Comparison of clustering methods for identification of outdoor measurements in pollution monitoring

Xu Yang\textsuperscript{1}, Lingxi Zhu\textsuperscript{1}, Sio Lam\textsuperscript{2}, Laurie Cuthbert\textsuperscript{2}, Yapeng Wang\textsuperscript{2}

\textsuperscript{1} Macao Polytechnic Institute, School of Public Administration, Macao SAR, China
\textsuperscript{2} Macao Polytechnic Institute, Information Systems Research Centre, Macao SAR, China
E-mail: xuyang@ipm.edu.mo

Abstract. This paper considers the problem of post-processing air pollution data to clearly identify outdoor clusters, by removing indoor data and “noise” caused by air from indoors mingling with air from outdoors. In this paper, several different clustering algorithms are compared using data from measurements in Macao. It is shown that \textit{X-means} generally outperforms the others for this purpose and can successfully separate data modified by noise. Such a technique simplifies the collection of large data sets since the person taking the measurements does not have to make any advance decisions about what is pure outdoor, or pure indoor, data. However, it is also shown in this work that setting up suitable procedures can be quite complex.

1. Introduction
Monitoring air pollution is now routine in many cities with governments striving to reduce the well-known risks to their citizens from outdoor pollution. However, a lesser considered effect is the air quality in the home or the office, with pollution arising from gases such as radon and carbon monoxide, mould and pollen, tobacco smoke, and materials used in the building such as asbestos and formaldehyde. In a dense area like Macao, these pollutants can mix with outdoor readings with air flow through windows and doors.

Monitoring of outdoor air pollution monitoring systems has traditionally used static stations located at points where either the authorities expect there to be a problem, or where they cover particular areas of interest. Such stations tend to be fairly sparsely distributed – as an example the Guangdong-Hong Kong-Macao Pearl River Delta Regional Air-Quality Monitoring Network has only 23 stations, of which 4 are in Hong Kong and only 1 in Macao (Taipa Grande) [1].

Another approach is mobile sensing, either using specially equipped vehicles or by operatives walking round taking miniature sensors with them. Compared to static monitoring stations, such mobile stations make it easier to cover a larger area in the city and have lower cost. They also have the advantage that (i) they allow a large amount of data to be collected and (ii) the data is collected at locations where people actually are, so allowing a better picture of how people’s health may be affected. These systems can use location services (like the location services in mobile phones) to record location and time information along with the pollution data, so that users can collect pollution data from both indoor and outdoor environments. However, in a large-scale deployment, it is difficult to identify whether a location is indoors or outdoors, or to expect every user to mark whether their current location is indoors or outdoors, so some means of separation has to be used.

However, the situation is more complicated than that because data can be affected by “noise”. Here we define noise to be the effect of mixing data from indoor and outdoor locations – for example, sitting
outside a restaurant may actually be outdoors, but the readings will be affected by pollutants from indoors coming through open windows and doors, and of course, *vice versa*. This sort of noise is generated by the data source itself, but there can be other sorts of noise, such as the warm-up times for sensors – good measurement practice should ensure that sensors are properly stabilised before taking readings, but it is possible that readings are taken early. [2]

In [3] and [4] it is claimed that the effect of noise in a dataset can be small if measures have been taken to minimise the noise. In this paper though, the noise can be very significant since we do not know in advance the locations to avoid because of contamination between indoor and outdoor data. It is not feasible to mark the location of every record as containing noise data or not – indeed, it would be impossible to tell. Hence there needs to be a method for identifying which points are (for example) outdoors and which represent outdoors plus noise. This is the subject of this paper.

In [5] the authors identified four techniques to enhance data analysis in the presence of high noise levels: (i) clustering-based, (ii) distance-based, (iii) an approach-based on whether an object can be considered an outlier and (iv) their proposed hyperclique technique. Here we are considering a clustering-based approach and note that the algorithm itself can detect outliers. In [6] and [7], the authors used the principle that, once the data has been clustered, noise objects are the farthest from the cluster centre. The authors in [8] took another approach and treated small clusters that are far from other major clusters, as sets of outliers. Obviously, the choice of clustering algorithm is important and can make a difference as to which objects are treated as outliers and the difficulty of deciding that may well depend on the type of dataset being considered and the characteristics of the noise.

