# Anomaly Detection for Air Quality Monitoring

Shreyas Ramakrishna
Department of EECS
Vanderbilt University
Nashville, Tennessee 37212
Shreyas.ramakrishna@vanderbilt.edu

Sanchita Basak
Department of EECS
Vanderbilt University
Nashville, Tennessee 37212
sanchita.basak@vanderbilt.edu

***Abstract*—In this paper an anomaly detection technique for air quality monitoring dataset is presented. K-Means clustering algorithm is applied to classify an entire region of Minneapolis into different zones and determine the hazardous chemicals present in a particular zone. Along with this we also apply CUSUM (Cumulative Sum Control Chart) detection algorithm to find faulty sensor node in a particular area responsible for the anomalous chemical readings. Ultimately, we aim at identifying the streets where faulty senor nodes are placed producing anomalous readings of any particular chemical in the same area.**

***Keywords—Anomaly Detection; K-Means: Clustring; CUSUM Algorithm; Air Quality Monitoring***

## I. INTRODUCTION

In this work we apply various anomaly detection techniques to determine anomaly in a given dataset. For the purpose of this work we consider a dataset on "Air Quality Monitoring-Minneapolis", provided by the Minneapolis Health Department (MHD). The dataset consists of 61 Volatile Organic Compounds readings which are commonly present in the atmosphere. There are particularly 4790 readings measured across the entire city over a period of 72 hours.

Anomaly refers to the deviation in data that do not match with its normal behavior. They are also commonly referred to as outliers, peculiarities, aberrations, etc. Finding out these deviations of the data are important as they provide significant and critical information [1]. Hence, anomaly detection is widely used in a suite of different application areas such as insurance, health care, credit card frauds, intrusion detection in cyber security, etc.

However, simple and straightforward it may look, there exists a number of challenges which make anomaly detection difficult to apply. One persistent problem is to create a boundary between the normal and anomalous data. For this we require a benchmark to segregate the data between the normal and the anomalous one. The other problem lies in creating the notion of anomaly as it may differ for different application domains [1]. A data may stand out as an anomalous one in one region but may not be in another, so this would make it all hard to define an exact anomaly.

In-order to understand the advantages and the challenges that anomaly detection techniques have to offer, we try to apply a few well-known anomaly detection techniques to our air quality monitoring dataset. Our aim here is to evaluate techniques and also determine the anomalous readings of the Volatile Organic Compounds (VOC) depending on the Health Benchmark Value provided.

The work here is two folded. First, we try to cluster the entire city of Minneapolis into eleven different zones with the longitude, latitude and zip code readings provided in the dataset, and then we try to segregate the 4790 chemical readings according to these zones and evaluate the chemical which is present in higher quantity above the health benchmark provided for each clustered zone. Thus, we find out the hazardous chemicals for every zone in Minneapolis. For this K-Means algorithm has been used.

In the second part, we apply CUSUM detection technique to determine the faulty sensors responsible for the anomalous chemical readings. Through this we try to find out the faulty sensor in every zone.

The remaining section of this paper is organized as follows. In Section II, we present the background for anomaly detection. In Section III, a detailed description about the methodologies used is given. Section IV, discusses about the implementation and simulation results. The concluding remarks are presented in Section V along with the future directions of work.

## II. RELATED WORK

Anomaly detection as a survey has been discussed in [1], where the authors have defined anomaly as those specific patterns in the data that do not conform to a well-defined notion of normal behavior. The work has also dealt in deeply with the topics of challenges and classification problems in anomaly detection. They classify the anomaly detection techniques into different categories, namely, classification based, clustering based, nearest neighborhood based, Statistical, Information theoretic and spectral. With relevant examples they classify and explain the various detection techniques which can be applied to different classifications. This provides an excellent theoretical understanding of the various concepts involved in anomaly detection, and discusses about its implementations in real life complex anomaly detection problems. The first part of our work to cluster the entire region and determine anomalous readings has been based on the understanding of clustering and classification techniques defined in this work

