variables L_{d2019} (**a**), L_{e2019} (**b**), L_{n2019} (**c**) and $sd_{2019}(L_{den1d})$ (**d**) by cluster for $k = 3$. Hierarchical Agglomeration clustering model. Clusters groups are colored as presented in Table 6.

A summary of the three discovered clusters obtained with federated data is presented in Table 6 in which the number of nodes per city is broken down together with the centroid of each acoustics parameter.

Table 6. Size and centroid of clusters using data collected during 2019 for $k = 3$ Hierarchical Agglomeration clustering.

Cluster	L_{d2019}	L_{e2019}	L_{n2019}	$sd_{2019}(L_{den1d})$	Size	#MAD [1]	#BCN [1]	Color
1	68.7	68.2	63.6	1.36	42	12	30	blue
2	67.0	67.7	65.1	2.93	29	0	29	red
3	60.5	60.0	55.7	2.08	30	19	11	green

[1] Abbreviations: MAD Madrid Stations, BCN Barcelona Stations.

It is remarkable that, as can be seen in Table 6, cluster number 2 only contains nodes belonging to Barcelona city, suggesting that this type of behavior is specific to this city. Moreover, the relative proportion of Madrid's nodes in cluster number 1 is lower than in cluster number 3, showing that Madrid has nodes with lower sound pressure levels on average than Barcelona.

Now, Hierarchical Agglomeration is applied for $k = 4$ clusters. Figure 4 shows that previous cluster 1 (blue) with 42 nodes, obtained with $k = 3$, is split into two groups with 21 nodes each (blue and magenta).

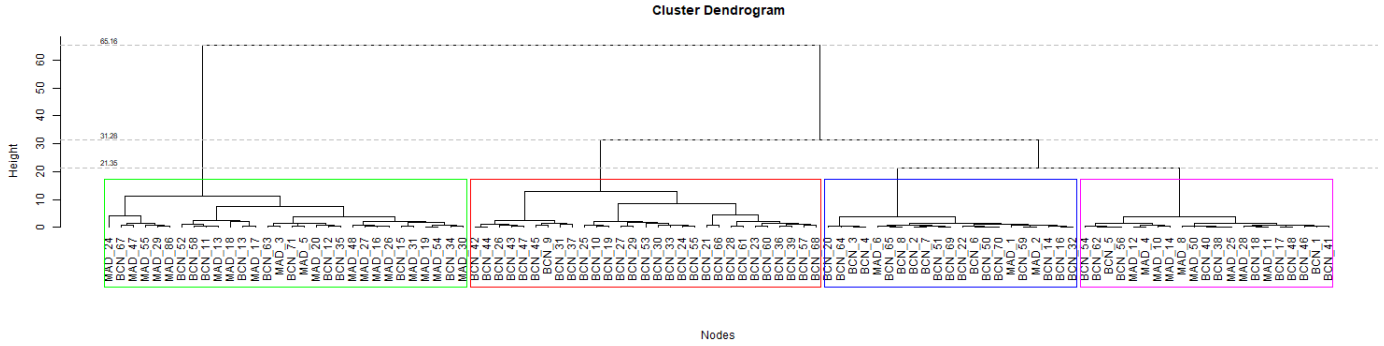


Figure 4. Nodes distribution within clusters by distance for $k = 4$ Hierarchical Agglomeration clustering. abbreviations: MAD_X Madrid Stations, BCN_X Barcelona Stations. Clusters groups are colored as presented in Table 7.

Table 7. Size and centroid of clusters using data collected during 2019 for $k = 4$ Hierarchical Agglomeration clustering.

Cluster	L_{d2019}	L_{e2019}	L_{n2019}	$sd_{2019}(L_{\mathrm{den}1d})$	Size	#MAD [1]	#BCN [1]	Color
1	66.7	66.0	61.5	1.41	21	9	12	blue
2	67.0	67.7	65.1	2.93	29	0	29	red
3	60.5	60.0	55.7	2.08	30	19	11	green
4	70.7	70.4	65.8	1.31	21	3	18	magenta

[1] Abbreviations: MAD Madrid Stations, BCN Barcelona Stations.

