

Cloud Based Sewerage Monitoring and Predictive Maintenance Using Machine Learning

Siddharth Tripathi¹, Ashwin Budhraja²

¹Student, Dept. of ECE, RV College of Engineering, Bengaluru, KA, India

²Student, Dept. of CS, Rasoni GH Rasoni College of Engineering, Nagpur, MH, India

Abstract - An estimated 65000 million litres of sewage is generated per day in urban areas of India, while the authorities are capable of treating only 28000 million litres per day (MLD). This accounts for just 30% of sewage generated. That means almost 70% of sewage generated in urban areas remains untreated. In this paper an intelligent real-time sewerage monitoring system has been proposed. The system follows a holistic approach in monitoring the network of sewerage tubes through the manholes by functionalizing them with an array of sensors. The system provides real-time data including flow-level, toxic gas concentration, pH level, pressure, turbidity and temperature inside the sewer at any time, to the waste management authorities. For the safety of the device, it is equipped with a loud buzzer. Any fault detected in the flow of the sewerage is communicated through the sensors to the motherboard of the system which is then updated to the cloud database instantly. Machine Learning algorithm is used over the collected data from the sensors which is transmitted to the cloud and then worked upon to make suitable predictions on the status of the system for maintenance beforehand. PCA is used to identify the sensor which captures the maximum variance in the data set. Raspberry Pi (2 B) is the motherboard of the system which has 40 GPIO pins and 4 USB slots that makes interfacing easy. Solar cells with a panel are used to power the device attached to the manhole cover. R tool is used to conduct machine learning algorithms on the data set. The system proves to be highly effective, real-time, scalable, low-power consuming and cost-effective.

Key Words: Intelligent System, Sewerage Monitoring, Machine Learning, Predictive Maintenance, PCA, Raspberry Pi, MQ3, Cloud, R

1. INTRODUCTION

Rapid urbanization and industrialization along with a drastic population growth in the last decade has led to an inestimable amount of waste products in the world. While only some of the developed countries have the facility or awareness to segregate dry and wet household wastes, majority of the countries lack the basic understanding. It is surprising to note that only 27% of the world population used private sanitation facilities connected to sewers, according to a report on sanitation by WHO [1]. Even the ones connected to sewers are not treated properly by the

authorities in most countries. The biggest hurdle in the treatment of sewerage waste is the underground monitoring of the sewers that pass through the middle of the busiest roads.

The sewage is not safely disposed if the sewage disposal system is not properly maintained i.e. all the faulty parts have to be monitored and repaired as soon as possible. The faulty parts which are ignored for a long duration accumulate the sewage and other suspended particles in the flow. Over the time, it creates the breeding grounds for many different bacteria, viruses and worms. Some of the most common diseases caused by the unmonitored and unrepaired sewerage systems are: Diarrhea, Trachoma, Gastroenteritis, Hepatitis A, Giardiasis and other worm infections.[2]

The sewerage monitoring and maintenance in most countries is done manually, either with hands or using machines. This paper proposes a system which follows a holistic approach towards monitoring and also maintenance. The proposed system is divided into two parts or functions:

- Sewerage Monitoring through a network of sensors connected to Raspberry Pi, attached to manhole covers.
- Predictive Maintenance through Recursive PCA and Decision tree, at the cloud server control room.

A wireless sensor network model based on Zigbee is proposed by the authors in [3]. This model employs RF antennas and relays to transmit sensor data to a centrally located control room. Many disadvantages in using antennas with low range are also discussed which lowers the efficiency of the model significantly. Gao Hongyan in his paper [4] uses a multi-sensor approach which fuses the data of all the similar sensors and then principal components are deduced. This approach lacks the sensors variability as all the sensors used are of the same kind. Another approach followed by Tianqi Yu involves the aggregation of different sensor data and analysing it using Recursive PCA to find out the outliers [5]. The paper is not extended to prediction algorithms for predictive maintenance. IoT involves the simultaneous working and data acquisition by various sensor networks over the time. The trends in the data are critically analysed to develop a stringent pattern of anomalies. Such an approach is used by Weihua Li [6], which

focuses on a trend-adaptive PCA analysis of the sensor data and detects any faulty data.