In [2], the authors work on determining the faulty sensors deployed on streets for traffic monitoring. The application being monitored here is route planning for which the sensor values are absolutely critical. The complete work revolves around deploying detection algorithm to determine the faulty sensors relaying erroneous data. The detection algorithm being used here is called CUSUM, which is a statistical quality control algorithm for finding out the faulty node. Along with this the work also

introduces an algorithm for optimal selection of detection threshold which can be useful to balance between false positives (FP) and false negatives (FN). Our, second part of anomalous sensor detection is based on the CUSUM detection algorithm introduced in this work. We, implement this detection algorithm to determine the faulty sensor nodes responsible for erroneous readings.
In the next section we will discuss the implementation of the different detection techniques.

## III. Methodologies Used

### *A) K-Means Algorithm:*

K-Means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. It is a popular cluster analysis tool used in data mining. It aims to partition *n* observations in *k* clusters defined by their centroids. The number of clusters *k* is chosen before the algorithm starts. Each observation belongs to a cluster with having its value nearest to the cluster's mean. As a result the data gets partitioned into Voronoi cells.

K-Means clustering is computationally difficult and is an NP-Hard problem.

Let X= $\{x_1, x_2, x_3, \dots \dots \dots x_n\}$ be the set of data points and V= $\{v_1, v_2, v_3, \dots \dots \dots v_k\}$ be the set of cluster centers. The *k* cluster centers are randomly selected initially. Next, we aim to assign each data point belonging to a dataset to a cluster. This is done by calculating the distance between each data point and all the cluster centers. So, the main objective of the algorithm is to minimize an objective function known as the squared error function given by:

$$J(V)=\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i - v_j)^2 \quad (1)$$

Where, $(x_i - v_j)$ is the Euclidean distance between $x_i$ and $v_j$, *n* is the total number of data points in the dataset and *k* is the number of clusters.

Each data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers. Then the new cluster center is recalculated using:

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i \quad (2)$$

Where $c_i$ represents the number of data points in $i^{th}$ cluster. After that, the distances between each data point and newly calculated cluster centers are recalculated and reassignment of data points are done based on the smallest distance between a data point and a cluster center. This process is repeated until no data gets reassigned for several number of iterations or the predefined limit of maximum number of iterations is reached.

### *B) CUSUM Algorithm:*

To identify small incremental changes in the mean of a process, CUSUM algorithm can be used to produce CUSUSM control chart (Cumulative Sum Control Chart). It is a sequential analysis technique developed by E.S. Page [4] and specially used for change or anomaly detection.

If a sequence is given by $x_1$, $x_2$, $x_3$……$x_n$, and its average or mean is estimated as $m_x$, and the standard deviation comes out to be $\sigma_x$, then the upper and lower cumulative process sums are defined as:

- Upper Cumulative Sum:

$$U_i=\begin{cases}0, & i=1\\ \max(0, U_{i-1}+x_i-m_x-0.5*n*\sigma_x), & i>1\end{cases} \quad (3)$$

- Lower Cumulative Sum:

$$L_i=\begin{cases}0, & i=1\\ \min(0, L_{i-1}+x_i-m_x+0.5*n*\sigma_x), & i>1\end{cases} \quad (4)$$

[iupper,ilower]=cusum(x,climit,mshift,tmean,tdev)(6)

The variable *climit* in equation (6) specifies the number of standard deviations, the upper and lower cumulative sums are allowed to drift from the mean. The minimum detectable mean shift *mshift*, the target mean *tmean*, the target standard deviation *tdev* can also be provided as arguments in the *cusum* function.

The variable *n*, in equation (3) and (4), denotes the number of standard deviations from the target mean, when a shift is discernable or noticeable and is denoted by the argument *mshift* as discussed in equation (6). The CUSUM criterion is violated in a process at a sample $x_i$, if the condition $U_i > c\sigma_x$ or $L_i < -c\sigma_x$ is satisfied, where *c* is the control limit denoted by the argument *climit* as shown in equation (6). While implementing it in MATLAB, the function returns the first violation detected, by default. To verify all the violations the flag 'all' must be specified.

