Design Of A Predictive Algorithm For Early Monitoring Of Drinking Water Networks

Santiago Rubiños, Juan Apesteguia, Juan Grados, Cesar santos, Wilver Auccahuasi

Abstract: A leak is an uncontrolled outflow of water into any of the components of the drinking water distribution system; most often occurs at pipe joints, elbows, conduit and valve breaks. The service of detection of non-visible leaks of emergency leaks as its name indicates is reported by the Sedapal Zones themselves (service and collection center offices strategically located in various districts of the capital). The information reported by the areas is collected by various means: telephone calls, inspections carried out by Sedapal, face-to-face complaints, breakage of tracks due to the passage of heavy vehicles, etc. These Emergency leaks are taken care of by the mobile units according to their importance and / or urgency and it is determined if there is indeed a water leak or not, after which it proceeds, makes its respective report, cataloging and classifying the leak and its report is entered to the system database for subsequent repair and / or filing in case the leak has not been detected. That is why the design of a predictive control algorithm system is proposed that can be anticipated in which the next emergency leaks are distributed, and what type they could be, analyzing the information obtained to date, processing what is known like Big Data.

Index Terms: Predictive Algorithm, Detection of non-visible leaks, Sedapal, Geophone, Trends, Big Data, Software R.

1. INTRODUCTION
S. edapal, as a company providing the drinking water distribution service, is responsible for ensuring the supply of the liquid element in the city of Lima. Users, for their part, are responsible for maintaining their factories, universities, schools, offices, homes, cisterns water, etc. without damage to its pipes, connections and accessories to avoid wasting water unnecessarily and thus avoid the high amounts in billing of receipts at the end of the month and the possible shortages generated by said damages. There are more than 14000 km of drinking water pipes distributed in the great Lima, which due to its length, its age and many other factors; they are vulnerable to damage, call failures which imply substantial losses of drinking water which affect the normal supply by SEDAPAL, these faults can, according to their magnitude of water leakage: small, medium, large and sometimes enormous when the pipes affected are Main or main pipes causing cuts in the water supply in complete districts. Design of a predictive algorithm for early monitoring of drinking water networks in the city of Lima, 2019. It is taken as part of a strategy to improve the early prediction process of where the emergency leaks not visible in the city of Lima. To carry out this process, there is a database with information collected and recorded in a determined period of time from 2014 to 2019, which will be analyzed for the present study.

2 METHODS Y MATERIALS
Drinking water as a natural resource: The problem of the shortage of drinking water for human consumption is a challenge for future generations, the world population is increasing and, in contrast, water resources are scarcer year by year. Only very little water is used for human consumption, since: 90% is seawater and has salt, 2% is ice and is at the poles, and only 1% of all the water on the planet is sweet, being in rivers, lakes and underground mantles. In addition, water as it is found in nature, to be used without risk for human consumption requires treatment, to eliminate particles and organisms that can be harmful to health. And finally it must be distributed through pipes to the houses, so that it can be consumed without any problem or risk. As already indicated, the water lost due to these leaks is called “unbilled water” since these pipes and / or connections for drinking water are located before the water meter and therefore cannot be accounted for and charged to the user by SEDAPAL, all that drinking water from said leaks could be recovered and reinjected into the system, allowing the reduction of the shortage rates.

Figure 1. Cross section of a household connection and the stages where drinking water leaks occur.

EI ECRF
The emergency leak detection process is carried out manually in the field with qualified personnel called ECRF (non-visible water leak control and reduction equipment) who travels to any of the 50 districts of Lima to carry out the inspection of the leak and its subsequent classification with respect to the type of leakage caused, the diameter of the pipe, the flow of drinking water lost by said leak in liters / day, as well as its...
subsequent repair by Sedapal, all in one process. It takes about 3 to 7 days from when the emergency is notified until it is repaired. Then this information is entered into Sedapal’s servers for later consultation and as a history of the service. Currently there is already a database with information of more or less 10 years old which is the basis of this research.