As it can be observed in the boxplots in Figure 5, the new blue cluster presents a lower sound pressure level than magenta and red, with a significant reduction in level during the night period. However, the new magenta cluster is the one with the highest sound pressure level and the lowest variance of the 4 clusters. In this case, the red cluster has the highest standard deviation.

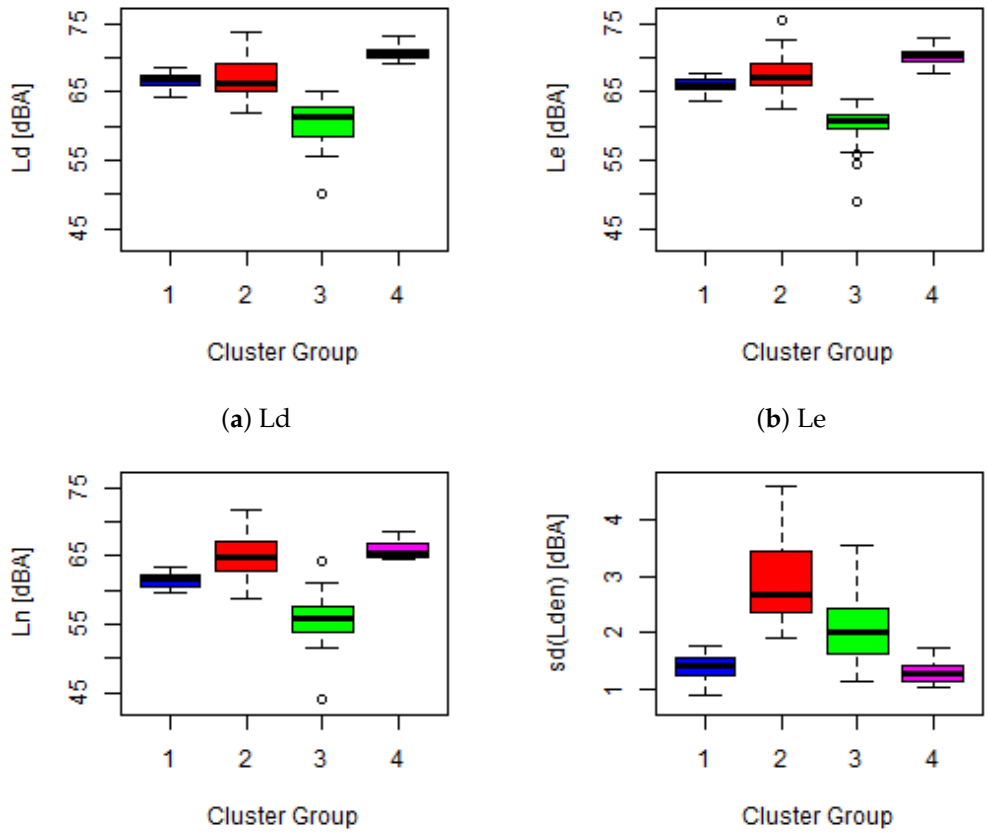
Table 7 summarizes the clusters showing distribution and centroids. As in the previous case, Madrid only presents three of four behaviors, explaining the outputs of the Silhouette Width metric of $k = 3$ for Madrid data, and $k = 4$ for both Barcelona and federated data, see details in Section 3.

Regarding the selection based on the Dunn index, from a noise pollution management perspective, it is neither useful nor easy to handle 12 clusters with only 2.6 nodes on average in Madrid and 5.8 nodes on average in Barcelona. This requires establishing 12 strategies with their associated action plans, therefore K-means for $k = 12$ is discarded from this analysis.