2. METHODOLOGY

The block diagram of the proposed system is shown in Fig 2.1.

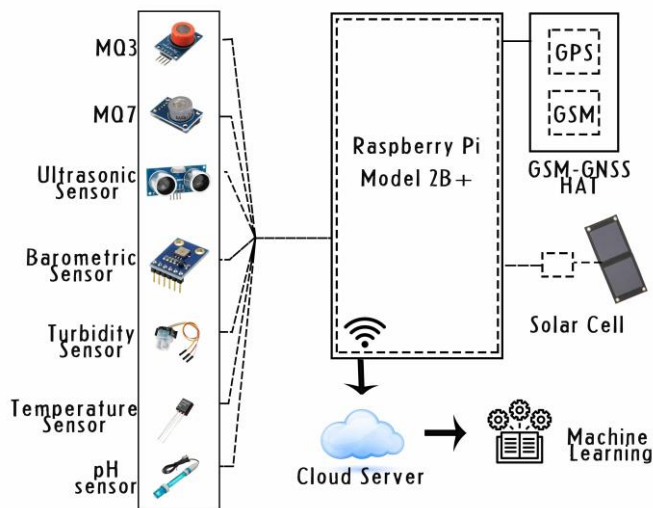


Fig.2.1. Block diagram of the proposed system.

The proposed system follows an automated approach towards monitoring the sewers through the manhole covers. The manual intervention is limited to cleaning and maintenance works only. The system employs an array of sensors including the gas, proximity, temperature, pressure, turbidity and pH sensors. These sensors are interfaced with a motherboard which is responsible for providing real-time data to a centrally located cloud server. Any discrepancies noticed are reported instantly to the cloud with the unique ID of the manhole line, using the GSM/GNSS module.

The other part of the system deals with the data set received from all the lines or networks of manholes. The data set is stored at the cloud server and is then analysed for forming predictions over maintenance using machine learning. The Principal Component Analysis (PCA) is carried out on the data set to lower the dimensionality such that the sensors with most captured variance are identified from the array of sensors. The variance matrix shows the sensors which have shown signs of abnormality over the time. With the help of decision trees applied on the identified principal components, the status of the system is predicted against the test data set. The predictions help in the maintenance of the sewerage beforehand.

2.1 Sewerage Monitoring

The sewerage monitoring is the first part of the proposed system. It employs an array of sensors of different kinds namely: gas sensors (MQ3, MQ7), temperature sensor (LM35), ultrasonic Sensor (HC-SR04), barometric sensor (BMP180), turbidity sensor (SEN0189) and pH sensor

(SEN0161). This network of sensors is interfaced with the motherboard i.e. Raspberry Pi (2B+) with the help of an ADC (Analog-to-Digital Converter) i.e. MCP3008. The real-time data acquired through the sensors is transmitted to the cloud server with the help of a GPS/GNSS HAT (Hardware-Attached-On-Top) for Raspberry Pi. The location of the specific manhole cover with its unique ID (UID) is also transmitted against the corresponding data. This enables the authorities to monitor the sewer lines through a centrally located control room with the cloud server. Furthermore, the ultrasonic sensor works in two ways in the system. First, it monitors the water level inside the sewer and, second, it is programmed to trigger a buzzer if the manhole cover is opened without permission. The system is powered by a PiJuice HAT which consists of a 3.7V Li-ion battery and a solar panel.

2.2 Predictive Maintenance

The second part of the proposed system comprises of data analysis procedures at the control room where the cloud server is located. Once the sensor data is collected at the server, it is divided into training and testing sets. The principal component analysis is done recursively on the data such that the most significant sensor or the sensor with the most captured variance is identified. With the help of prediction algorithms like the decision tree, the testing dataset is predicted with the status of the system. The authorities thus can conduct maintenance exercises beforehand.