**Big Data**

Although the term “big data” is relatively new, the action of collecting and storing large amounts of information for subsequent analysis has been taking place for many years. The concept gained momentum in the early 2000s.

**Some factors to consider when analyzing data:**

- **Volume.** Organizations collect data from a variety of sources, including business transactions, social media, and information from sensors or transmitted from one machine to another. In the past, storing them would have been a problem, but technologies have lightened the task.

- **Variety.** The data comes in all kinds of formats from structured numerical data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

- **Complexity.** Today’s data comes from multiple sources, making it difficult to link, tie, debug, and transform data between different systems. However, you need to connect and correlate relationships, hierarchies, and multiple data links, or your data can get out of control in a second.

**3 RESULTS**

**Base de Datos Sedapal GIOC**

The original database was extracted from a Sedapal application in SQL environment with infinity of both numerical and text data, the data has 67 columns and 8226 rows of which several are empty and / or duplicates which we will proceed to eliminate by the algorithm and keep only the most important fields. The exported data is in CSV format (separated by commas) which is the most recommended for working and processing in the R software. The first thing was to run with the data and the algorithm a dispersion matrix to evaluate the initial data and find patterns and relationships. For this the definition would be $y = X\beta + \varepsilon$.

![Figure 3. Where $y$ is the vector of the observed values, $X$ the data matrix, $\beta$ is the regression coefficient and $\varepsilon$ would be the error.](image)

**Process**

**Software R**

R currently in version 3.6.3 is free software that allows statistical analysis and is the most widely used in the scientific community. R is an instruction program, and therefore it is not entirely “user friendly” for users who are not used to this type of operation. Currently there is an interface that allows the management of the R program through a menu window, this interface is called RCommander. This program is available on the website: http://www.r-project.org and consists of a central application and libraries of many themes that can be installed as needed. As a first step we must install the R software and load some libraries without which it is not possible to process the information.

![Figure 4. Dispersion Matrix Emergency Leaks 2014-2019](image)
Our predictive model should tell us of the amount of scattered and unrelated data several parameters, such as:

- Districts with the highest incidence of reported emergencies
- Types of Leaks plus incidence of reported emergencies
- Variation of the data based on a specific date
- Relate these variables in a single table
- Relate the variables to each other

After the investigation, it was concluded that it should be a function of 3 variables or fields to consider DATE, TYPE LEAK and DISTRICT 2 of which were text-type data fields and only the date field was numeric, so the algorithm was designed based on those variables finally finding a pattern which we execute through the R software command window. First the data was filtered, we proceeded to convert the data to a matrix and then they were processed in R obtaining results for Districts and Types of Leaks which were plotted for better visualization.

![Figure 5. Dispersion Matrix Results obtained by the algorithm in R.](image)

The data obtained with our predictive algorithm was contrasted with the control data for the year 2020, January to March, obtaining the following result.

![Figure 6. Data processed in January R - March 2020](image)

### 4 CONCLUSIONS

It is possible to develop a Predictive Algorithm in the R Software for early monitoring of drinking water networks in the City of Lima, 2019 using open source software. It was verified that it is possible to analyze Big Data with unrelated data to generate an algorithm that can predict in which Districts new potable water leaks will occur for emergencies and of what types these leaks will be. Use this information by the service provider to project itself and try to optimize the service and reduce its operating costs.

![Figure 7. Contrasting results in Districts.](image)

### 5 RECOMENDACIONES

This study could improve if more data needed for the projections were known, such as the number of water meters installed per year in the city of Lima to estimate the growth curve of the service. The processing of Big Data for estimates and projections is a field that is constantly evolving, so this research must continue to improve and exceed the success rates obtained. PROMOTE THE USE OF THIS TYPE OF BIG DATA TOOLS BY COMPANIES IN ORDER TO MAKE ESTIMATES OF THEIR PRODUCTS AND SERVICES IN ORDER TO OPTIMIZE THEM AND PRODUCE PERCENTAGE SAVINGS.

### REFERENCES

[19] softw a de analisis estadistico r https://cran.r-project.org/bin/windows/base/