ISSN 1870-4069

Open Data Analysis to Identify Behavior in Water Quality for Mexico City

Emmanuel Alejandro Martínez Casillas

Instituto Politécnico Nacional,
Mexico

ecasillas.emc@gmail.com

Abstract. This manuscript is based on the project "Prototype of monitoring network and forecasting model to know water quality in Mexico City", with register number "TTT-2021/1-15", the project is made up of three modules: 1) a data collection module, 2) a data processing module, 3) a data presentation module. Module 1) is made up of IoT sensors, which will collect data on water quality, considering the chemical, physical and biological characteristics of water, which are considered suitable for human use in accordance with NOM-012- SSA1-1993. Once the information is collected, it will be stored in 2) a data repository with a data manager in the cloud, the communication mechanisms between the modules will be web services. Module 2) will oversee calculating the water quality forecast for a specific date or period, for this, machine learning algorithms will be used for getting the predictions and for detecting the anomalies (Outlier values), the detection of outliers in the geographic analysis, will represent the origin of a problem. Finally, module 3) will display the graphs and maps using the dashboard technique (dashboard in a responsive web interface).

Keywords: Exploratory, water quality, clustering.

1 Introduction

Currently many refer to the current era, as the data era [1], this since people currently consider information as the "new gold" [6], so it can be understood, the great importance of data in our times, as well as what can be done and achieved with them is mentioned, of course, for that we must first go through the real challenge, which is to be able to analyze and extract the information from said data.

In Mexico, the culture about data has been growing, however, it is still very little, since science in our country there is an endless amount of data, which has not yet been treated and its potential has not been discovered, since Mexico is in the 76th place within a list of 140 countries, in the index of adoption of information technologies favorable to economic development.

For example, one of the sectors of which there is a lot of information data still unprocessed on the water system, day by day many people within the CDMX consume water with a quality, which in some cases is harmful, if it is consumed or used in a certain amount [10]. Then people question the National Water Commission

(CONAGUA), the reason why the quality of the water they receive has been deteriorating over the years. In many cases, the detection and correction of this problem is very slow, and it is until the water quality has decreased by half that a true notion that there is a problem is obtained, so we can conclude that It is necessary to have a system that predicts or is capable of recognizing according to the successive behavior of water quality, what problem could be the cause of said decrease in water quality and combat it before it spreads and reaches a critical point at which the water quality is lower than that recommended for human consumption and use.

The problem of calculating or knowing the quality of water in CDMX began a long time ago, serious consequences have already occurred, as happened in the 90's, where a large part of the population suffered from a disease called cholera, which is a bacterial disease, spread through contaminated water, causing severe diarrhea and dehydration in people, which can be fatal within hours.

Based on articles and reliable sources such as scientific studies by universities and government reports [9], the idea arises of being able to generate a system that helps to solve this problem, so important, since we are talking about water, which is an invaluable resource for the daily life of all human beings and whose conservation and care, also within the factors that define what quality of life is, is the basic service of water supply, in times like the pandemic, having good quality water becomes a basic requirement to be able to face diseases.

The principle of the first law of geography is understood, which says that the objects closest to each other in space have more relationship [7], and this is important when we talk about a network such as the CDMX, the which connects all municipalities, but does not obtain the same water quality results in each municipality, This concept is further deepened, but it is also seen that within a network there is also a relationship through time [8], this is very helpful for understanding more, how this problem has advanced within the network through space and time, so that thanks to concepts we can better understand how the problem of water quality is a problem that grows over the years and spreads across the network.

1.1 Problem Statement

At present only 33.2 percent of the water quality monitoring sites operated by the National Water Commission (CONAGUA) meet the acceptable limits of water quality, this means that 66.8 percent of the sites have low water quality. This represents a severe problem, since there are people who can be intoxicated without being informed of what they are really consuming, in the water of the CDMX, because the information is not easy to obtain and the statistics found in the government reports, they are statewide, so, in smaller regions, the information is not accurate. Currently Mexico has a national water quality measurement network, but this only monitors the main bodies of water in the country, that is, it monitors very large ranges, so it is more difficult to obtain accurate information from smaller communities, such as the case of the CDMX mayors, where it is not known whether the affected area is the entire network or only a part of it.

