"Science Stays True Here"
Advances in Ecological and Environmental Research (ISSN 2517-9454, USA) | Science Signpost Publishing

Collecting and Processing Crowdsourced River Measurements Using Innovative Sensors

Rigos Anastasios*, Krommyda Maria, Theodoropoulos Theodoros, Tsiakos Valantis, Tsertou Athanasia, Amditis Angelos

Institute of Communication & Computer Systems (ICCS), National Technical University of Athens, Athens, Greece.

Received: October 09, 2019 / Accepted: November 12, 2019 / Published: Vol. 5, Issue 01, pp. 13-25, 2020

Abstract: This study presents the process of crowdsourcing and processing of environmental data collected by volunteers. The data consist of river measurements collected during the pilot activities (experiments) of SCENT, an EU-funded research project. The data were collected by volunteers during organised campaigns at the Kifisos river (Attika, Greece). The volunteers used their smartphones/tablets and low-cost portable sensors to collect the measurements. They collected images of water level indicators for the water level, videos of a pre-defined floating object for the water surface velocity and sensor measurements for air temperature and soil-moisture. Since the campaigns involved volunteers with different social and educational backgrounds, in their majority not familiar with the strict process of collecting scientific data, the collected data are further processed in order to ensure a high quality degree.

Key words: Flood modelling, innovative sensors, crowdsource data, data analysis, data quality control.

1. Introduction

Expensive and hard to maintain in-situ monitoring systems, producing consistent and accurate data in the course of time, are used for the creation of flood risk prediction models (Sorooshian et al., 2008). Selecting the area where such in-situ monitoring systems are installed is a time consuming process that takes into consideration previous incidents of extreme weather phenomena, the affected areas and the local population. As the global climate changes, extreme weather phenomena become more often and the need to monitor wider areas that have not been previously studied becomes imperative. This need has raised the question of whether the development of flood models can be disengaged from the in-situ sensors to cover larger areas of interest with a minimized cost (Kei & Kawaguchi, 2016; Mousa et al. 2016; Lo et al., 2015).

Corresponding author: Rigos Anastasios, Institute of Communication & Computer Systems (ICCS), National Technical University of Athens, Athens, Greece. Email: anastasios.rigos@iccs.gr

Due to both the climate change and the raise of the frequency of extreme weather phenomena, more people are affected by environmental issues and become aware of the need to monitor them (Hayes, et al. , 2008, Ferdoush & Li, 2014, Penza et al.,2014). As a result, they are willing to offer their time to support the scientists in monitoring areas of interest and collect measurements as needed aiming to gain a better understanding of their environment, the potential dangers and the needed mitigation measures. The utilization of crowdsourced data, however, comes with the challenge of data quality. The volunteers differ in knowledge and experience in the data collection procedure, especially when they are required to follow strict guidelines to ensure the accuracy and consistency needed for scientific purposes. More often than not, this results in data that vary in accuracy and quality, may be biased or include noise and invalid values. At the same time, scientists are used to depend on consistent and accurate data that are produced by methods that eliminate any noise and protect their research from invalid input.

To minimize the potential bias and inconsistency of asking the volunteers to collect measurements in a scientific way, they are asked to perform simplified tasks and collect multimedia that are then processed to extract the needed measurements. To begin with, the volunteers are asked to collect images containing a Water Level Indicator (WLI) sub-merged into water, which are processed by the Water Level Measurement Tool (WLMT) that uses image recognition techniques to extract the water level. They are also asked to capture video containing a pre-defined floating object moving on the surface of a water body, which are processed by the Water Velocity Calculation Tool (WVCT) that uses video processing algorithms to extract the water surface velocity. Finally, they are asked to use the Scent Measure, a mobile application that connects their devices with easy-to-use portable sensors and collects measurements for air-temperature and soil-moisture. The tools have been designed so that a high accuracy of trust can be achieved from images, videos and sensor measurements taken from smartphones and low-cost portable sensors.

To further alleviate the barrier of the quality of the collected data, they are further analyzed and processed to identify results that are invalid or of low quality and should not be used in the scientific research. Different methods/tests are applied in order to identify these measurements among the useful ones.