Another way to compare the results is using an external clustering validity index. If federated data clusters are considered the ground truth partition of the nodes, external evaluation metrics can select the more appropriate clustering completed in isolation. The Chi index is an external clustering validity index based on the chi-squared statistical test, very competitive that, on average, beats other external evaluation metrics [49]. The Chi index takes a value in [0, 2], where 0 is given by the worst clustering solution, and 2 is the best value that the Chi index can achieve. Chi index results are clear to read, require no further interpretation, and help to select the optimal number of clusters based on the ground truth class.

Table 8 shows results in cross tables a comparison between the $k = 3$ Hierarchical Agglomeration clustering considering the federated dataset, in rows, and the optimal cluster models trained considering isolated datasets, which are $k = 3$ and $k = 4$ K-means for Barcelona dataset, upper-left and lower-left, respectively, and $k = 3$ K-means and $k = 2$. Hierarchical for Madrid dataset, upper-right and lower-right, respectively. For instance, the upper-left cross table shows the distribution of the Barcelona nodes considering $k = 3$ K-means using federated data in rows and $k = 3$ K-means Barcelona data in columns, so

the number 9 in the second row and third columns represents the number of Barcelona nodes belonging to cluster number 2 in $k = 3$ K-means model using federated data and belonging to cluster number 3 in $k = 3$ K-means model using Barcelona data.



(**a**) Ld (**b**) Le

(**c**) Ln (**d**) sd(Lden)

Figure 5. Boxplot representation of the statistical distributions of the variables L_{d2019} (**a**), L_{e2019} (**b**), L_{n2019} (**c**) and $sd_{2019}(L_{den1d})$ (**d**) by cluster for $k = 4$. Hierarchical Agglomeration clustering model. Clusters groups are colored as presented in Table 7.

For Barcelona city, the federated dataset improves the result of the clustering compared with the Barcelona isolation data (1.104 maximum Chi index). So $k = 4$ K-means clustering is the best algorithm for Barcelona city based on the federated dataset clusters (chi index 1.104 versus Chi index 0.759 for $k = 3$ K-means algorithm). For Madrid city, the $k = 3$ K-means clustering is the best algorithm with a Chi index of 1.777 (compare to $k = 2$ Hierarchical Agglomeration with 0.714 Chi index). In this case, a smaller improvement has been made with the federated dataset ($0.223 = 2 - 1.777$), concluding that the clustering created with Madrid data in isolation gives almost the same information that the one created with the federated dataset.

Table 8. Comparison Federated Data $k = 3$ Hierarchical Agglomeration (in rows) with isolation BCN [1] or MAD [1] data clustering optimal models.

	(a) BCN [1] Data $k = 3$ K-Means			(b) MAD [1] Data $k = 3$ K-Means		
Cluster	**1**	**2**	**3**	**1**	**2**	**3**
1	25	5	0	0	0	12
2	10	10	9	0	0	0
3	0	11	0	5	13	1
	Chi index: 0.759			Chi index: 1.777		

Table 8. *Cont.*

	(c) BCN [1] Data k = 4 K-means				(d) MAD [1] Data k = 2 Hierarchical	
Cluster	**1**	**2**	**3**	**4**	**1**	**2**
1	18	12	0	0	12	0
2	5	14	9	1	0	0
3	0	1	0	10	9	10
		Chi index: 1.104			Chi index: 0.714	

[1] Abbreviations: MAD Madrid, BCN Barcelona.

4. Conclusions

Noise pollution is a major concern in cities around the world and wireless acoustic sensor networks are being deployed to acquire information about sound pressure level in many locations and during long-term. Sharing data between administrations in a big data infrastructure, as the EU commission is promoting, can help to obtain better insights and create a common framework. Machine Learning techniques are being applied to learn and analyze these datasets.

