Detecting Leakages in Water Distribution Using Data Analytics Approach

A Thesis

Submitted For the Degree of

Master of Management

in the Interdesciplenery Science

by

Priy Ratan Mishra

Department of Management Studies

Indian Institute of Science

BANGALORE – 560 012

June 2017

TO

My Mom and Dad

Acknowledgements

Many thanks to all the persons who made this style file. It will certainly live long! Detailed acknowledgements are available within the style file itself.

Abstract

The leakages of water during the supply is the major non revenue loss for BWSSB in Bangalore city,there are methods such as detect the change of pressure every junction by flow meter,but this method is often cumbersome and very difficult to do at the more granular level.

This project aims to detect the leakages in the pipelines by looking at the consumer monthly consumption volumes and employing the simple and advanced statistical techniques and Data Mining approach.

The method employed were simple residual analysis,exponential moving average and Modified ARIMA in presence of outliers for time series analysis,and for the Data Mining approach the method like Local Outliers Factor and Mahalonobis Distance was employed.

Various algorithms generally picked same common outliers(discordant points)

This analysis revealed that instead of using the flow meter method these techniques can also be employed to identify the discordant points or place where there can be leakage.

Contents

Chapter 1

Introduction

1.1 Bangalore's Great Water Distribution Problem

According to the report of Ministery of Urban Development(MOD) 20 out of 32 cities are water scacre,they provide water around 4.3 hours in a day on an average. According to the MOD the permissible level to the water loss during distribution is around 20%.But indian cities are doing far worse eg Chandigarh waste around 30%,while Bangalore waste around 46% and Indore is the worst off at 60%.

The major cause of losses is

1. Leakage in Pipe Line Burst
2. Water Theft
3. Private Tankers tapping the supply
4. Illegal Connection

In Bangalore the problem is more acute the water is distributed in normal monsoon times one is two days, while during bad monsoon times it is twice in week. The problem is aggravated in the new part of Bangalore where still there is no proper connection and citizens had to depend on the private tanker mafias.

These problem results in the water war sometimes in slums and Govt. has to announce project to carry water from 100-150 km from hinterland to Bangalore.

The dams construction also results in the catastrophic Ecological and Forest areas losses.According to the one estimates 12 dams which are responsible for supply of water in Mumbai has submerged 7000 ha out of which 705 ha where forest land, it also displaces the Aadivasis which are the most neglected part of Indian society.

1.2 Non Revenue Loss

While Developed Countries focuses on the Reuse,Recycle and improving the efficiencies,cities of the developing countries are failing to provide the adequate water to the masses.

According to one study if the loss on distribution at BWSSB reduced to 46% to 20%, the water litres per person(lpcd) which around 75 will increase to 200 lpcd.This will also help in equitable distribution of water among the slums and high rise apartment which is very unequitable now.The Present distribution in slums is 28 lpcd and in high rise apartment is 200 lpcd.

1.3 Why Non Revenue Loss is High

Non Revenue loss is defines as the difference between the water supplied in the system and water is billed at the consume level.

Industry wise NRW is defined as four types:-

1. Pumped and Billed(Where water meters do not exist).
2. Unbilled and Authorize(Unmetered Use).
3. Apparent Losses(Water Theft).
4. Real Losses(Leaks and Burst).

While there are many factors for high NRW(Non Revenue Loss), but the prominent few are as follow:-

1. Lack of Meters
2. No systematic record of Assets
3. Low Cost Recovery in Project
4. High Subsidies

Bangalore has good meter connections and there are District Meter Areas(DMA), there can be case of matching the DMA reading with consumer reading and check for the mismatch,but these approaches are not always reliable and human effort is also more and most importantly the not every area of Bangalore has DMA.

Chapter 2

Data Description

2.1 Data Description

BWSSB divide the Bangalore into twelve Region and they have the consumer billing data of each region. There was total of 10 lakh or 1 Million records was there,where each record represent the one consumer detail along with the month consumption of water.

The total record have 4 year of data from 2006 to 2009 and idea was to use this information for devising techniques to find the discordant regions or subregions.

2.2 Data Preprocessing

The entire data was divided into 12 Region at the beginning based on the unique Region number in the dataset, the idea was to divide the data into region and then subregion wise and then employ various data analytics algorithm to detect the discordant region and subregion in dataset.

The entire following steps was performed during Data Preprocessing:-

1. Divide the Entire Data into 12 Regions
2. Aggregate the consumption of users, month wise across 48 months in each Region.
3. Further Divide the data into 18 subregions based on the address field in the Region data set.
4. Aggregate the consumption of users month wise across 48 months in each Region.

The idea was to go more granular level but because the more granular detail was missing at the address fields. The granular level to street was dropped and analysis was done on the subregion granularity.