The crowdsourced river measurements are used to develop improved flood models with dramatically reduced equipment cost, as the water level indicators, the floating object and the sensors are low-cost and re-usable, while effectively covering large areas of interest.

This work contributes in the processes of:

(a) collecting environmental data from un-experienced users using only common smartphones and low-cost sensors

(b) transforming of the collected data into a numerical dataset

(c) determining measurements in the collected database that can be the result of failures on the data collecting procedure.

It has to be mentioned that this study focuses only in the process of collecting and evaluating the data collected from volunteers and doesn't focus with they later usage, such as in flood modeling algorithms.

2. Methodology

The Scent project incorporates pilot-demos/experiments for demonstrating the concept of citizen observatories in practice and for demonstrating the use of the developed applications. In such experiments, participants range in age, professional expertise and familiarity of collecting scientific data. There are several challenges in organizing and running such large-scale open field activities that include non-expert people and span over several months. Such issues include the prompt dissemination of the event, informing the general public and engaging them to get involved and participate, organizing groups per day & site, selecting routes to visit, providing a short first-day training and application installation workshop, minimizing transportation costs and delays as well as guiding them throughout the activity. In this work, the data from the study area of the Kifisos river (Attica, Greece) collected at the campaign organized at 15-17 of Nov/2018 are processed.

The volunteers collect data using a smartphone application which was created for this purpose. This application collects images, videos and measurements from a sensor that communicates with the Bluetooth Low Energy (BLE) protocol. Afterwards, all the collected data are sent to a central server and are processed by the three tools. The analysis of the three tools follows.

2.1 Water Level Measurement Tool (WLMT)

The volunteers are asked to capture images of water level indicators at areas of interest. The images are then forwarded to the WLMT to get processed. The first step is the pre-processing of the image through a state-of-the-art enrichment process in order to improve the image quality as much as possible. This is done as most of the images are captured at an angle, contain shadows and light reflections elements that may affect the quality of the results.

The next step is the identification of the numbers available in the image. Initially the algorithm identifies all the numbers that are available on the image as single digits. Then it tries to make pairs based on the spatial distribution of the numbers over the image. Finally the number located at the lowest part of the image (Fig.1) is identified. If this number is the lowest number identified among all pairs then this is the water level measurement. In case this is not the smallest number identified then the image is invalidated. The identification process is relying on a pre-trained Cascade Classifier (Hu & Collomosse, 2013), which was trained with a

series of images taken at the water level indicators at the Kifisos river.

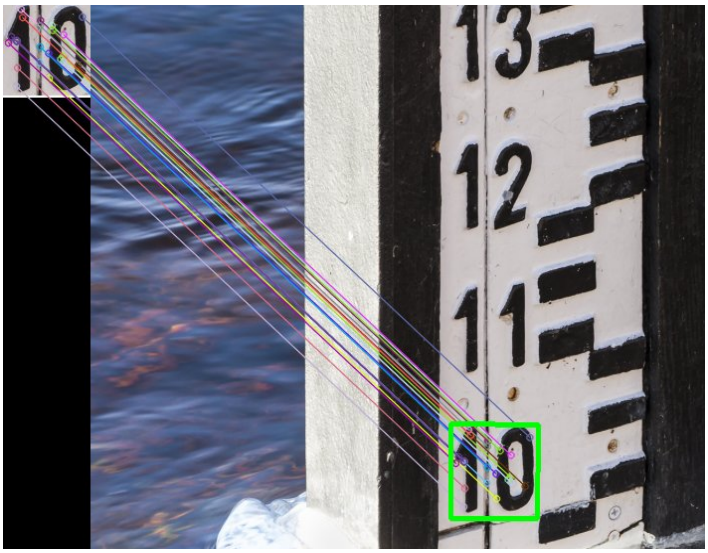


Figure 1. Number extraction from a water meter installed inside the river.