Within the network that CONAGUA currently manages, in CDMX there are few monitoring sites and most of these do not comply with the acceptable limits of water quality, this shows that in CDMX the quality is poor, but the current network does not consider a model that forecasts for a specific period. With these data on a dashboard, you can support decision-making and know or infer causes of the problem or be able to monitor the data, to know how the water quality has improved or worsened over time and in certain areas.

This problem is a reality in Mexico, where cases of diarrhea have been increasing, this because the presence of contaminants in the water has increased, between 7 and 12 percent, this is serious and means that the risk of finding contaminants in water is comparable to the risk in European countries and some parts of China.

2 State of the Art

In this space there will be a comparison of scientific articles which have the same objective as the proposed project, but they use different techniques, also at the bottom of this section is the table 1, with a deeper comparison.

2.1 Spatial Variation Impact of Landscape Patterns and Land Use on Water Quality Across an Urbanized Watershed in Bentong, Malaysia

This research aimed to quantify and illustrate the effects of land use and landscape configuration on water quality in Bentong River, Malaysia. The study sampled 22 sites during the normal and wet season in 2018. FRAGSTATS was used to analyze the spatial change of landscape metrics. The results showed that water quality was closely associated with landscape configurations and land cover proportions. It also indicated that the susceptibility of water to degradation increased with a great interrelation of different land uses [2].

2.2 Quantifying the Contributions of Structural Factors on Runoff Water Quality from Green Roofs and Optimizing Assembled Combinations Using Taguchi Method

In this study, runoff plots of extensive green roofs with Taguchi designed structural factors and levels were constructed and simulated rainfall experiments were conducted. Influences of structural factors on outflow water quality of green roofs were statistically assessed and quantified. Runoff water quality of green roofs with assembled combinations at specific levels were optimized and predicted by using the Taguchi method [3].

Table 1. Comparison of projects.

Projects	Sensor type	Type of data analysis	Presentation of the results	Applied on
P1	Different	Different	Different	Malaysia
P2	Different	Equal	Different	China
P3	Different	Equal	Equal	China
P4	Different	Equal	Equal	

2.3 A Holistic Assessment of Water Quality Condition and Spatiotemporal Patterns in Impounded Lakes Along the Eastern Route of China's South-to-North Water Diversion Project

Water quality is one of the key determinants for assessing effectiveness and success of water diversions, but rarely studied at a spatial scale that crosses large river basins. Multiple statistical methods and the water quality index (WQI) were used to assess overall condition and detect spatiotemporal patterns of water quality in a series of impounded lakes along the Eastern Route of China's South-to-North Water Diversion Project [4].

2.4 Water Quality Related to Conservation Reserve Program (CRP) and Cropland Areas: Evidence from Multi-Temporal Remote Sensing

Therefore, aiming to quantify the relationship between CRP enrollment, cropland area, and the downstream water quality, we propose an approach that combines archived survey data, water quality monitoring data (total nitrogen content, TN), and remote sensing observations. By constructing the long-term datasets (1999–2014 annually) in Google Earth Engine and conducting multiple linear regression, they explained 79% variation in TN by the area of total CRP enrollment (CRP_all), area of corn and soybeans croplands, and discharge [5].

3 Methodology

A prototype was built to know the quality of the water that is delivered to a house every day, as well as the quality that the water will have in the future, through a model that accurately predicts the quality of the water, in a way that people can more quickly and easily know what they consume at that time and in the near future.