## IV. Implementation And Results

The selected air quality monitoring dataset has about 61 VOCs and 4790 chemical readings which includes the longitude and latitude, Object ID that uniquely identifies a row, date of sample collection, Sample_ID: a unique identifier of the sample, Parameter (VOCs), Results: The amount of VOC present in the sample, Units: µg/$m^3$, CAS registry number, Health Risk Value (HRV), Units: µg /$m^3$, HRV types (HRV, REL-Recommended exposure limits , PEL-Permissible exposure limits), Name of location, volunteer and business, Description- the location where canister(sensor) was placed, Address of the sample collection area, State and Zip code.

From the dataset the longitude, latitude, zip-code, parameters, results and the HRV values (Health Benchmark) are taken into consideration. Health Benchmark is defined as the upper limit above which a chemical is considered to be hazardous. As the first step, the VOCs having values above the specified health benchmark are identified and separated from the other VOCs. For this purpose we use the health benchmark as the threshold to separate the hazardous and non-hazardous VOCs.

It is observed that among the 61 VOCs which are present, 46 VOCs have values within the health benchmark and 15 VOCs have their values above the suggested health benchmark. For the remaining part of this work we will be considering these 15 VOCs containing a total of 1462 samples out of which 1225 samples have values below health benchmark and the remaining 237 samples have their values above the health benchmark.

With this filtering of dataset, we apply the K-means clustering technique to segregate the area into eleven different zones and also determine the hazardous chemical present in each particular zone.

### *A) Clustering the area into different zones:*

From the dataset, the longitudes, latitudes, postal codes and threshold values are taken as inputs in K-Means clustering algorithm, which generates a plot with eleven clustered regions showing the total number of samples of those fifteen VOCs present in a particular region. Along with this it generates zone numbers for particular chemical readings, which are used for further analysis.

Next, we go on to identifying only the hazardous samples of those 15 VOCs present in each region. So, the number of chemicals unsafe for a particular zone can be found out.

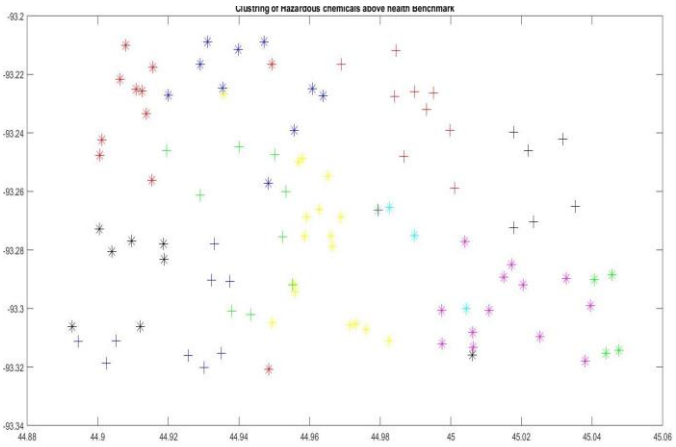


Figure 1: Plot of clustering of Hazardous chemicals above health Benchmark using K-Means

The figure (1) above shows the plot of the different zones obtained after the clustering technique.

After clustering the data using K-Means algorithm, we obtain the cluster number corresponding to all the chemical readings in the dataset. With the help of this we can segregate our data in the dataset, zone wise.

Having the threshold segregated chemical values and the zone numbers we determine the hazardous chemical and their quantity in a given zone. The main aim of this part of the work is to determine the hazardous chemical in a given region, which has been carried out successfully.