2.2 Water Velocity Calculation Tool (WVCT)

The volunteers are asked to capture video of a pre-defined floating object (a yellow tennis ball) as it passes in front of them following the river's course (Fig.2). The collected video is first stabilized to ensure that any intentional or unintentional movement of the camera is eliminated as much as possible, removing the associated noise. After the video stabilization, the video-frames are examined one-by-one until 2 consecutive frames that include the tennis ball are discovered. The detection of the tennis ball is based on a color filter that looks for the bright yellow color of the ball. The filtering is expected to have very accurate results as the tennis ball has a regulated color, a very distinct bright and artificial yellow that it is unlikely to be found in a natural environment such the course of a river. Using these two consecutive frames the object's displacement is calculated. The displacement is calculated by computing the average optical flow of the area of the frame that includes the tennis ball. The average optical flow of the rest of frame is subtracted from this calculation to eliminate any camera movement that is still present in the video frames. The process is repeated until a frame without the pre-defined floating object is reached; when such frame is located, it determines the end of the video. The ball's displacement (which is calculated in pixels) is then calibrated using the dimension of the pre-defined object. This is easily support once more by the choice of the tennis ball as the floating object, given that it is always 6cm in diameter. An example with the extraction procedure of the tennis ball from video frames is presented in Fig.2.

In case that the pre-defined object is not present in the video or it is presented in fewer frames than the frames per second of the video, giving less than a second of useful material, or the calculated displacement of the tennis ball is less than the identified size of the ball then the video is classified as invalid and is not used to calculate water velocity. In the case of a valid video the water velocity is extracted along with a confidence level. The confidence level is calculated based on the duration video and the time that the floating object was detected as well as the proportion of the identified size of the tennis ball to the overall displacement. The confidence level was calculated as:

$$ConfidenceLevel = \frac{1}{2}\left(\frac{ValidVideoFrames}{TotalVideoFrames} + \frac{TennisBallSize}{Displacement}\right) \quad (1)$$

(a) (b)

Figure 2. (a) A video frame containing the tennis ball. The tennis ball extracted from the previous frame using the WaterVelocity Calculation Tool is encircled. (b) The previous image focused around the area of the ball.

2.3 Scent Measure

The Scent Measure is an application designed for the volunteers that collect data during the experiment. The application is supported in the back end by an algorithm that evaluates the collected data after the experiment ends.

Volunteers use the *Flower care* by *Xiaomi* sensor (Fig.3) in order to collect environmental measurements. From the data that the sensor is able to collect (Fig.3a), only the air-temperature and soil-moisture values are used in this study. An Android OS application was developed specifically to use with the aforementioned sensors (Fig.4). This application is available for mobile devices with Android OS, it requires devices with Bluetooth version 4.1 (Bluetooth Low Energy) connectivity and offers a maximum measurement update frequency of 1/15 Hz (i.e. one measurement per 15 seconds). The application is available on the *Google play Store*. Using the application, measurements from the sensor are collected, enriched with information from the android device such GPS and timestamp and afterwards uploaded to a central server.

Figure 3. The *Flower care* sensor that collects the air-temperature and soil-moisture data.

(a) Description of the included sensors. (b) Using the sensor.