In this paper we consider a situation where a noise cluster may be quite close to a cluster of data that is needed – for example, an indoor location and outdoors immediately adjacent to it, but with an open window in between as illustrated in Figure 1.

![Figure 1. Illustration of the problem](image)

In this simple case it is assumed that it is summer in Macao and the outside temperature is much hotter than indoors, where air conditioning is used; the room is filled with people, so the CO\(_2\) level is higher than outside. Humidity in Macao is generally high so the use of air conditioning inside would lead to lower values indoors.

At other times of the year, the temperature levels will be different, and the CO\(_2\) level will depend on the use of the room. Hence, an algorithm that is capable of separating data into noise clusters (where they exist) as well as the main indoors and outdoors, would be very useful.

Automatically determining the number of clusters has been one of the most difficult problems in data clustering. Most methods for automatically determining the number of clusters cast it into the problem of model selection. Usually, clustering algorithms are run with different numbers of cluster types \(K\); the best value of \(K\) is then chosen based on a predefined criterion.

In the set we have chosen, the X-Means and the Cascading K-Means do not need to have a value of \(K\) specified, just the range in which \(K\) is expected to lie. The algorithms themselves find the best value of \(K\) as well as the clusters. This allows a comparison of automatic choice of \(K\) to be compared with the more traditional approach mentioned above.

In this paper, the performance of the algorithms is compared, and it is shown that X-Means is clearly the best choice, although there are caveats on how the tests are conducted.
2. Clustering algorithms

Because these algorithms are generally well-known, they are not described in any detail in this paper, the comments here being restricted to a brief mention of the main features. A very comprehensive description of clustering techniques is given in [9] and that book covers all the methods below with the exception of Cascading K-Means.

2.1. Algorithms requiring k to be specified in advance

- **K-Means** is a well-known and popular iteration-based clustering method with the term stemming from [10]. The performance of the algorithm performance depends on the initial settings and there is, therefore, no guarantee of it converging to the global optimum. Because of this, there is a common practice of running the algorithm repeatedly with different starting conditions. Additionally, this algorithm is very sensitive to outliers, which have great impact on the updates of the cluster centres. However, it does have the advantage of being fast.

- **Agglomerative Hierarchical Clustering** [11] creates a hierarchical structure that reflects the order in which groups are merged or divided. With this approach, the choice of linkage algorithm is important as they do have different characteristics.

- **Bisecting K-Means** [11] combines K-Means and Hierarchical Clustering. From an initial position with all objects in a single cluster, it splits this cluster into two datasets. One is then selected for the second data partition and the algorithm repeats. This algorithm is superior to the K-Means algorithm in the selection of initial centre points. However, the initial two cluster centres are randomly selected, and this can lead to large deviations if the noise points or outliers happen to be chosen as the initial cluster centres; it is also computationally quite intensive.

- **K-Medoids** [12] is thought to be more resilient to noise and outliers than K-Means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances, but this approach has a higher computational complexity.

2.2. Algorithms that determine the value of K

- **X-Means** [13] searches the space of cluster locations and number of clusters efficiently to optimize the Bayesian Information Criterion [14] or the Akaike Information Criterion [15]. In this algorithm, the number of clusters is computed by the algorithm itself, the user only having to specify the limits of the range of values that K can take.

- **Cascading K-Means** [16]. In this approach, the user specifies the range K can take, and the algorithm creates partitions forming a “cascade” from the smallest value of K to the largest. The best partition is selected using the Calinski-Harabasz [17] criterion.