In this work, several machine learning clustering techniques have been applied to identify different acoustic environment patterns from sound pressure level datasets. A comparison of clustering techniques for modeling acoustic data from wireless acoustic sensor networks of the cities of Barcelona and Madrid (Spain) has been made. This evaluation has been performed using isolated data and federated data and three parameters as metrics: Connectivity, Dunn index, and Silhouette Width.

From the results, it is observed that both Hierarchical Agglomeration clustering and K-means have the best performance, in both federated and non-federated data. Therefore, they are the more suitable algorithms to fit environmental acoustics parameters, such as sound pressure levels during different periods of the day.

In general, the Connectivity and Silhouette indexes tend to select a low amount of clusters, whereas the Dunn index suggests a large number of groups. Regarding the use case of noise monitoring and management of the noise plans, a small amount of clusters is recommended, therefore the Connectivity or Silhouette index has been used to select the optimal clustering algorithm.

An external clustering validity index, the Chi index, has been also calculated, obtaining insight into the relevance of using federated data to do the clustering. More datasets will be incorporated in future works to further analyze the benefits of using federated datasets instead of isolated datasets.

It has been shown that these techniques can help the local administrations to dynamically detect different patterns of sound pressure level behavior and update the definition of acoustic zones. Moreover, this information can be publicly shared with citizens to know about the acoustic typology of the area in which they live or are planning to buy a house, allowing them better decisions.

Possible future work can continue this research along the following lines:

1. Design a methodology for monitoring the evolution of the acoustic zones to be able to measure the effect of the actions carried out by the consistories included in their action plans.
2. Create an acoustic open data spaces for federated data to identify common clusters.
3. Develop an algorithm to identify the cluster in which belongs to a city spot considering only a small sample of data.

Author Contributions: Conceptualization, A.P. and J.M.N.; data curation, A.P.; investigation, A.P., J.M.N. and F.J.R.; methodology, A.P.; software, A.P.; supervision, J.M.N. and F.J.R.; validation, J.M.N. and A.P.; writing—original draft, A.P.; writing—review and editing, A.P., J.M.N. and F.J.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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The BCN dataset used in this research has been provided by the city council of Barcelona, but authors of this research do not have permission to publish it. The MAD dataset has been downloaded from the Acoustic Pollution repository in the Madrid's open data platform [25].