Ratio of number of Hazardous samples and the number of total samples for each chemical present in a particular region are plotted in bar charts as shown below:

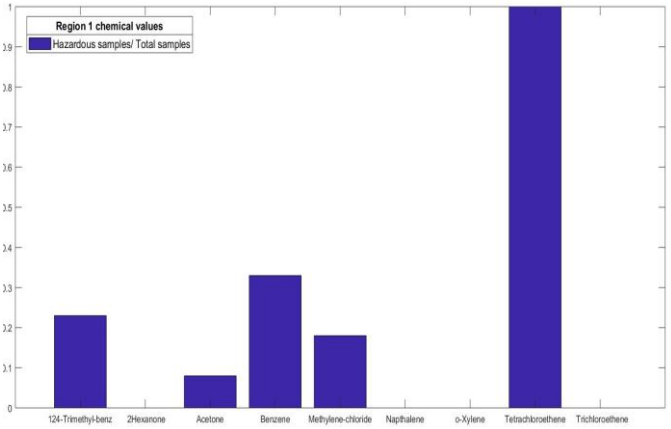


Figure 2: Region 1 Chemical Values

Figure (2) shows a bar chart of the different chemicals present in region 1, whose values are above the health benchmark. It is evident from the chart that the chemical Tetrachloroethene is present in an anomalous proportion compared to other hazardous chemicals present in the area.

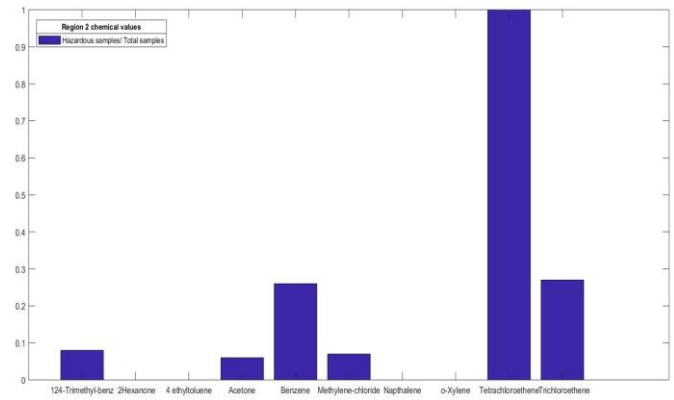


**Figure 3: Region 2 Chemical Values**

Figure (3) shows a bar chart of the different chemicals present in region 2, whose values are above the health benchmark. It is evident from the chart that the chemical Tetrachloroethene is present in an anomalous proportion compared to other hazardous chemicals present in the area.

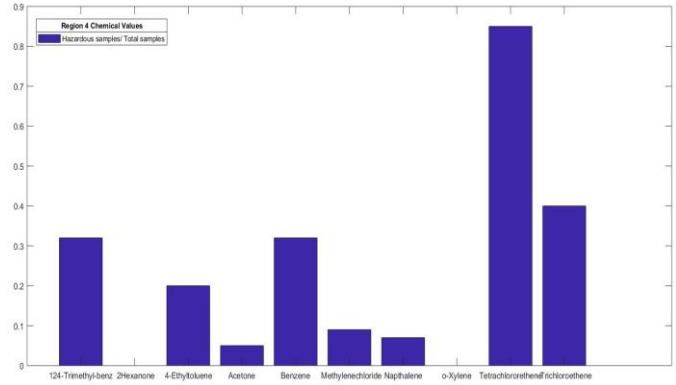


Figure 4: Region 4 Chemical Values

Figure (4) shows a bar chart of the different chemicals present in region 4, whose values are above the health benchmark. It is evident from the chart that the chemical Tetrachloroethene is present in an anomalous proportion compared to other hazardous chemicals present in the area.

From all the bar charts it can be concluded that Tetrachloroethene is present in an anomalous proportion in almost all the regions.

### *B) Detection of faulty Sensors:*

The second part of this work is to determine the faulty sensors which provided the anomalous readings for a particular chemical in a specific region. As discussed earlier we will be implementing the CUSUM control chart for the detection of these faulty sensor nodes.

The theory for CUSUM algorithm has been discussed in section III, it can be noted that the detection threshold plays a prominent role in reducing the false positives and false negatives. As the detection threshold, we vary the standard deviation for every chemical reading to determine the anomalous sensor reading.

