inter.noise
21-24 AUGUST
SCOTTISH EVENT CAMPUS
GLASGOW
2022

Using clustering methods to detect quality data in a smartphone-based crowd-sourced database for environmental noise assessment

Ayoub Boumchich[1]
Univ Gustave Eiffel, CEREMA, UMRAE
Allée des Ponts et Chaussées, CS 5004, F-44344 Bouguenais Cedex, France

Judicaël Picaut[2]
Univ Gustave Eiffel, CEREMA, UMRAE
Allée des Ponts et Chaussées, CS 5004, F-44344 Bouguenais Cedex, France

Erwan Bocher[3]
Lab-STICC CNRS UMR 6285, IUT de Vannes
8, Rue Montaigne, BP 561, F-56017 Vannes Cedex, France

ABSTRACT

*Environmental noise is a major source of annoyance with serious effects on health. Therefore, noise assessment is crucial to reduce these impacts. An alternative approach, based on noise measurements with smartphones, may overcome the limitations of classical assessment methods, such as simulation tools. In this way, the NoiseCapture approach was developed, consisting of measuring and sharing data, in order to produce community noise maps. Nevertheless, collected data may suffer from problems such as a lack of calibration, which lowers its quality. Quality control is therefore very important to enhance the data analysis and the relevance of the noise maps. Having trustworthy data as a reference can help in assessing the database, for example using machine-learning methods. With NoiseCapture, such data can be collected thanks to a NoiseCapture Party, an organized event, on limited space/time (*i.e. *a cluster of data). Because not all events are known by the research team in charge of NoiseCapture, and since the corresponding data can be considered of better quality, so their detection is a relevant task to trust the database. In the present communication, a clustering methodology is then proposed to automatically detect data that have been produced in such events.*

1. INTRODUCTION

Noise is a major source of pollution with impacts on health, especially in urban areas. The public authorities are trying to solve, this essential societal and health issue by putting regulations in place. In Europe, for example, directive 2002/49/EC seeks to establish an inventory of noise annoyance, to

[1] ayoub.boumchich@univ-eiffel.fr

[2] judicael.picaut@univ-eiffel.fr

[3] erwan.bocher@univ-ubs.fr

propose actions to reduce this annoyance and to communicate to citizens about their noise exposure. In this regulatory context, the key tool for decision-makers is to generate strategic noise maps.

Instead of using noise prediction software, with their inherent limits, an alternative method is to use more affordable sensor networks, allowing to densify the observation points [1], and in particular, to consider the participation of citizens as data collectors, in a crowd-sourcing approach, using smartphones as a sensor (*i.e.* a measuring instrument). NoiseCapture (NC) approach [2, 3], a part of the Noise-Planet project [4], is an example of such method.

However, a recent analysis [3] have shown that collected data with NC may suffer from several problems due to the technical limitation of the smartphone, the respect of a relevant measurement protocol, the acoustic calibration, the GPS accuracy…Quality control such as anomalies detection or data correction is therefore a very important task to enhance the database's quality and make it more fit for a relevant use. In order to develop such method, a reference database is needed. In the NC approach, such data may occurs when considering NoiseCapture Parties, *i.e.* an event that is organized, by experts, on limited space and time period to collect simultaneously a large number of contributions. Data collected in such event can be considered of better quality, since the contributors are supervised and because the smartphones are generally calibrated. Consequently, the data be considered as references. If most of the time, such events are organized by the NoiseCapture project team, it also happens that they are organized outside the framework of the project. Therefore, it would be of particular interest to be able to identify such events in the NC database, in order to feed a reference dataset. More explicitly, such method that consists in regrouping data with similar spatial and temporal characteristics, is known as spatial clustering/temporal clustering.

The objective of the present paper is therefore to implement a clustering method, using the Density-Based Spatial Clustering of Applications with Noise approach (DBSCAN) [5], to identify in the database similar events to NoiseCapture Party. Note that the term 'Noise' that is employed in the name of the method does not refer to the environmental 'noise', but to 'noise' in the data.

2. MATERIALS AND METHODS

2.1. NoiseCapture approach

NoiseCapture is an android application that was developed to let people participate in collecting noise data to generate noise maps. After starting the measurement (figure 1), the user moves along a path of his choice. An acoustic measurement is realized each second (1 s equivalent sound level and other acoustic indicators) as well as additional information such as the date/time of the measurements, the GPS location and accuracy of the measurement point, the speed of the user…When the user stops the measurement, he can provide optional information, such as the presence of sound sources (using 'tags'), the conditions of measurements (using 'tags') and the appreciation of the pleasantness (according to a scale of values). Lastly, the collected data, which are anonymous, is uploaded to the NC remote server and integrated into the NC database. The raw data collected by from the entire community is then used to create a first noise map, which can be displayed on the NC application and on the NC web platform.