2.3 Recursive PCA

Since the sensors acquire data on a real-time basis, applying PCA on just one set of training data will prove to be useless and give many false alarms. So a different approach is adopted in this paper which is based on the model developed by [6] that recursively updates the correlation matrix. This adaptive process is called *Recursive PCA*. The newly acquired values get updated instantly in the matrix and the predictions are thus formed recursively too. In the conventional PCA model, the data is analyzed block wise. If one block of data is used to generate an initial PCA, we are required to update it with the new block when it is available. Let $X_1^0 \in \mathcal{R}^{n_1 \times m}$ be the initial block of raw data. We calculate the mean of each of the column as:

$$b_1 = \frac{1}{n_1} (X_1^0)^T 1_{n_1} \tag{1}$$

where $1_{n_1} = [1, 1, \dots, 1]^T \in \mathcal{R}^{n_1}$. The data that is scaled to unit variance and zero mean is calculated as:

$$X_1 = (X_1^0 - 1_{n_1} b_1^T) \Sigma_1^{-1} \tag{2}$$

where,

$$\Sigma_1 = \text{diag}(\sigma_{1.1}, \dots, \sigma_{1.m})$$

Where the i th element is the standard deviation of the i th sensor. The correlation matrix calculated as:

$$\mathbf{R}_1 = \frac{1}{n_1 - 1} \mathbf{X}_1^T \mathbf{X}_1 \quad (3)$$

When the new data block is available, it will update the data matrix and also recursively calculate correlation matrix. Let's assume that \mathbf{b}_k , \mathbf{X}_k and \mathbf{R}_k are already calculated from the k th block of data. Now we have to calculate \mathbf{b}_{k+1} , \mathbf{X}_{k+1} and \mathbf{R}_{k+1} when we have the next block of data available i.e

$$\mathbf{X}_{n_{k+1}}^n \in \mathcal{R}^{n_{k+1} \times m}$$

Denoted as,

$$\mathbf{X}_{k+1}^0 = \begin{bmatrix} \mathbf{X}_k^0 \\ \mathbf{X}_{n_{k+1}}^0 \end{bmatrix}$$

Corresponding to all $k+1$ blocks, the mean vector \mathbf{b}_{k+1} has a relation with \mathbf{b}_k given by:

$$\left(\sum_{i=1}^{k+1} n_i \right) \mathbf{b}_{k+1} = \left(\sum_{i=1}^k n_i \right) \mathbf{b}_k + \left(\mathbf{X}_{n_{k+1}}^0 \right)^T \mathbf{1}_{n_{k+1}} \quad (4)$$

Taking $N_k = \sum_{i=1}^k n_i$

We have,

$$\mathbf{b}_{k+1} = \frac{N_k}{N_{k+1}} \mathbf{b}_k + \frac{1}{N_{k+1}} \left(\mathbf{X}_{n_{k+1}}^0 \right)^T \mathbf{1}_{n_{k+1}} \quad (5)$$

Calculating recursive for \mathbf{X}_{k+1} ,

$$\begin{aligned} \mathbf{X}_{k+1} &= [\mathbf{X}_{k+1}^0 - \mathbf{1}_{k+1} \mathbf{b}_{k+1}^T] \Sigma_{k+1}^{-1} \\ &= \left[\begin{bmatrix} \mathbf{X}_k^0 \\ \mathbf{X}_{n_{k+1}}^0 \end{bmatrix} - \mathbf{1}_{k+1} \mathbf{b}_{k+1}^T \right] \Sigma_{k+1}^{-1} \\ &= \left[\begin{array}{c} \mathbf{X}_k^0 - \mathbf{1}_k \Delta \mathbf{b}_{k+1}^T - \mathbf{1}_k \mathbf{b}_k^T \\ \mathbf{X}_{n_{k+1}}^0 - \mathbf{1}_{n_{k+1}} \mathbf{b}_{k+1}^T \end{array} \right] \Sigma_{k+1}^{-1} \\ &= \left[\begin{array}{c} \mathbf{X}_k \Sigma_k \Sigma_{k+1}^{-1} - \mathbf{1}_k \Delta \mathbf{b}_{k+1}^T \Sigma_{k+1}^{-1} \\ \mathbf{X}_{n_{k+1}} \end{array} \right] \end{aligned} \quad (6)$$

where,

$$\begin{aligned} \mathbf{X}_k &= (\mathbf{X}_k^0 - \mathbf{1}_k \mathbf{b}_k^T) \Sigma_k^{-1} \\ \mathbf{X}_{n_{k+1}} &= (\mathbf{X}_{n_{k+1}}^0 - \mathbf{1}_{n_{k+1}} \mathbf{b}_{k+1}^T) \Sigma_{k+1}^{-1} \\ \Sigma_j &= \text{diag}(\sigma_{j,1}, \dots, \sigma_{j,m}), j = k, k+1 \\ \Delta \mathbf{b}_{k+1} &= \mathbf{b}_{k+1} - \mathbf{b}_k \end{aligned}$$

Note that $\mathbf{1}_k = [1, \dots, 1]^T \in \mathcal{R}^{N_k}$.