The first step for the development of this system will consist of choosing the sensors that will be used for data collection, these sensors must have at least the possibility of measuring concentrations of total dissolved solids (TDS), pH level, and turbidity, once the sensors have been chosen, they will be placed, one will be placed in each city hall of the CDMX, giving a total of 16 sensors, said sensors will be found within the government sites of each city hall, the installation within these sites will be thanks to an agreement with the CDMX secretariat of science and technology, they will be installed within these sites for two reasons; The first is the security of the

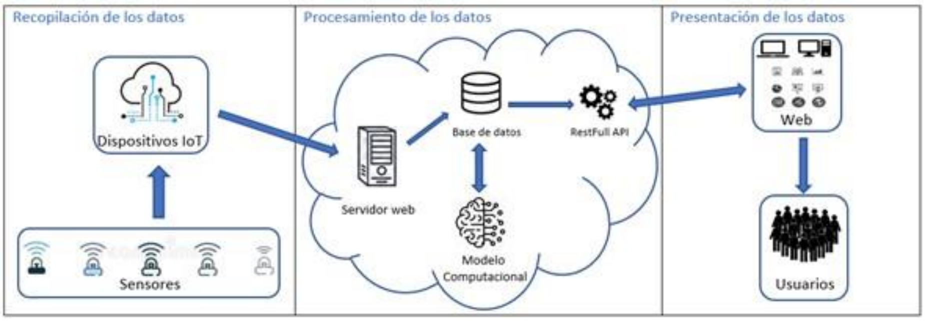


Fig. 1. Data sequence.

sensors and the second is to take advantage of the internet within said sites to send the data captured by each sensor to the database, through IoT devices, said the database will store the information for then be able to process it through a model, which will be helpful for predicting [15], according to the data of each sensor, how the water quality may improve or worsen in the future, in each municipality, then through geographic information systems, the results will be shown, occupying an Outlier type analysis, this analysis will be occupied since as we are talking about a network it is important to know what is the relationship of the water quality in one municipality with respect to the water quality in other municipalities, in order to be able to use the system, the user needs a connection that allows access to the internet, the system will not ask the user for permissions, but will first show a map of the sensor network s and according to the element on which the user clicks, the respective graphs corresponding to that municipality will appear, showing information on both the current quality of water that is being consumed, as well as the quality that is predicted. It is expected that with this system users will have more information about the water service they are receiving, but also that with this it will be easier and faster to detect where the quality of water is declining and could put the quality of life at risk. the users who consume it, figure 1 illustrate the connection between the modules that make up the system.

3.1 Data and Water Quality

Once each one of the properties has been analyzed, take into account the area to be measured, it is time to characterize all this to the CDMX water system, which is the area that the project will cover, for this we must take into account that the CDMX has an area of land of 1485 square km, as well as the number of 8.92 million inhabitants.

The CDMX has a drinking water coverage of 98%: It is important to highlight that 7.5% of the country's population resides in CDMX, as well as that 17% of the country's economic activity passes through here. Considering the previous data and in accordance with the SACMEX.

As reported by SACMEX, every year in CDMX the quality of the water suffers a deterioration since each year, the water level of the aquifer decreases by one meter and as the water is extracted from a greater depth the water undergoes certain changes, in addition to the fact that this also encourages the city to sink 30 cm each year.

In 2019, several actions were undertaken to improve the service of the CDMX water system, one of these actions has been the sectorization of the secondary network in all municipalities to improve the service to users with pressure control schemes, hydraulic balances, and increased efficiency, also considering the CDMX wells that are active, restored or that continue to have problems.

The concept of water quality focuses on the use for a city or for the use of people daily in their homes, for this CONAGUA is based on the following parameters, to say that the quality of the water is, the qualities of the water that CONAGUA has defined are:

- Green: for those that correctly meet all the parameters for water quality.
- Yellow: for those that violate one or more of the following parameters E_COLI, CF, SST and OD%.
- Red: for those that fail to comply with one or more of the following parameters BOD5, COD, TOX and ENTEROC.

The water quality limits, so that it can be considered as potable for human use and consumption, are established in the current Official Mexican Standard NOM-127-SSA1, said standard defines limits such as those exemplified in table 2.

The data was obtained directly from reports by CONAGUA and SACMEX of wells and aquifers that feed the CDMX, this data is public online and include data from 2011 to 2020, it contains information on parameters such as E_COLI, CF, SST, OD%, pH, turbidity, TDS, among others.