Since the launch of the application in September 2017, a large amount of data is available, offering the possibility to analyze the data on over a large spatial area and over a long period of time. At that time, this database, which is distributed under the Open Data Base License [4] and updated everyday at 3:30 a.m., represents the equivalent of more than 1,080 days of 1 second measurements (more than 375,000 tracks and 93 million measurement points) collected by over 90,000 contributors over more than 200 countries. The DBSCAN algorithm that is presented is the next section, has been performed on a 3-years extraction of this NC database, from 29/08/2017 to 28/08/2020, starting from the official release version '28' of the NC application [6]. Any measurement point in the NC database without geo-localization (may be a full track or only a part of a track) have been deleted before processing

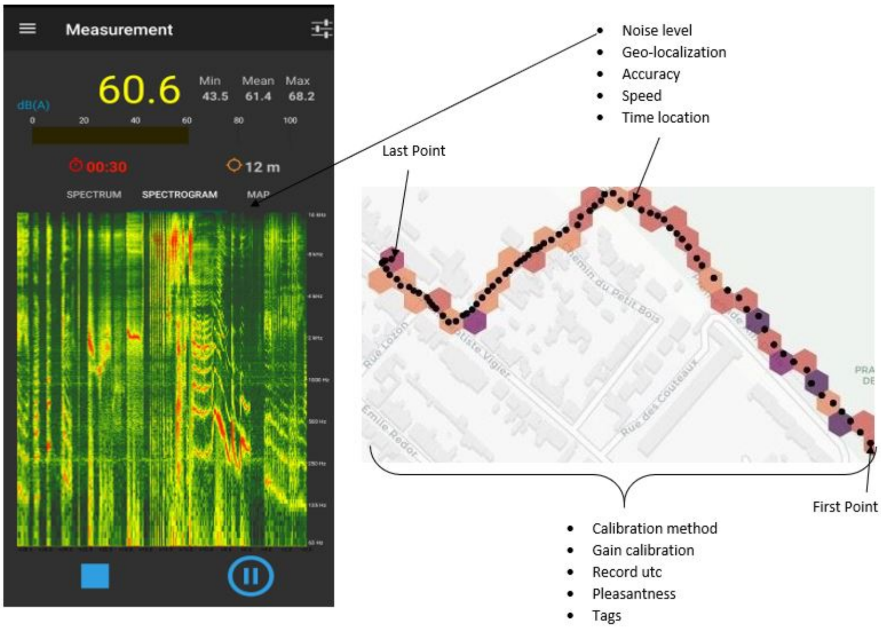

Figure 1: NoiseCapture approach. Left: screenshot of the NoiseCapture application ('Spectrogram' view). Right: Representation of a NoiseCapture measurement: track and measurement points.

the DBSCAN method. Using this NC database extraction, all the data can be organized on relational database, which can be manipulated using GIS tools.

2.2. DBSCAN method

The objective of the method is to be able to detect a cluster of points, based on the detection of a higher density of measurement points in certain spatial areas. If the density is 'abnormally' more important over a short period of time than in other areas, then it is possible to consider that these measurements were generated during an event.

The principle of the DBSCAN approach is relatively simple (figure 2). Starting from a measurement point, chosen randomly, it is sought in a circle of radius `Eps` the presence of a minimal number of points (`MinPts`). Then, starting from each of these points, new points are searched following the same principle. The procedure stops when the cluster thus built does not manage to progress any more. The process will restart after selecting a new starting point randomly. When the clustering is complete, multiple clusters might be formed .

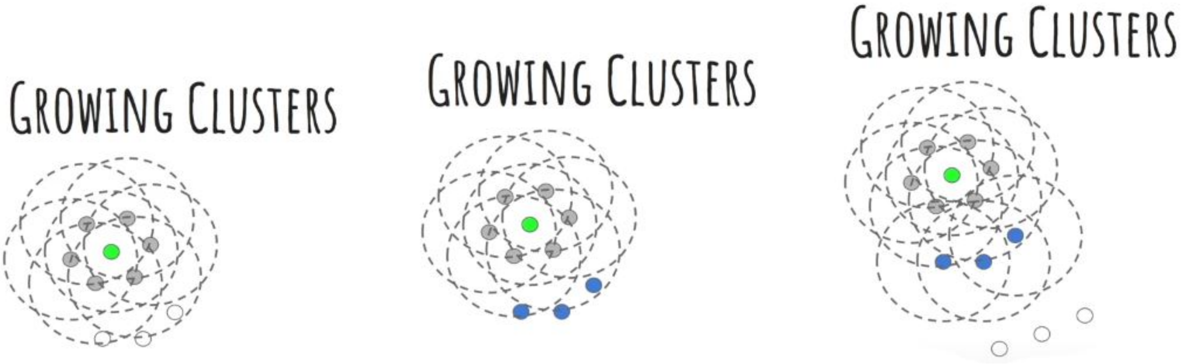

Figure 2: Graphical representation of the DBSCAN approach.