Acknowledgments: We would like to thank Júlia Camps Farrés, Cap de Secció del Departament d'Avaluació i Gestió Ambiental in Ajuntament de Barcelona and Alejandro Aparicio Estrems, Tècnic Dep. d'Avaluació i Gestió Ambiental in Ajuntament de Barcelona, for their invaluable contribution providing the BCN dataset used in this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. European Commission. *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. Murphy, E.; King, E.A. Strategic environmental noise mapping: Methodological issues concerning the implementation of the EU Environmental Noise Directive and their policy implications. *Environ. Int.* **2010**, *36*, 290–298. [CrossRef] [PubMed]
3. Licitra, G.; Ascari, E. Noise Mapping in the EU: State of Art and 2018 Challenges. In Proceedings of the Comunication in Internoise, Chicago, IL, USA, 26–29 August 2018.
4. European Commission. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions a European Strategy for Data*; European Commission: Brussels, Belgium, 2020.
5. European Commission. *Annex to the Commission Implementing Decision; on the Financing of the Digital Europe Programme and the Adoption of the Multiannual Work Programme for 2021–2022*; European Commission: Brussels, Belgium, 2021.
6. Zanella, A.; Bui, N.; Castellani, A.; Vangelista, L.; Zorzi, M. Internet of Things for Smart Cities. *IEEE Internet Things J.* **2014**, *1*, 22–32. [CrossRef]
7. Alías, F.; Alsina-Pagès, R.M. Review of Wireless Acoustic Sensor Networks for Environmental Noise Monitoring in Smart Cities. *J. Sens.* **2019**, *2019*, 7634860. [CrossRef]
8. Martínez, R.; Vela, N.; el Aatik, A.; Murray, E.; Roche, P.; Navarro, J.M. On the Use of an IoT Integrated System for Water Quality Monitoring and Management in Wastewater Treatment Plants. *Water* **2020**, *12*, 1096. [CrossRef]
9. Yi, W.Y.; Lo, K.M.; Mak, T.; Leung, K.; Leung, Y.; Meng, M. A Survey of Wireless Sensor Network Based Air Pollution Monitoring Systems. *Sensors* **2015**, *15*, 31392–31427. [CrossRef] [PubMed]
10. 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. [CrossRef]
11. Puyana-Romero, V.;Ciaburro, G.; Brambilla, G.; Garzón, C.; Maffei, L. Representation of the soundscape quality in urban areas through colours. *Noise Mapp.* **2019**, *6*, 8–21. [CrossRef]
12. Maijala, P.; Shuyang, Z.; Heittola, T.; Virtanen, T. Environmental noise monitoring using source classification in sensors. *Appl. Acoust.* **2018**, *129*, 258–267. [CrossRef]
13. Ye, J.; Kobayashi, T.; Murakawa, M. Urban sound event classification based on local and global features aggregation. *Appl. Acoust.* **2017**, *117*, 246–256. [CrossRef]
14. Alsina-Pagès, R.M.; Alías, F.; Socoró, J.C.; Orga, F. Detection of anomalous noise events on low-capacity acoustic nodes for dynamic road traffic noise mapping within an hybrid WASN. *Sensors* **2018**, *18*, 1272. [CrossRef]
15. Ciaburro, G.; Iannace, G. Improving Smart Cities Safety Using Sound Events Detection Based on Deep Neural Network Algorithms. *Informatics* **2020**, *7*, 23. [CrossRef]
16. Zambon, G.; Benocci, R.; Brambilla, G. Cluster categorization of urban roads to optimize their noise monitoring. *Environ. Monit. Assess.* **2016**, *188*, 26. [CrossRef] [PubMed]
17. Pita, A.; Rodriguez, F.J.; Navarro, J.M. Cluster Analysis of Urban Acoustic Environments on Barcelona Sensor Network Data. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8271. [CrossRef] [PubMed]
18. Camps, J. Barcelona noise monitoring network. In Proceedings of the EuroNoise, Maastricht, The Netherlands, 31 May–3 June 2015; pp. 218–220.
19. Farrés, J.C.; Novas, J.C. Issues and challenges to improve the Barcelona Noise Monitoring Network. In Proceedings of the 11th European Congress and Exposition on Noise Control Engineering, Heraklion, Greece, 27–31 May 2018; pp. 27–31.
20. CESVA TA120 Noise Measuring Sensor for Smart Solutions. Available online: https://www.cesva.com/en/products/sensors-terminals/TA120/ (accessed on 15 May 2021).