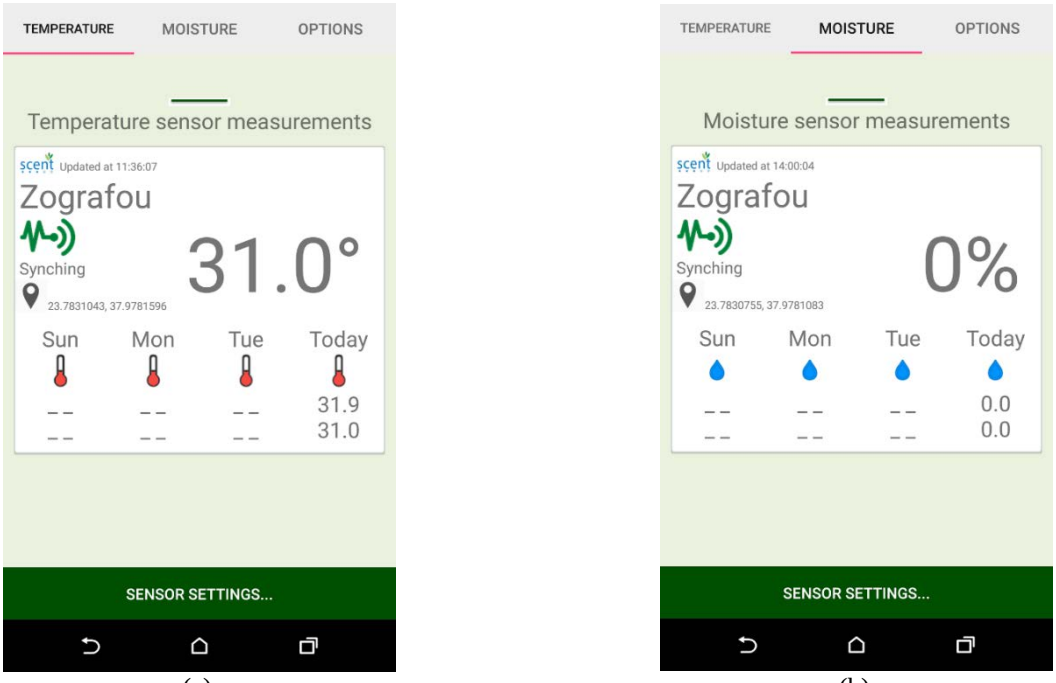
Each measurement contains a unique ID, the time that the data were collected, the ID of the user that collected them, the GPS coordinates (latitude, longitude and altitude) and GPS accuracy (a parameter taken from the Android OS, measured in meters), the air-temperature measured in °C and the soil-moisture as a percentage (in the interval 0-100%). After the experiment ends, the collected data pass a first-stage filtering in order to remove data considered as invalid. For both the temperature and soil-moisture measurements, the

criteria-tests that define fault data are:

- A measurement being the only one provided in a radius of 10 meters.
- A specific user being the only one providing measurements in a radius of 10m.
- GPS accuracy being greater than 20 meters.

For the temperature data, an additional range test is conducted. Temperatures outside of a (predefined) range are automatically invalidated. This range was calculated as $\overline{T} \pm 10°C$, where $\overline{T}$ is the daily mean temperature the volunteers provided.

For the soil-moisture measurements, an additional test called the *15"-test* is applied. The analysis of this test follows. While searching the soil moisture database for erroneous measurements, data with soil moisture values equal to zero were discovered. The challenging task is to discover if values of soil moisture equal to zero have occurred by measuring while the sensor isn't inserted inside the soil or if they occurred by measuring dry soil. As it has been checked in laboratory tests, the sensor can provide a valid measurement of 0% moisture level when measuring dry soil without any roots of alive plants inside, which is reasonable since the soil in such cases behaves as an infinite resistance (zero conductivity). In order to validate the zero moisture levels as fault data or not, the following test was applied: if a user provided a 0% moisture level and within the next 15 seconds (the sensors provide a new measurement every 15 seconds) he provided a measurement greater than 0%, then his first measurement is considered as fault. The aforementioned test is called the *15"-test* in this study and but is also known as the 'delta test' in bibliography (Taylor & Loescher, 2013).



(a) (b)

Figure 4. Screenshots of the smartphone application that collects the sensor's measurements.

3. Results and Discussion

During the experiments, small groups of people (3-6) were collecting data in nearby places. Each of these users was collecting the data of water level, water velocity, temperature and soil-moisture at nearby places within small time intervals with other users. So the collected data should consist of groups of similar measurements except of the soil-moisture; since these measurements can differ in nearby places, depending if the soil is absolutely dry or not. So the collected data were clustered into a total of K groups (called clusters) by clustering their coordinates (Latitude and Longitude, taken from GPS coordinates) and their timestamps. The clustering process was performed, using the K-means clustering algorithm (Lloyd, 1982, Forgy, 1965). Afterwards, data inside each cluster were examined in order to locate users that provided measurements dissimilar to the ones provided by the rest of the nearby users. These methods reveal outliers in the collected data.