The recursive computation of standard deviation under Appendix A, is given by the following relation:

$$\begin{aligned} (N_{k+1} - 1) \sigma_{k+1,i}^2 &= (N_k - 1) \sigma_{k,i}^2 + N_k \Delta b_{k+1}^2(i) \\ &+ \left\| \mathbf{X}_{n_{k+1}}^0(:, i) - \mathbf{1}_{n_{k+1}} b_{k+1}(i) \right\|^2 \end{aligned} \quad (7)$$

Where $\mathbf{X}_{n_{k+1}}^0(:, i)$ is the i th column of the corresponding matrix; $b_{k+1}(i)$ and $\Delta b_{k+1}(i)$ are taken as the i th elements of the vectors

Similarly the recursive computation of the correlation matrix, derived under Appendix B, has the relation:

$$\begin{aligned} \mathbf{R}_{k+1} &= \frac{1}{N_{k+1} - 1} \mathbf{X}_{k+1}^T \mathbf{X}_{k+1} \\ &- \frac{N_k - 1}{N_{k+1} - 1} \Sigma_{k+1}^{-1} \Sigma_k \mathbf{R}_k \Sigma_k \Sigma_{k+1}^{-1} \\ &+ \frac{N_k}{N_{k+1} - 1} \Sigma_{k+1}^{-1} \Delta \mathbf{b}_{k+1} \Delta \mathbf{b}_{k+1}^T \Sigma_{k+1}^{-1} \\ &+ \frac{1}{N_{k+1} - 1} \mathbf{X}_{n_{k+1}}^T \mathbf{X}_{n_{k+1}} \end{aligned} \quad (8)$$

With respect to the recursive relation above, not that:

- The effect of mean changes $\Delta b_{k+1}(i)$ on the correlation matrix is only rank-one modification.
- If the model has to be updated after every new sample is available, the recursive relationship reduces to the form:

$$\begin{aligned} \mathbf{R}_{k+1} &= \frac{k-1}{k} \Sigma_{k+1}^{-1} \Sigma_k \mathbf{R}_k \Sigma_k \Sigma_{k+1}^{-1} \\ &+ \Sigma_{k+1}^{-1} \Delta \mathbf{b}_{k+1} \Delta \mathbf{b}_{k+1}^T \Sigma_{k+1}^{-1} + \frac{1}{k} \mathbf{x}_{k+1} \mathbf{x}_{k+1}^T \end{aligned} \quad (9)$$

which can be called two rank-one modifications. We have the algorithms to compute the eigenvectors of rank-one modifications [6].

- As old data samples do not represent the current process, they are exponentially ignored. Thus, recursive computations for Eqs. (5), (7) and (8) with a forgetting factor are as follows:

$$\mathbf{b}_{k+1} = \mu \mathbf{b}_k + (1 - \mu) \frac{1}{n_{k+1}} \left(\mathbf{X}_{n_{k+1}}^0 \right)^T \mathbf{1}_{n_{k+1}} \quad (10)$$

$$\begin{aligned} \sigma_{k+1,i}^2 &= \mu (\sigma_{k,i}^2 + \Delta b_{k+1}^2(i)) + (1 - \mu) \frac{1}{n_{k+1}} \\ &\times \left\| \mathbf{X}_{n_{k+1}}^0(:, i) - \mathbf{1}_{n_{k+1}} b_{k+1}(i) \right\|^2 \end{aligned} \quad (11)$$

and

$$\begin{aligned} \mathbf{R}_{k+1} &= \mu \Sigma_{k+1}^{-1} (\Sigma_k \mathbf{R}_k \Sigma_k + \Delta \mathbf{b}_{k+1} \Delta \mathbf{b}_{k+1}^T) \Sigma_{k+1}^{-1} \\ &+ (1 - \mu) \frac{1}{n_{k+1}} \mathbf{X}_{n_{k+1}}^T \mathbf{X}_{n_{k+1}} \end{aligned} \quad (12)$$

for $N_k \gg 1$.