This project will focus on parameters such as turbidity, pH, as well as a total of all the solids dissolved in the water, following the CONAGUA color code, if only one parameter is met, the point will be displayed in red, in case of meeting two parameters of yellow and in case of three of green color.

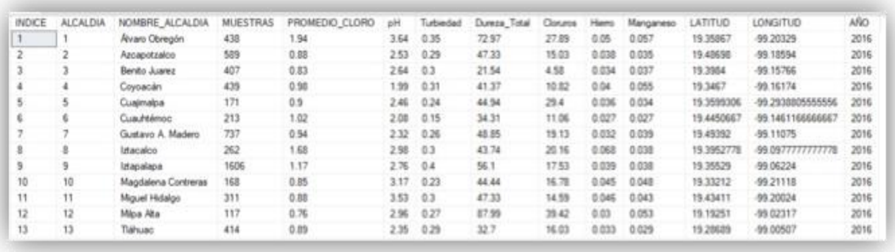
In figure 2, it can be seen the data taken from the source of the government base of SACMEX, which contains data by year of each measure for water quality considered in CDMX, these records are divided by municipalities, the measurements of the sensors and the year; latitude and longitude data were added for space exploration.

From this table, queries were generated, where it was divided only by mayor's office to facilitate analysis of the data as shown in figure 3, the possibility that there was a pattern between the growth or decrease of some measures throughout the period was analyzed as can be seen, in the municipalities that are closest to each other, such as the Gustavo A. Madero mayor's office and the Azcapotzalco mayor's office, despite their proximity, there is no similar pattern of variation of the measures.

Queries were made showing the processes of inserting information to the table of concentrates and the calculation of the error, as can be seen, the insertion to the table of concentrates will occur each time information from all monitoring points arrives, while the calculation of the error will be given once the date of the insertion of data from the monitoring points coincides with the date of the prediction made by means of the forecast model, where the database manager itself by means of a trigger that It

Table 2. Parameters and their limit allowable at Official Mexican Standard NOM-127-SSA1.

Parameters	**Limit Allowable**
Total coliform organisms	2 MPN / 100 ml
Fecal coliform organisms	Not detectable MPN / 100 m
Color	20 true color units on the platinum-cobalt scale.
Odor and Taste	Pleasant (those that are tolerant for most of the consumers)
Turbidity	5 nephelometric turbidity units (NTU) or its equivalent in another method.
Aluminum	0.20
Free residual chloride	0.2-1.5
Copper	2.00
Total hardness as (CaC03)	500
pH	6.5-8.5
Pesticides in microorganisms	0.03
Global alpha radioactivity	0.1
Zinc	5.00
Sodium	200.00
Global beta radioactivity	1.0
Mercury	0.001



INDICE	ALCALDIA	NOMBRE_ALCALDIA	MUESTRAS	PROMEDIO_CLORO	pH	Turbiedad	Dureza_Total	Cloruros	Hierro	Manganeso	LATITUD	LONGITUD	AÑO
1	1	Álvaro Obregón	438	1.94	3.64	0.35	72.97	27.89	0.05	0.057	19.35867	-99.20329	2016
2	2	Azcapotzalco	589	0.88	2.53	0.29	47.33	15.03	0.038	0.035	19.48698	-99.18594	2016
3	3	Benito Juarez	407	0.83	2.64	0.3	21.54	4.58	0.034	0.037	19.3984	-99.15766	2016
4	4	Coyoacán	439	0.98	1.99	0.31	41.37	10.82	0.04	0.055	19.3467	-99.16174	2016
5	5	Cuajimalpa	171	0.9	2.46	0.24	44.94	29.4	0.036	0.034	19.3599306	-99.2930805555556	2016
6	6	Cuauhtémoc	213	1.02	2.08	0.15	34.31	11.06	0.027	0.027	19.4450667	-99.1461166666667	2016
7	7	Gustavo A. Madero	737	0.94	2.32	0.26	48.85	19.13	0.032	0.039	19.49392	-99.11075	2016
8	8	Iztacalco	262	1.68	2.98	0.3	43.74	20.16	0.068	0.038	19.3952778	-99.0977777777778	2016
9	9	Iztapalapa	1606	1.17	2.76	0.4	56.1	17.53	0.039	0.038	19.35529	-99.06224	2016
10	10	Magdalena Contreras	168	0.85	3.17	0.23	44.44	16.78	0.045	0.048	19.33212	-99.21118	2016
11	11	Miguel Hidalgo	311	0.88	3.53	0.3	47.33	14.59	0.046	0.043	19.43411	-99.20024	2016
12	12	Milpa Alta	117	0.76	2.96	0.27	87.99	39.42	0.03	0.053	19.19251	-99.02317	2016
13	13	Tláhuac	414	0.89	2.35	0.29	32.7	16.03	0.033	0.029	19.28689	-99.00507	2016