3.1 Analysis of the water level data

The water lever data provided by the WLMT consist of 114 measurements of water level. The first challenge is to identify if a topological or a hydrological indicator was used. While efforts were made during the campaigns to ensure that the data were used using only hydrological indicators, in some cases images contained topological rods. This is an issue as the way the values are depicted in the two rods are different. The hydrological rods display the levels in centimeters while the topological in decimeters, so for a hydrological rod an identified value of 9 corresponds to 9 cm while for the topological to 90cm (9dm).

The collected water-level data were clustered into $K = 13$ clusters using their coordinates and timestamps. The number K of the clusters was determined using the average silhouette method (Kaufman & Rousseeuw, 1990) (Fig.5). This method indicated that the number of clusters should be $K = 13$ (Fig.5).

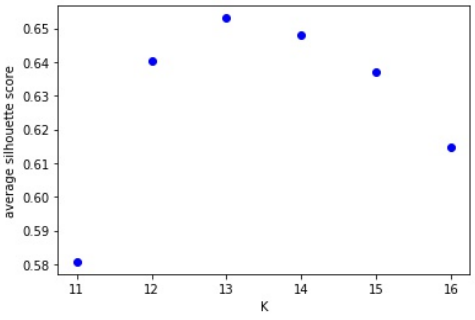


Figure 5.The average scores by the silhouette method for different values of clustering centers K.

In the resulting clustered data, a great variance was observed together with a lot of data that are considered to be outliers (Fig.6a). To avoid such data distribution, the 'sigma' test was applied to these data (Taylor & Loescher, 2013). According to the sigma test, the sampling distribution of the collected data in nearby timestamps, should follow the normal distribution $N(\mu,\sigma^2)$ and the mean value and standard deviation can be approximated by the characteristics of the sample, i.e. $\mu \simeq \overline{x}$ and $\sigma^2 \simeq s^2$. In order to remove outliers and possible fault measurements, for each cluster the data were filtered by keeping only those in the interval (μ-σ, μ+σ), so about the 32% of data of each cluster were removed and the resulting clusters do not contain extreme values (outliers) (Fig.6b). As Fig.6b presents, more than half of the resulting clusters still have a great variance; this should not be expected by a river with steady flow. Manual inspection of the outliers showed that there were cases where digits were incorrectly identified by the classifier. In most cases this was mainly due to brought material that was dirtying the water level indicator making the lower parts of it unreadable, resulting on higher extracted readings, and in some others it was due to the wear of the letters printed on the indicators by the weather and time that made some digits resemble others such as a 7 resembling a 1 or an 8 resembling an 0.