The final outcome of this part of the work is to determine the regions having the faulty sensor. The dataset we are working here consists of 139 streets with each of them having a sensor to collect the VOC values. So, using the CUSUM algorithm we can exactly track the street in which the faulty sensor node was placed.

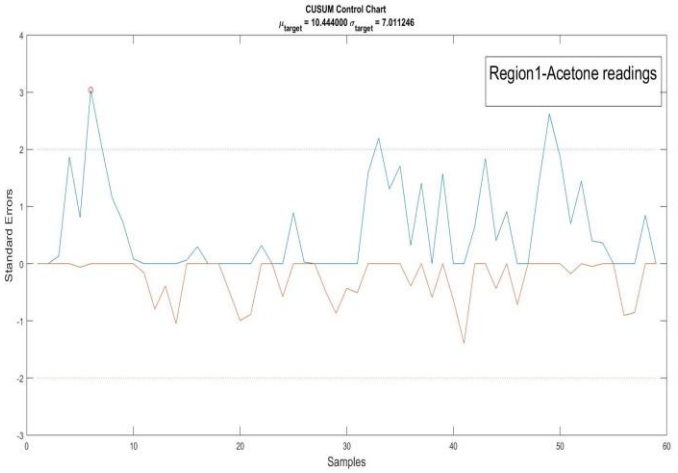


Figure 5: Showing faulty sensor values in Region 1 for Acetone readings

The figure (5) above denotes that there are three faulty sensors in region 1 showing the anomalous acetone readings above the threshold of predefined standard deviation. The faulty sensors in region 1, corresponding to these three readings are placed in 2344 River Pointe Circle, 3419 4th St N and 4242 Webber Parkway.

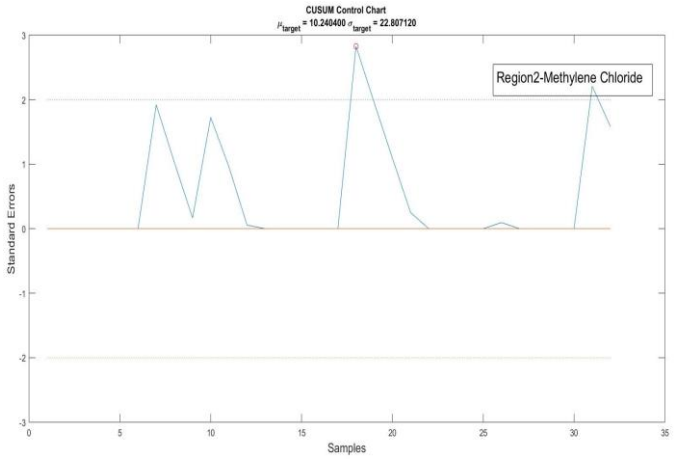


**Figure 6: Showing faulty sensor values in Region 2 for Methylene Chloride readings**

The figure (6) above denotes that there are two faulty sensors in region 2 showing the anomalous Methylene Chloride readings above the threshold of predefined standard deviation. The faulty sensors corresponding to these two readings are placed in 3712 Snelling Ave and 502 24th St E.

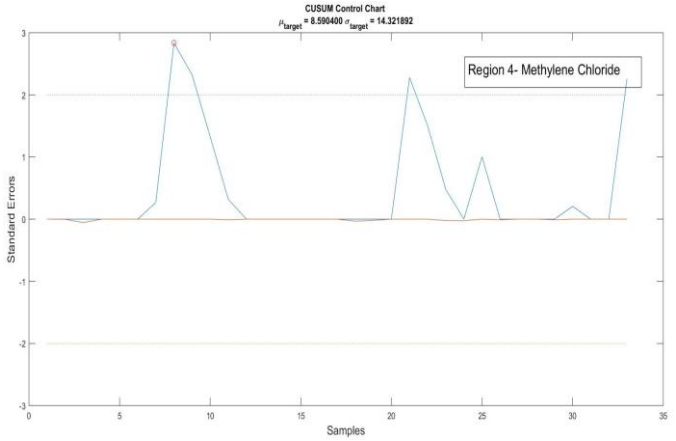


**Figure 7: Showing faulty sensor values in Region 4 for Methylene Chloride readings**

The figure (7) above denotes that there are three faulty sensors in region 4 showing the anomalous Methylene Chloride readings above the threshold of predefined standard deviation. The faulty sensors corresponding to these three readings are placed in 3118 Lake St W, 5157 Oakland Ave S and 5615 21st Ave.