Forgetting Factor is denoted by μ ,

where $0 < \mu \leq \frac{N_k}{N_{k+1}} < 1$

i.e. smaller μ tends to forget the data more quickly, while,

$$\mu = \frac{N_k}{N_{k+1}} \quad (13)$$

is the case of no forgetting. The forgetting factor acts like a tuning parameter which determines how fast the process gets changed. It works in the similar approach as a moving window. Once the principal components are computed, the prediction algorithm namely decision tree helps in computing the status of the system for the test dataset. The computations were done using R tool on a system running Windows 7.

3. COMPONENTS USED

3.1 Raspberry Pi 2 B+

The Raspberry Pi is a powerful, small and lightweight single-board ARM-based computer. Apart from the 40 GPIO pins, it provides USB, Ethernet, HDMI and WiFi connectivity as well. For supplying analog inputs, a simple ADC converter (MCP3008) is used.

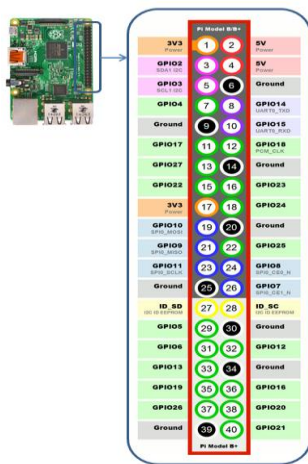


Fig3.1. GPIO pin diagram of Raspberry Pi 2B+ [7]

3.2 Sensor Network Array

The sensors network array is the data acquiring network which comprises of total seven different sensors. Each of the sensor has a specific parameter to monitor. The different sensors employed in this network are as follows:

1. Gas sensor: MQ3 and MQ7: For realtime monitoring of gases inside the closed sewer system.
2. Temperature Sensor: LM35: For monitoring the temperature inside the closed sewer system.
3. Ultrasonic Sensor: HRC-SR04: For monitoring the water level inside and device theft detection in case of manhole has been opened by any unauthorized person.
4. Barometric Sensor: BMP180: For monitoring the air pressure inside the closed sewer system.
5. Turbidity Sensor: SEN0189: For monitoring the turbidity of the sewer water inside the closed sewer system.
6. pH Sensor: SEN0161: For monitoring the pH level of the sewer water.

3.3 GSM/GNSS HAT

The proposed system uses a SMARTELEX GSM/GNSS Hardware-Attached-On-Top for Raspberry Pi. This HAT enables the motherboard to transmit data over the internet to the cloud server. The integrated GNSS module helps send the location coordinates of a manhole too.

3.4 Solar Cell with Panel

The complete system is powered using a solar panel (12V-4W) with a built-in solar charging regulator. The panel is connected to a 12V-1.3Ah Lead Acid Rechargeable Battery. The battery is connected further to an UBEC DC/DC Converter which provides 5V input power supply to the Raspberry Pi.

3.5 Thingspeak Cloud Server (SaaS)

The data acquired by the edge device is transmitted continuously to the cloud server. The Thingspeak server [8] stores the data and provides many general insights. The data is then sent to the R tool for further processing and application of machine learning algorithms.

3.6 R Tool

The R programming tool is a famous statistical computing software. The datasets stored in the cloud server are continuously updated. These dynamic training and test data sets are fed into the R tool for applying R-PCA and Decision tree algorithms.

4. RESULTS

The prototype of the proposed system was made using breadboards. Both the parts of the system were tested for their respective functionalities. It was observed that sewerage monitoring edge device worked efficiently by transmitting data from different sensor nodes simultaneously without any delay. Sample dataset of 400 different values was divided into 300 and 200 samples respectively for training and test dataset.