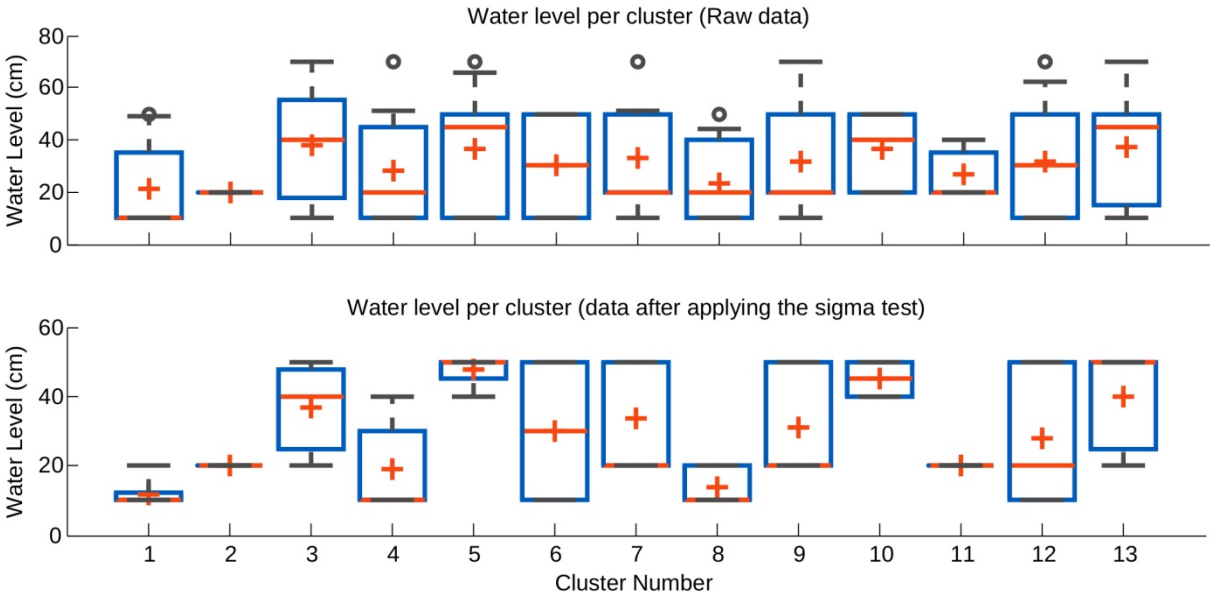


Figure 6. Box-plots of the Water lever per cluster from: (a) raw data from the WLMT and (b) after applying the sigma test. The red crosses represent the mean values.

3.2 Analysis of the water velocity data

In order to evaluate the quality of the velocity extracted by the WVCT, these values were separated into *K* groups. For this separation, the K-means clustering algorithm was applied and the data were grouped into *K* = 12 clusters according to their location. The four of these 12 clusters were re-clustered afterwards according to their timestamps, because they contained data collected on different days. So finally the collected data were separated into 16 clusters. Figure 7 presents the box plots that show the range of the collected data. In some cases, there are clear outliers; these were examined further to identify the causes. The main cause was identified as the camera stabilization and the noise in the captured frames.

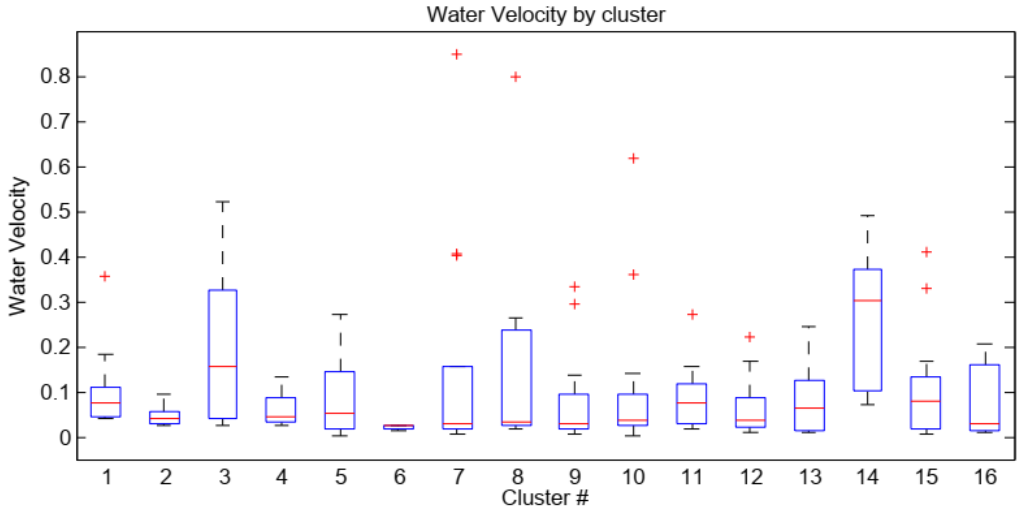


Figure 7. The Water velocity data clustered into *K*=16 groups.

3.3 Analysis of the sensors' data

The sensors data consist of *N* = 248 measurements of temperature and soil-moisture collected at 16/Nov/2018. These were clustered into *K* = 4 clusters, with spatial distribution the one presented in Fig.8. Exploring the temperature data in each cluster, in some clusters the temperature seems to be uniform distributed inside its range (Fig.9a) but in some other (Fig.9b) the temperature doesn't seem to be so uniform distributed, so it is suspected that one user or one sensor is collecting wrong measurements. Since each user used a different sensor, locating the user is enough.