Fig 2. Basic Exploration of the Concentrate Table.

will calculate the difference between the result of the prediction and that measured by the IoT device [14].

3.2 Data Exploration

In addition, a simulation of the Outlier analysis was carried out, it should be remembered that as already mentioned earlier in this document, the Outlier analysis consists of not only considering the information of the place itself but also of the neighbors and although in the statistics these points are neglected in geospatial analysis is highly relevant [13], since they can indicate possible sources of failures or

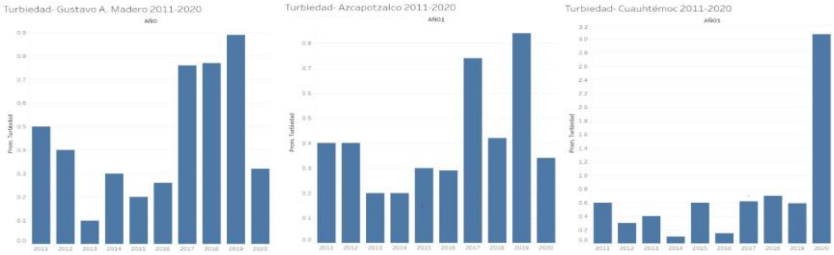
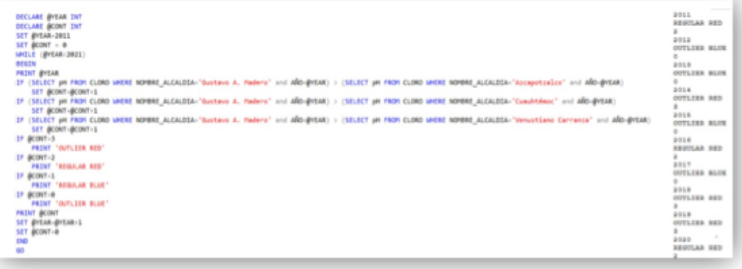


Fig. 3. Graphs on the distribution of turbidity in municipalities (GAM, Azcapotzalco and Cuauhtémoc) from 2011 to 2020.



```
DECLARE @YEAR INT
DECLARE @CONT INT
SET @YEAR=2011
SET @CONT = 0
WHILE (@YEAR<2021)
BEGIN
PRINT @YEAR
IF (SELECT pH FROM CLORO WHERE NOMBRE_ALCALDIA='Gustavo A. Madero' and AÑO=@YEAR) > (SELECT pH FROM CLORO WHERE NOMBRE_ALCALDIA='Azcapotzalco' and AÑO=@YEAR)
    SET @CONT=@CONT+1
IF (SELECT pH FROM CLORO WHERE NOMBRE_ALCALDIA='Gustavo A. Madero' and AÑO=@YEAR) > (SELECT pH FROM CLORO WHERE NOMBRE_ALCALDIA='Cuauhtémoc' and AÑO=@YEAR)
    SET @CONT=@CONT+1
IF (SELECT pH FROM CLORO WHERE NOMBRE_ALCALDIA='Gustavo A. Madero' and AÑO=@YEAR) > (SELECT pH FROM CLORO WHERE NOMBRE_ALCALDIA='Venustiano Carranza' and AÑO=@YEAR)
    SET @CONT=@CONT+1
IF @CONT=3
    PRINT 'OUTLIER RED'
IF @CONT=2
    PRINT 'REGULAR RED'
IF @CONT=1
    PRINT 'REGULAR BLUE'
IF @CONT=0
    PRINT 'OUTLIER BLUE'
PRINT @CONT
SET @YEAR=@YEAR+1
SET @CONT=0
END
GO
```

```
2011
REGULAR RED
2
2012
OUTLIER BLUE
0
2013
OUTLIER BLUE
0
2014
OUTLIER RED
3
2015
OUTLIER BLUE
0
2016
REGULAR RED
2
2017
OUTLIER BLUE
0
2018
OUTLIER RED
3
2019
OUTLIER RED
3
2020
REGULAR RED
2
```