The DBSCAN method has been implemented using the DBeaver universal database tool [7] with the specific 'PostGIS' function `ST_ClusterDBSCAN` [8], which is a 2D implementation of the

DBSCAN algorithm. This function takes 3 entries: geom as geometry data type, Eps as float data type and MinPts as an integer data type. The results are saved in table called NoiseCapture_cluster, which contains the measurement point Id (pk_point) of coordinates (the_geom) and cluster Id c_id. An other variable (pk_party) is also added, to verify if the corresponding point is already part of a known NoiseCapture Party event, and, if yes, the corresponding NoiseCapture Party Id. Finally, the data and the clusters can be plotted in GIS software, such as QGIS in the present case. Note that all the tools used in this study are Open Source.

The NC database is coded using the WGS 84 (World Geodetic System 1984) format [9], which is an horizontal component of a 3D system, used, for example, by the GPS satellite navigation system. In other words, the type of the coordinates is geographic. This encoding uses 'degree' as measure distance, while the ST_ClusterDBSCAN function only takes coordinates of type 'geometry'. Therefore, a transformation to the metric projection EPSG:3857 (Pseudo-Mercator) [10] is required before applying the algorithm. This projection can only be used for data between 85.06°South and 85.06°North, which is in agreement with the NC database (NC data is bounded between 74.49°South and 78.73°North). The transformation has been done using the PostGIS function ST_Transform [11].

The entire DBSCAN processing takes around 16 hours to perform the clustering on the 3-year NC database using a low personal computer.

3. RESULTS

3.1. NoiseCapture data clusters identification

In a first step, a preliminary study (not detailed here) allowed to identify the 'best' sets of parameters (Eps and MinPts) to identify known NC Party type events. This work was done by applying the DBSCAN algorithm on spatial areas on which known NC Parties had been organized in the past. The 'best' parameters are finally those which allow to obtain the best success rate of detection of these events (in number of points and in number of tracks), which gives Eps=3000 m and MinPts=5000 points. In concrete terms, this means searching, step by step, for clusters for which each measurement point is close to 5,000 other measurement points within a radius of 3 km.

When applying the DBSCAN approach on the whole global database, 2,046 clusters are found in 68 countries; in particular, 975 clusters in the United States, followed by France (297 clusters) and United Kingdom (111 clusters). Note that these 3 countries are considered among the top 3 contributors to the NC database [3]. Among these 2,046 clusters, 1,567 (76.59%, in 40 countries) are generated by only one single user (1,155 users). Also, 252 (12.32%) clusters were generated by 2 users across 31 countries (434 users), in which 65 were users that were sole collectors of some clusters (65 of 434 are part of the 1,155 users that collected clusters with one user contribution). This leaves 227 clusters (representing 24,280 tracks and 4,548,638 points) with at least 3 users contributing to the collect of data, in France with 95 clusters, in United States with 36 clusters and in Switzerland with 9 clusters (figure 3). As expected, 19 of these identified clusters were known as NoiseCapture Party events.

3.2. Elements of validation

The analysis of the literature shows that scientific studies have been conducted on the basis of the collection of measurement data using the NC application, by teams without any link with the NC project team. At this stage of our study, it seems interesting to check if corresponding clusters have been found for these specific experiments. This would constitute a complementary validation element on the developed approach.

A first study, published in 2020, was carried in Japan [12] in order to evaluate the effect of COVID-19 pandemic lockdowns in the eastern edge of the city of Kobe. A NC measurement campaign was set up over a period of one hour (10:00-11:00 a.m.), with an averaging time of 30 seconds, at 6 locations in an urban area (as well as at a fixed position in front of a building), on 2 different days in