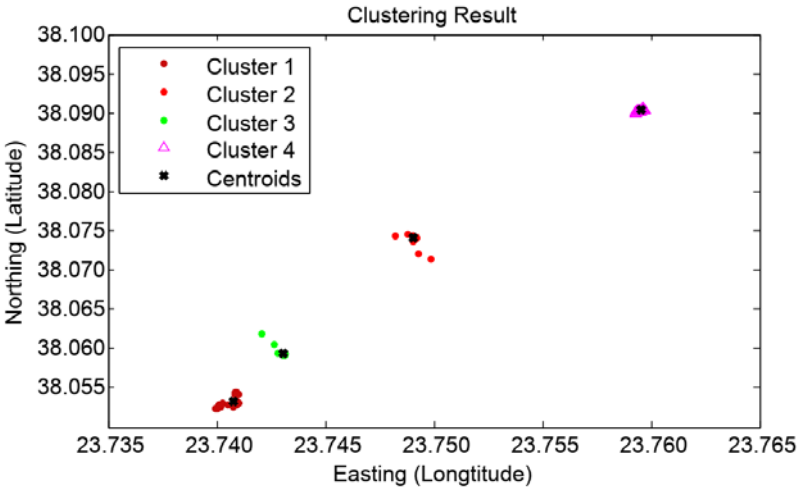
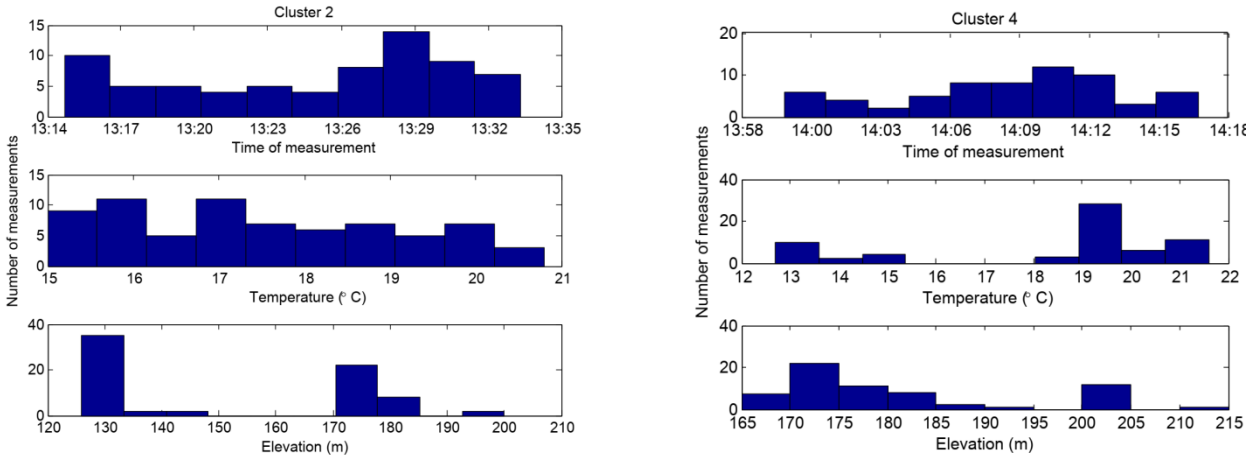


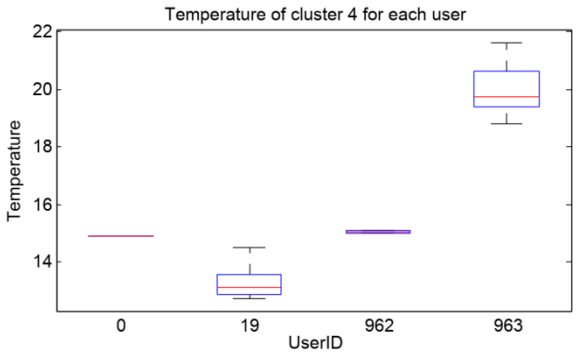
Figure 8. The spatial distribution of measurements collected using sensors and their clustering into $K = 4$ groups.



(a) (b)

Figure 9. The distribution of time, temperature and elevation in two clusters of the data. Time refers only to the 16/Nov/2018.