Fig. 4. Example and result of an Outlier analysis simulation.

also points where the sensors are not sending the correct information. For this simulation, several exercises were carried out and an example and its result are shown below in figure 4.

4 Tests and Results

For the test scenario, the homes of people close to the developer were taken as monitoring points. If the minimum concentrations of chlorine in the detections made at the monitoring points want to be guaranteed, consider that the chlorine concentration decays once the water leaves the purification plant, in places far from the plant, residual chlorine may be absent, thus allowing the increase in bacterial levels. During the tests, the parameters of pH level, turbidity and total dissolved solids (TDS) will be measured.

As can be seen in figure 5, the apparatus in the bottom line was used to represent the modems of the houses where the sensors are located, as the communication between the sensors and the cloud system or even between them is not observed. direct, it requires several routers that communicate to the entire network, the main node was also added above since being the coordinator it was easier to exemplify it like this, although this does not mean that it is in a network space superior to that of the other nodes.

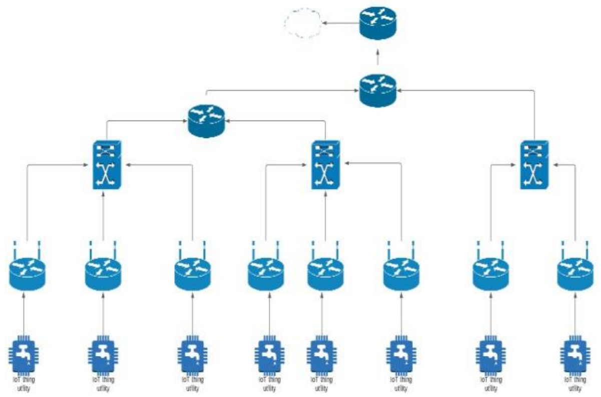


Fig. 5. Exemplification of the CDMX network to which the sensors are connected.

In the second part of the tests, the section of the model and its predictions will be seen, the clustering technique will be used, it is a technique whose purpose is to be able to group the objects by similarity, in groups or sets so that the members of the same group have similar characteristics, the geospatial data will be the variable on which the grouping criterion will be based.

To apply an Outlier type analysis, it is important to define a numerical characteristic, which will help to make the correct division of characterization by zone, for example the level of pH, turbidity, and the number of total dissolved solids can be taken as a variable.

As the Outlier analysis can be seen to be more complete, finally, figures 6 and 7 are shown, some Dashboards generated with the Tableau tool, which serve to further explore the measures.

The web application where the results will be projected is a responsive web dashboard, giving the user options to choose the type of map on which he wants to see the sensor network, what parameter he wants to observe on his screen to measure water quality, view the graphs of the data of each sensor, the options offered by the web application are:

- Location: highlighted points that are the locations of the sensors. Each of these points will have two suboptions, which will be if you want to see the data from the sensors or from the forecast model.

- Charts: the following options bars, pie, histogram, and scatter.

- Report: the application will generate a PDF report of the information, whether it is general or a specific point that the user has selected, adding both the data from the sensors and the data from the predictions made by the model.