A. Sewerage Monitoring (Edge Device)

The edge device comprising of sensor array interfaced with Raspberry Pi, transmits real-time data to the cloud server at the control room. The general insights from the cloud monitor is shown in Fig1.1.

Time	Line	Serial	MQ3	MQ7	Temp	Ultrasonic	Pressure	Turbidity	pH	Battery_Sl	Status	
2	12-11-2011	20	59	70	28	33.73	96	33	121	7.2	HIGH	1
3	12-11-2011	20	60	71	26	33.68	97	23	118	7.3	HIGH	1
4	12-11-2011	20	61	71	27	33.65	99	23	118	7.3	HIGH	1
5	12-11-2011	20	62	72	27	33.6	99	7	118	7.3	LOW	0
6	12-11-2011	20	63	74	27	33.6	99	7	117	7.3	LOW	0
7	12-11-2011	20	64	74	28	33.57	98	8	114	7.3	LOW	0
8	12-11-2011	20	65	74	28	33.57	99	8	116	7.3	LOW	0
9	12-11-2011	20	66	75	30	33.51	123	7	117	7.3	LOW	0
10	12-11-2011	20	67	75	30	33.45	124	22	142	7.3	HIGH	1
11	12-11-2011	20	68	76	30	33.51	123	22	126	7.3	HIGH	1
12	12-11-2011	20	69	76	25	33.58	125	23	130	7.3	HIGH	1
13	12-11-2011	20	70	82	25	33.56	125	20	136	7	HIGH	1
14	12-11-2011	20	71	82	26	33.43	125	20	119	7	HIGH	1
15	12-11-2011	20	72	81	26	33.4	126	21	135	7	HIGH	1
16	12-11-2011	20	73	80	26	33.43	126	23	146	7	HIGH	1
17	12-11-2011	20	74	78	26	33.45	128	23	124	7	HIGH	1
18	12-11-2011	20	75	78	28	33.4	128	23	126	7	HIGH	1
19	12-11-2011	20	76	77	26	33.39	128	24	133	7	HIGH	1
20	12-11-2011	20	77	77	2	33.4	0	25	129	7	LOW	0
21	12-11-2011	20	78	77	2	33.37	0	23	124	7	LOW	0
22	12-11-2011	20	79	76	2	33.3	0	23	121	7	LOW	0
23	12-11-2011	20	80	73	25	33.23	125	22	120	7	HIGH	0
24	12-11-2011	20	81	73	26	33.18	125	22	107	0	LOW	0
25	12-11-2011	20	82	73	22	33.19	127	23	114	0	LOW	0

Fig4.1. Screenshot of the data transmitted to cloud server.

The cloud monitor shows general insights of the data in the form of a line chart. Real-time advancing line chart can be analyzed for each of the sensors. The chart for MQ3 and LM35 sensors are shown in Fig 4.2.

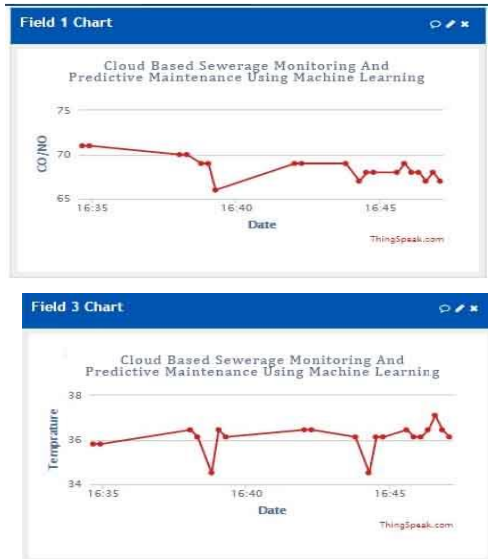


Fig.4.2 Line charts showing values of MQ3 and LM35

B. Predictive Maintenance (R-PCA)

The level of correlation that is of importance is kept at 0.38. Therefore the correlation above 0.38 is deemed as important. The first principal component is computed to be strongly correlated to two of the original variables i.e. temperature and pH level sensors. The second principal component corresponds to the MQ7 and Ultrasonic sensors.