Details fo the 12 Sub Region is as follows:-

1. Banshankari Phase 1
2. Banshankari Phase 2
3. Banshankari Phase 3
4. Chamrajpet
5. Sadashivnagar
6. SanjayNagar
7. Geddelahalli
8. IndiraNagar
9. Malleswaram
10. LN_Pura
11. Srirampura
12. Yeswanthpur
13. Binnamanagla
14. Non_Thippasandra
15. Prakash_Nagar
16. RajajiNagar
17. Okalipura
18. R_C_Pura

Chapter 3

Time Series Approach

3.1 Introduction

A time series is a sequence of observations taken sequentially in time. Many sets of data appear as time series: a monthly sequence of the quantity of goods shipped from a factory, a weekly series of the number of road accidents, hourly observations made on the yield of a chemical process, and so on.

Examples of time series abound in such fields as economics, business, engineering, the natural sciences, and the social sciences.

An intrinsic feature of a time series is that, typically, adjacent observations are dependent. Time series analysis is concerned with techniques for the analysis of this dependence.

There are three major area of application of time series and these are:-

1. The forecasting of future values of a time series from current and past values.

2. The determination of the transfer function the determination of a dynamic inputoutput model that can show the effect on the output of a system of any given series of inputs.

3. The use of indicator input variables in transfer function models to represent and assess the effects of unusual intervention events on the behavior of a time series. Application of indicator variable in transfer function models in presence of the unusual event was utilized in one of the Algo. during our analysis. Usually the series is decomposed into Trend seasonality and Random part and we are interested in modeling the Random Part of series ARIMA model

Typical univariate modeling of time series in the absence of outliers includes the following:-

(a) Identification of Order of ARIMA.

(b) Estimation of Parameters.

(c) Residuals Analysis

textwidth:	450pt (approx. 16cm)
textheight:	635pt (approx. 8.8cm)
side margins:	3.5cm
headsep:	40pt

3.2 Simple Algorithm for detecting the Outliers(Inliers)

The basic idea here is to find the robust estimate of trends and Seasonality and subtract them. Then find the outliers,The test for residual Outliers is the same as for the standard box plot ,points greater than 1.5 IQR above or below the upper and lower quartiles are assumed outliers. The number of IQRs above/below these thresholds is returned as an Outlier "score". So the score can be any positive number, and will be zero for non-outliers.

R implementation

In R this can be acheived by the STL function. The seasonal component is found by loess smoothing the seasonal sub-series (the series of all January values, ...); if s.window = "periodic" smoothing is effectively replaced by taking the mean. The seasonal values are removed, and the remainder smoothed to find the trend. The overall level is removed from the seasonal component and added to the trend component. This process is iterated a few times. The remainder component is the residuals from the seasonal plus trend fit. This test perform well when there is very small time dependence among the residuals.

1. New environments "singlespace", "onehalfspace" and "doublespace" are provided, within which single, onehalf and double spacing will apply.
2. Double spacing is turned off within table of contents, footnotes and floats (figures and tables).
3. Proper double spacing happens below tabular environments and in other places where LaTeX uses a strut.
4. Slightly more space is inserted before footnotes.
5. Fixes spacing before and after displayed math.

This is a sample single spaced text created using the `singlespace` environment. Would you prefer this or the default double spaced one? You can have intermediate effect by `\setstretch{1.3}`.

Chapter 4

Fine Tuning

IISc thesis style defines the following macros to fine tune the typesetting according to your taste.

4.1 Renaming bibliography

Use `\bibtitle{References}` to get "References" (or whatever argument you give) as the heading for the Bibliography section. "Bibliography" is the default.

4.2 Index

To produce an index for your thesis, mark index entries in the text by using `\index` command. Then run `makeindex` like `bibtex` after the first pass of LaTeX. This will produce a file `jobname.ind` which will get included automatically in the subsequent passes. If you do not run makeindex, no index gets created. For more details see the LaTeX book [1]and `makeindex` documentation.

4.3 Page Headings

Now you can have page headings in boldface instead of slanted and upper-cased chapter/section headings. It also underlines the headings as in the LaTeX book [1]. See the heading on this page. To use this feature, place `\pagestyle{bfheadings}` in your preamble.

Chapter 5

Conclusions

IISc thesis style provides a simple way to typeset theses in an excellent and pleasant manner. Its use is highly recommended. Additions or modifications are most welcome. Send them (and the bug reports) to

The Almighty
<almighty@admin.iisc.ernet.in>

Appendix A

My Appendix

Bibliography commands as in the LaTeX book [1] may be used. For more details on LaTeX, please see [1].

References

[1] Lamport, L. LaTeX: A Documentation System, Addison-Wesley Publishing Company, 1986.