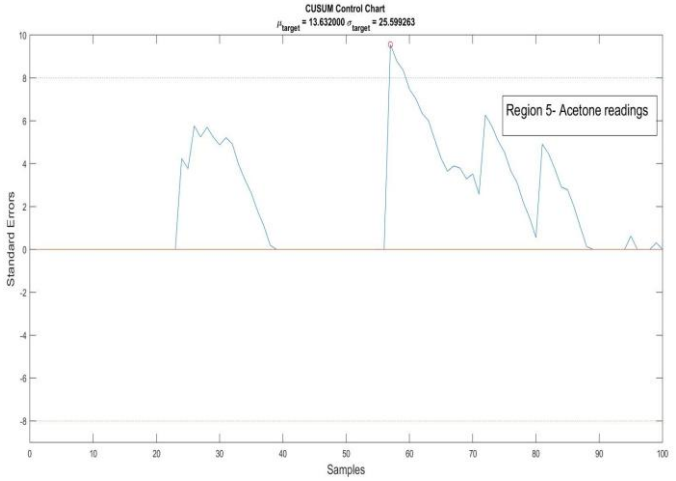


**Figure 8: Showing faulty sensor values in Region 5 for Acetone readings**

The figure (8) above denotes that there are one faulty sensor in region 5 showing the anomalous acetone readings above the threshold of predefined standard deviation. The faulty sensor corresponding to this reading is placed in 3440 19th Ave S.

The sensors which were found to be faulty, by the CUSUM Algorithm, have been traced back in the original dataset, to verify the authenticity of the work. The data samples for a particular chemical which were detected as erroneous by the algorithm, originally had values far different from the other ones in the same region, as verified from the dataset. So those sensors definitely stood out to be anomalous and faulty. This could also be verified by introducing some noise in the sample values, and obtain those particular values as faulty sensor readings by the CUSUM Algorithm.

## V. Conclusion and Future Work

In this work we apply different anomaly detection techniques for finding out hazardous chemical readings from the air quality monitoring dataset. The work here is two folded, in the first part we apply K-Means clustering technique to cluster the entire city of Minneapolis into eleven different zones and then determine the hazardous chemical in each zone.

In the second part of the work we applied a statistical quality control chart called CUSUM for detection of faulty sensor nodes in a particular zone. Here, we had to choose the detection threshold for the algorithm in-order to balance between the false positives and false negatives.

Applying the techniques of K-Means clustering and CUSUM algorithm we could successfully determine the anomaly in our dataset. However, in the process of our implementation we have found a few improvements that could be applied as an enhancement to this work, they are:

(1) Implementing an algorithm to fix the detection threshold for our CUSUM algorithm. This observation was due to the simple reason that we had to manually fix the threshold value for different zones and every chemical.

(2) We could get better results on clustering if we could apply the K-means algorithm powered by some computationally efficient optimization algorithm rather than a simple K-means clustering. We had a tradeoff, when we had to cluster the city into different zones, the clustering area was significantly changing with the clustering number.

## References


[1] Chandola, V., Banerjee, A., and Kumar, V. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages.

[2] Amin Ghafouri, Aron Laszka, Abhishek Dubey, and Xenofon Koutsoukos. 2017. Optimal Detection of Faulty Tra• c Sensors Used in Route Planning. In Proceedings of the 2nd Workshop on Science ofSmart City Operations and Platforms Engineering, Piˆsburgh, PA USA, April 2017 (SCOPE 2017), 6 pages.

[3] Air quality: A neighborhood approach report.

[4] ES Page. 1954. Continuous inspection schemes. Biometrika 41, 1/2 (1954),100–115