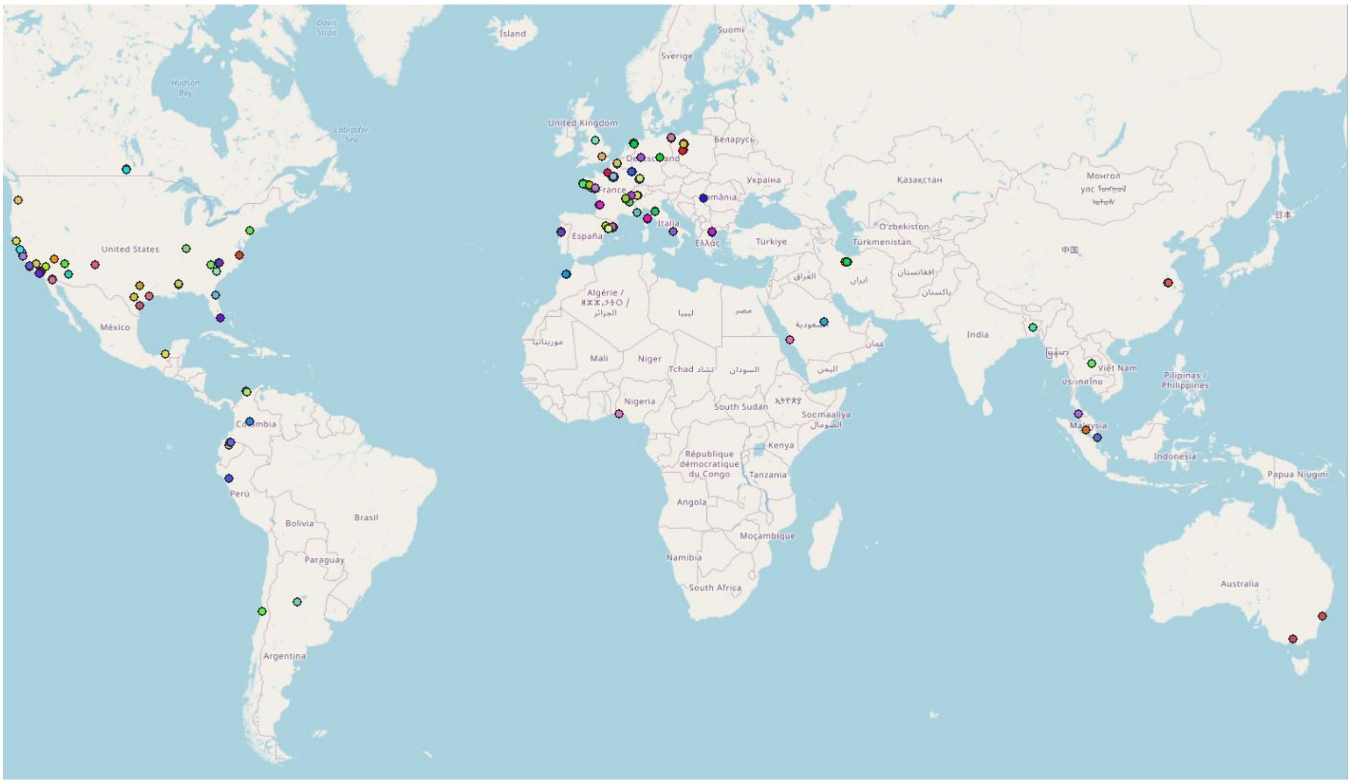

Figure 3: Representation of the NoiseCapture clusters found around the globe (the cluster points may overlay).

May 2020, during and after the lockdown period. Using our clustering approach, several clusters have been found in Japan in 2020, but none that could correspond to this measurement campaign (however, 'individual' measurements have been identified over this period in this area by some NoiseCapture users, which could be associated *a posteriori* to this measurement campaign). This is probably due to a number of measurement points (2 days x 7 locations x 120 measurements (*i.e.* 1 hour with an average on 30 s) = 1680 points) that is *a priori* lower than the number of measurement points for detecting clusters (`MinPts`=5000 points). It is however interesting to note that two clusters have been detected on this area in July 2020 and in August 2020 (may be additional measurements to the initial experimentation), both collected by the same 2 users that carried out the event on May 2020.

Another NC measurement campaign was carried out in `India` [13], in three urban zones, corresponding to specific noise ambiances of the Lucknow city: 'Polytechnic chauraha', 'Hazrat ganj chauraha' and 'Haniman chauraha'. In total, measurements were made at 14 locations, over 3 time periods of a day (morning, afternoon, evening), every 10 minutes. Here again, the number of measurement points was *a priori* not sufficient for our clustering methodology to detect this event as a cluster.