The variance of all the components is computed to identify the principal component with maximum variance i.e. to identify the most significant sensors. The results show that the first principal component explains almost 50 percent of the variance. The second principal component explains 23 percent of the variance and so on. The resultant principal components are plotted as shown in Fig 4.3.

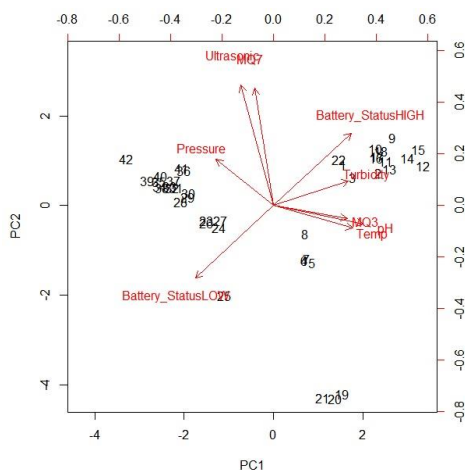


Fig 4.3. Plot of the resultant principal components.

A scree plot showing the proportion of variance of principal components is shown in Fig. 4.4.

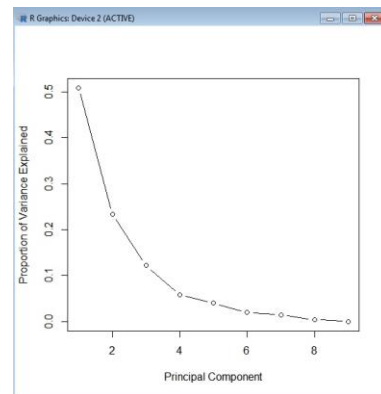


Fig 4.4. Scree plot of the variability of data

The plot shows that almost 3 components explain around 90 percent variance of the data. A confirmation check can be done using cumulative variance computation. Therefore, the three most significant sensors are identified which have the maximum correlation and explains the maximum variance. Similarly, the principal components are calculated for the testing set. With the help of a decision tree algorithm, the status predictions for the test dataset are computed.

5. CONCLUSION

In this paper, implementation of an Intelligent Monitoring System is shown and a detailed explanation has been given. The system prototype introduced in this paper has been tested for its feasibility and can prove to be revolutionary, if adopted in our current organization of urban planning. The system follows a holistic approach and uses cloud-computing for real-time monitoring as well as machine learning for predictive maintenance. Although, the prediction accuracy is less, the most significant sensors are identified. It is very user-friendly and an easy-to-adapt system.

6. ACKNOWLEDGEMENT

The author would like to thank the workers at Shock Labs for providing necessary modules to build the prototype of the system.

REFERENCES

- [1] <https://www.who.int/news-room/fact-sheets/detail/sanitation>
- [2] Mara D, Lane I, Scott B, Trouba D. Sanitation and health. *PLoS Med.* 2010;7(11):e1000363. Published 2010 Nov 16. doi:10.1371/journal.pmed.1000363
- [3] C. H. See *et al.*, "A Zigbee based wireless sensor network for sewerage monitoring," *2009 Asia Pacific Microwave Conference*, Singapore, 2009, pp. 731-734. doi: 10.1109/APMC.2009.5384245
- [4] Hongyan, Gao. (2009). A Simple Multi-sensor Data Fusion Algorithm Based on Principal Component Analysis. 423 - 426. 10.1109/CCCM.2009.5267459.
- [5] Yu, Tianqi & Wang, Xianbin & Shami, Abdallah. (2016). A novel R-PCA based multivariate fault-tolerant data

aggregation algorithm in WSNs. 1-5.
10.1109/ICC.2016.7511419.

- [6] Weihua Li, H.Henry Yue, Sergio Valle-Cervantes, S.Joe Qin, Recursive PCA for adaptive process monitoring, Journal of Process Control, Volume 10, Issue 5, 2000, Pages 471-486, ISSN 0959-1524, [https://doi.org/10.1016/S0959-1524\(00\)00022-6](https://doi.org/10.1016/S0959-1524(00)00022-6).
- [7] <https://www.raspberrypispy.co.uk/2014/07/raspberry-pi-b-gpio-header-details-and-pinout>
- [8] <https://www.mathworks.com/help/thingspeak/>