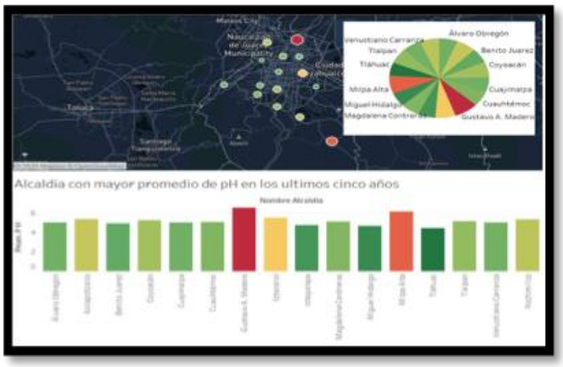


Fig. 6. Dashboard of the pH measurements in the last 5 years in each of the municipalities of the CDMX.

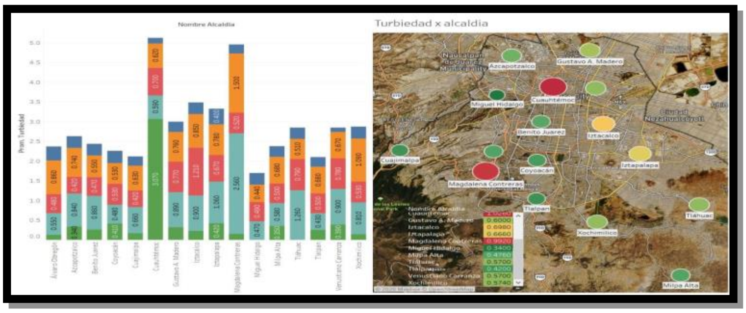


Fig. 7. Dashboard of the turbidity measurements in the last 5 years in each of the municipalities of the CDMX.

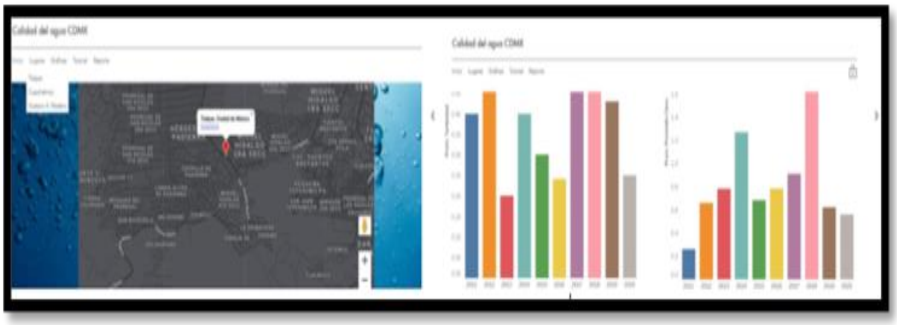


Fig. 8. Visualization of the web application on a computer.

5 Conclusions and Future Work

This project has been interesting from the first stage of collecting information, consulting the reports of both SACMEX and CONAGUA, in this way it was possible

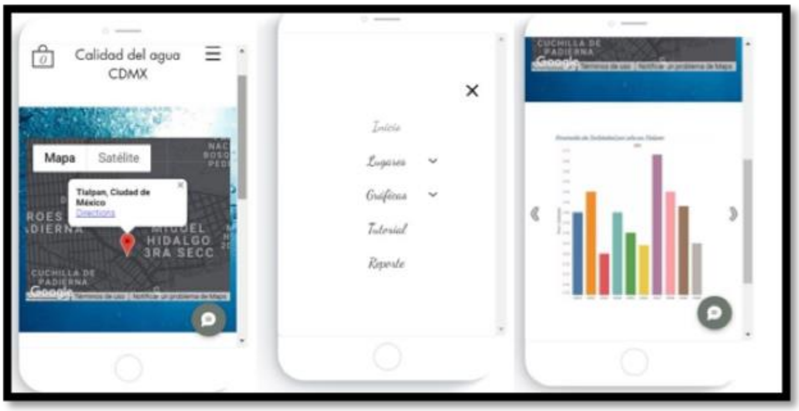


Fig. 9. Visualization of the web application on a cell phone.

to observe that none of the municipalities at the time delegations took continuous measures that allowed to have a good management on the control of water quality, there are also data that show a series of consequences due to this carelessness of control, going from being close to the permissible limit to being almost double, thus exceeding what is adequate for consumption or human use, but without informing the population of these major problems.