Another experimentation, involving the comparison of several smartphone noise measurement applications was conducted in 2018 [14]. Measurements were performed over 3 periods of one day (7:00-9:00, 15:00-17:00 and 19:00-21:00), two times, in an area of the city of Zagreb, Croatia. The reference [14] (a student report) does not give much indication about the sampling of the measurements and the exact date of the measurements, but the application of our methodology has identified an event that could correspond to a part of this experimentation, located at the same place, between March 12 and 18, 2018, and consisting of 3 tracks for 6,417 measurement points.

The last event that can be found in the scientific literature took place in the City of Cairo in Egypt on August 2018, in order to study the effect of noise pollution on patient undergoing surgery [15]. The corresponding measurement were correctly found by the clustering approach, composed of 5 tracks with 9,418 points, and collected by only 1 user.

4. CONCLUSION

The objective of this study was to set up a method to detect and group NC data (*i.e.* a clustering approach) that could have been realized in the context of a specifically organized event, such as the NC Parties organized by the NC project team. The final objective is to be able to use this data, which is judged to be of better quality, to create a reference database which would then be used, *via* quality control methods, to give a quality criterion to the other data in the NC database.

The proposed approach uses the DBSCAN algorithm, a density-based spatial clustering method, which aims to group data in function of the density of points, meaning that higher density zone can be considered as a cluster of data, in comparison with lower density zones. Following a preliminary study to define the most appropriate parameters of the DBSCAN method, the approach was used on 3 years of NC data and has identified about 2,000 clusters over the globe. Most of the NC Parties were naturally found (which was expected, since the parameters of the method were optimized on these events), but other clusters, corresponding to experiments using the NC application, published in the literature, were also found.

It is however very clear that not all the clusters that are found correspond specifically to the expected type of events. It could be useful to be more demanding on the DBSCAN parameters, to reduce the number of clusters obtained, by limiting ourselves to larger amounts of data, even if it means missing smaller events.

REFERENCES

[1] Judicaël Picaut, Arnaud Can, Nicolas Fortin, Jeremy Ardouin, and Mathieu Lagrange. Low-Cost Sensors for Urban Noise Monitoring Networks—A Literature Review. *Sensors*, 20(8):2256, january 2020. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Judicaël Picaut, Nicolas Fortin, Erwan Bocher, Gwendall Petit, Pierre Aumond, and Gwenaël Guillaume. An open-science crowdsourcing approach for producing community noise maps using smartphones. *Building and Environment*, 148:20–33, 2019.

[3] Judicaël Picaut, Ayoub Boumchich, Erwan Bocher, Nicolas Fortin, Gwendall Petit, and Pierre Aumond. A smartphone-based crowd-sourced database for environmental noise assessment. *International Journal of Environmental Research and Public Health*, 18(15), 2021.

[4] Noise-Planet website. Noise-Planet - Data. https://data.noise-planet.org/index.html, 2021. (Accessed on 2021-01-13).

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, pages 226–231, 1996.

[6] Judicaël Picaut, Nicolas Fortin, Erwan Bocher, and Gwendall Petit. Noisecapture data extraction from august 29, 2017 until august 28, 2020 (3 years). https://doi.org/10.25578/J5DG3W, 2021.

[7] DBeaver. https://dbeaver.io/.

[8] ST_ClusterDBSCAN. https://postgis.net/docs/ST_ClusterDBSCAN.html.

[9] WGS 84 - WGS84 - World Geodetic System 1984, used in GPS - EPSG:4326. https://epsg.io/4326.

[10] WGS 84 / Pseudo-Mercator - Spherical Mercator, Google Maps, OpenStreetMap, Bing, ArcGIS, ESRI - EPSG:3857.

[11] ST_Transform. https://postgis.net/docs/ST_Transform.html.

[12] Kimihiro Sakagami. How did the ‘state of emergency’ declaration in Japan due to the COVID-19 pandemic affect the acoustic environment in a rather quiet residential area? *UCL Open Environment*, August 2020. Publisher: UCL Press.

[13] R. Dubey, S. Bharadwaj, M. I. Zafar, V. Bhushan Sharma, and S. Biswas. Collaborative noise mapping using smartphone. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B4-2020:253–260, 2020.

[14] Antonio Njegovan. Analysis of free applications for noise measurement. https://www.bib.irb.hr/949969?rad=949969, 2018.

[15] Hany Mohammed El-Hadi Shoukat Mohammed, Sahar Sayed Ismail Badawy, Ahmed Ibrahim Hussien Hussien, and Antony Adel Fahmy Gorgy. Assessment of noise pollution and its effect on patients undergoing surgeries under regional anesthesia, is it time to incorporate noise monitoring to anesthesia monitors: an observational cohort study. *Ain-Shams Journal of Anesthesiology*, 12(1):20, June 2020.