21. *IEC 61672-1:2013*; IEC-International Electrotechnical Commission. Available online: https://webstore.iec.ch/publication/5708 (accesed on 15 May 2021).
22. *ISO 1996-2:2017*; Acoustics—Description, Measurement and Assessment of Environmental Noise—Part 2: Determination of Environmental Noise Levels. International Organization for Standardization: Geneva, Switzerland, 2017.
23. Plataforma BCNSentilo. Available online: http://connecta.bcn.cat/connecta-catalog-web/component/map (accessed on 16 April 2021).
24. Garrido, J.C.; Mosquera, B.M.; Echarte, J.; Sanz, R. Management Noise Network of Madrid City Council. In *InterNoise19, Proceedings of the Inter-Noise and Noise-Con Congress Conference, Madrid, Spain, 16–19 June 2019*; Institute of Noise Control Engineering: Madrid, Spain, pp. 1700–1711.
25. Portal de Datos Abiertos del Ayuntamiento de Madrid. Available online: https://datos.madrid.es/portal/site/egob (accessed on 20 February 2022).
26. Acoustic Pollution Historical Data Repository. Available online: https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=c035669177294610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default (accessed on 3 August 2022).
27. Pita, A.; Rodriguez, F.J.; Navarro, J.M. On the application of unsupervised clustering to sound pressure data from an acoustic sensors network. In *Workshops at 18th International Conference on Intelligent Environments (IE2022), Proceedings of the ISACA Conference, Biarritz, France, 20–23 June 2022*; Alvarez Valera, H.H., Luštrek, M., Eds.; IEEE: Hoboken, NJ, USA, 2022; pp. 170–179. ISBN: 978-1-64368-286-0. [CrossRef]
28. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
29. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley Series in Probability and Mathematical Statistics; John Wiley & Sons: Hoboken, NJ, USA, 1990; ISBN 9780471878766.
30. 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; pp. 281–297.
31. Kaufman, L.; Rousseeuw, P.J. Clustering by means of medoids. In *Statistical Data Analysis Based on the L1 Norm and Related Methods*; Dodge, Y., Ed.; North-Holland: Amsterdam, The Netherlands, 1987; pp. 405–416.
32. Kohonen, T. *Self-Organizing Maps*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1997.
33. Dopazo, J.; Carazo, J.M. Phylogenetic Reconstruction using a Growing Neural Network that Adopts the Topology of a Phylogenetic Tree. *J. Mol. Evol.* **1997**, *44*, 226–233. [CrossRef]
34. 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).
35. Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci.* **2015**, *2*, 165–193. [CrossRef]
36. Berry, M.J.A.; Linoff, G. *Data Mining Techniques For Marketing, Sales and Customer Support*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 1996.
37. Handl, J.; Knowles, K.; Kell, D.B. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **2005**, *21*, 3201–3212. [CrossRef]
38. Handl, J.; Knowles, J. Exploiting the trade-off–the benefits of multiple objectives in data clustering. In *Proceedings of the Third International Conference on Evolutionary Multicriterion Optimization*; Coello, L.A., Ed.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 547–560.
39. Brock, G.; Pihur, V.; Datta, S.; Datta, S. clValid: An R Package for Cluster Validation. *J. Stat. Softw.* **2008**, *25*, 4. [CrossRef]
40. Dunn, J.C. Well separated clusters and fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
41. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
42. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On Clustering Validation Techniques. *J. Intell. Inf. Syst.* **2001**, *17*, 2483. [CrossRef]
43. Halkidi, M.;Batistakis, Y.; Vazirgiannis, M. *Clustering Validity Checking Methods: Part I*; ACM SIGMOD Record: New York, NY, USA, 2002; Volume 31, pp. 40–45. [CrossRef]
44. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. *Clustering Validity Checking Methods: Part II*; ACM SIGMOD Record: New York, NY, USA, 2002; Volume 31, pp. 19–27. [CrossRef]
45. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of Internal Clustering Validation Measures. In Proceedings of the IEEE International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 911–916. [CrossRef]
46. Palacio-Niño, J.O.; Berzal, F. Evaluation Metrics for Unsupervised Learning Algorithms. *arXiv* **2019**, arXiv.1905.05667.
47. Al-Jabery, K.K.; Obafemi-Ajayi, T.; Olbricht, G.R.; Wunsch, D.C. 7-Evaluation of cluster validation metrics, In *Computational Learning Approaches to Data Analytics in Biomedical Applications*; Al-Jabery, K.K., Obafemi-Ajayi, T., Olbricht, G.R., Wunsch, D.C., Eds.; Academic Press: London, UK, 2020; pp. 189–208, ISBN 9780128144824. [CrossRef]
48. Statistical Software R. Available online: https://www.r-project.org/ (accessed on 1 June 2020).
49. Luna-Romera, J.M.; Martínez Ballesteros, M.; García-Gutiérrez, J.; Riquelme, J. External Clustering Validity Index based on chi-squared statistical test. *Inf. Sci.* **2019**, 487, 1–17. [CrossRef]