On the other hand, it could be observed thanks to the geospatial analysis that over the years no concrete and rapid solution has been given to water quality problems, which has led to a spread of problems around large parts of the network of Mexico City, However, this error has been reduced to only one unit in the pH parameter and for the other parameters the error is even smaller, so it will soon be a model with a fairly high percentage of reliability.

For future work we plan to reach a final agreement with the corresponding authorities to be able to use the system on their behalf and thus serve as a support to improve the quality of life of the inhabitants of Mexico City; in future work the idea is to expand the study area to other large cities in Mexico.

Acknowledgments. Special thanks to M. in C. Roberto Eswart Zagal Flores and Dr. Miguel Felix Mata Rivera for their counseling in the development of the project “Prototype of monitoring network and forecasting model to know water quality in Mexico City”.

References

1. Caldentey, F.: The Age of Data: What Benefits and Risks are there in Big Data. the ‘Big Brother’ of the 21st Century| UNIR (2020)
2. Zakariya-Nafi'Shehab, F., Nor Rohaizah-Jamil, S., Ahmad Zaharin-Aris, T.: Spatial Variation Impact of Landscape Patterns and Land Use on Water Quality Across an

Urbanized Watershed in Bentong, Malaysia. Ecological Indicators, 122, (2021) doi: 10.1016/j.ecolind.2020.107254.

3. WenLiu, F., Bernard A., Engel, S., Weiping Chen, T.: Quantifying the Contributions of Structural Factors on Runoff Water Quality from Green Roofs and Optimizing Assembled Combinations Using Taguchi Method. In: Journal of Hydrology, 593 (2021)
4. Xiao Qu, F., Yushun Chen, S., Han Liu, T.: A Holistic Assessment of Water Quality Condition and Spatiotemporal Patterns in Impounded Lakes Along the Eastern Route of China's South-to-North Water Diversion Project. Water Research, 185 (2020) doi: 10.1016/j.watres.2020.116275.
5. Dameng Yin, F., Le Wang, S., Zhenduo Zhu, T.: Water Quality Related to Conservation Reserve Program (CRP) and Cropland Areas: Evidence from Multi-Temporal Remote Sensing. International Journal of Applied Earth Observation and Geoinformation, 96, (2021) doi: 10.1016/j.jag.2020.102272.
6. Anonymous, F.: Data is the New Gold. Deloitte (2020)
7. Goodchild, F.: First Law of Geography. International Encyclopedia of Human Geography. Elsevier, pp. 179–182 (2009) doi: 10.1016/B978-008044910-4.00438-7.
8. Costa Rica, F.: Towards a Network Geography: A New Paradigm of Space Analysis Alternative to the Regional Approach. Redalyc.org (2020)
9. Agua, F.: Water Quality in Mexico. Government of Mexico (2020)
10. Villanueva, F.: Contaminated Water in Mexico Increased Hospitalization Due to Diarrhea by 12%. La Jornada (2020)
11. CONAGUA, F.: Water Quality Report. Government of Mexico (2020)
12. SACMEX, F.: Water Quality Reports. Government of Mexico City (2020)
13. Blasco Fernández, I: Outlier Detection Methodologies in Spatial, Temporal and Spatio-Temporal Data. Revista Cartográfica, 96, pp. 139–157 (2018)
14. Rose, K., Eldridge, S., Chapin, F.: The Internet of Things a Brief Review. Internet Society, (2015)
15. Sandoval, L.J.: Automatic Learning Algorithms for Analysis and Data Prediction. ITCA FEPADE, 11 (2018)