In order to validate if really some users are providing data in different ranges than the others, the only visual method to check it is using box-plots, as the ones of Fig.10a. This box-plot diagram suggests that possibly the user 963 is providing the 'different' data. In order to prove that, the one-way ANOVA test was used (Welch, 1952, Zar,1998) (Fig.10b). This test proves that the differences between the data of different users of Cluster 4 are significantly different ($p = 6.47 \cdot 10^{-34} < 0.01$), so it can be assumed that the user 963 provided fault measurements.



(a)

ANOVA table					
Source	SS	df	MS	F	Prob>F
Group (Between)	499.2	3	166.4	251.2	6.47e-34
Error (Within)	39.7	60	0.66		
Total	539.0	63			

(b)

Figure 10. (a) The distribution of time, temperature and elevation in two clusters of the data. (b) The One-way ANOVA table that compares the differences of values of the temperature data collected by different users in the Cluster 4.

4. Conclusion

In this study, the process of collecting of environmental data with the help of volunteers around a river is presented. Afterwards, the collected data are pre-processed by three tools to transform them into numerical data and make a first filtering on sensors data. Those data contain fault measurements caused by (a) the inexperience of the volunteers, (b) by the used sensors or (c) by the tools that extract data from videos and images. The results section presents different methods to make a second stage of filtering to the data and, as is presented, 'noisy' data caused by specific users or specific sensors can be identified. Since the data are collected by small groups of people located nearby each other at about the same time, data clustering methods provide a very effective solution for such cases.

Acknowledgements

This research has been financed by European Union's Horizon 2020 research and innovation programme under grant agreement no 688930, project SCENT (Smart Toolbox for Engaging Citizens into a People-Centric Observation Web). Content reflects only the authors' view and European Commission is not responsible for any use that may be made of the information it contains. For more information about the SCENT project visit the website https://scent-project.eu/

References

[1]. Sorooshian S., Hsu K.-l., Coppola E., Tomassetti B., Verdecchia M., Visconti, G.: Hydrological modeling and the water cycle: coupling the atmospheric and hydrological models, vol. 63. Springer Science & Business Media, (2008)

[2]. Hiroi Kei, and Nobuo Kawaguchi. "FloodEye: Real-time flash flood prediction system for urban complex water flow." 2016 IEEE SENSORS. IEEE, 2016.

[3]. Mousa Mustafa, Xiangliang Zhang, and Christian Claudel. "Flash flood detection in urban cities using ultrasonic and infrared sensors." IEEE Sensors Journal 16.19 (2016): 7204-7216.

[4]. Lo Shi-Wei, et al. "Visual sensing for urban flood monitoring." Sensors 15.8 (2015): 20006-20029.

[5]. Hayes, Jer, et al. Evaluation of a low cost wireless chemical sensor network for environmental monitoring. In: SENSORS, 2008 IEEE. IEEE, 2008. p. 530-533.

[6]. Ferdoush Sheikh and Li Xinrong. Wireless sensor network system design using Raspberry Pi and Arduino for environmental monitoring applications. Procedia Computer Science, 2014, 34: 103-110.

[7]. Penza Michele, et al. Towards air quality indices in smart cities by calibrated low-cost sensors applied to networks. In: SENSORS, 2014 IEEE. IEEE, 2014. p. 2012-2017.

[8]. Hu Rui, and Collomosse John. "A performance evaluation of gradient field hog descriptor for sketch based image retrieval." Computer Vision and Image Understanding 117.7 (2013): 790-806.

[9]. Taylor, J. R., & Loescher, H. L., “Automated quality control methods for sensor data: a novel observatory approach”. Biogeosciences, 10(7), 4957-4971, 2013.

[10]. Kaufman L., and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: John Wiley & Sons, Inc., 1990.

[11]. Lloyd S., “Least squares quantization in PCM,” IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[12]. Forgy, Edward W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics, 1965, 21: 768-769.

[13]. B. L. Welch, “On the comparison of several mean values: an alternative approach,” Biometrika, vol. 38, no. 3–4, pp. 330–336, Dec. 1951.

[14]. Zar, J. H. Biostatistical Analysis 5th ed New York. 1998.