3. Experiments Setting

3.1. Data normalization

Data Normalization standardizes the raw data by converting them into specific range using a linear transformation in order to generate good quality clusters and improve the accuracy of clustering algorithms. There is no universally defined rule for normalizing the datasets. [18]

In this paper, the attribute Min-Max normalization scheme [19] is applied for each dataset of each experiment. The experimental results in that paper show that Min-Max normalization leads to fewer errors than the Zero Mean and Decimal Scaling methods. In this work, we used the attribute Min-Max normalization method on the air pollution data, the results showing that this normalization improved the clustering accuracy.
The attribute Min-Max normalization method performs a linear transform on the original data so that if we know the maximum and minimum value of a given attribute \( v_i \), it is easy to transform the attribute into a normalised value \( x_i \) within the range \([0,1]\) by
\[
x_i = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)}.
\]

3.2. Data Clustering Platforms
In this work, four data clustering algorithms (K-Means algorithm, agglomerative hierarchical clustering, X-Means algorithm and Cascading K means algorithm) are performed using the well-known data clustering platform Weka [20]. The other two algorithms (Bisecting K-Means algorithm and K-Medoids algorithm) were implemented on another popular data clustering platform, ELKI 0.7.1[21].

3.3. Evaluation Methods
A good description of evaluation methods is given in [22]. Each air pollution data set in the experiments reported here is labelled with one of five class labels: (i) indoors, (ii) indoors with noise, (iii) at door, (iv) outdoors with noise, or (v) outdoors. From [23], evaluating these clustering results requires an external criterion and we choose two well-known external indices: Precision and Recall [24], [25]. Precision and Recall describe the accuracy and completeness, respectively, of the predicted clusters. These two measures can be calculated using the following:

\[
\text{Precision}(i, j) = \frac{n_{ij}}{n_j}, \quad \text{Recall}(i, j) = \frac{n_{ij}}{n_i}
\]

where \( n_{ij} \) is the number of objects of class \( i \) that are in cluster \( j \), \( n_j \) is the number of objects in cluster \( j \), and \( n_i \) is the number of objects in cluster \( i \).

Precision is the fraction of data put in the correct cluster (so a measure of what happens to the data), and Recall is the fraction of data in the cluster that belongs to it (so a measure of the cluster). Here we are interested in the outdoors cluster only and want to ensure that only outdoor data gets included – so we are looking for a Recall value of 100%. However, a Precision less than 100% means that some of the outdoors data will be lost to other clusters: with a large data set this is not a problem as long as the Precision is not too low.

4. Experiments
These experiments were carried out using sensors that recorded values of (i) PM1.0, (ii) PM2.5, (iii) PM10, (iv) HCHO, (v) CO\(_2\), (vi) Temperature and (vii) Humidity. Data points were recorded every 10s and each data set comprised at least 120 readings (i.e. 20 mins). The same sensor was used in each location, which meant that there was a time shift between the readings at each location, but experiments showed that the variation between sensors was likely to be greater than that over such time intervals and this is backed up by historical data.

Readings were taken in June 2018 around midday; in this season the temperature in Macao is usually hot (around 28°-34°) and there is widespread use of air conditioning indoors. CO\(_2\) depends very much on the number of people in an indoors locations and humidity can be very variable, especially if there is cooking indoors. Our results showed that PM values did not vary much between indoors and outdoors, and HCHO readings were zero. In different seasons the variation between indoors and outdoors would be different.

Figure 2 and Table 1 show example results, averaging the data set at each location. In Figure 2 the data is normalised using the scheme explained in Section 0, whereas Table 1 shows the actual average values at each location. The CO\(_2\) and temperature values follow the general outline of Figure 1 but notice the humidity dips inside the room, but nearer the door – this appears to be because of the air flow with the location of the air conditioning unit in the room, but it serves as an example as to the variations that might be expected in practical situations.
Figure 2. Example results showing difference between indoors and outdoors (normalised data values)

|                  | CO₂  | Temperature | Humidity |
|------------------|------|-------------|----------|
| Indoors          | 506  | 25.2        | 43.8     |
| indoor + noise   | 487  | 26.2        | 35.5     |
| noise            | 446  | 29.0        | 50.8     |
| outdoor + noise  | 445  | 28.8        | 53.6     |
| Outdoor          | 410  | 32.2        | 52.9     |

There are 5 possible clusters in this scenario, although the “at door” and “outdoors + noise” have similar values. Initially the different algorithms were applied to data sets containing all the attributes, i.e. including PM1, PM2.5, PM10 and HCHO, with the results summarised in Table 2. Each data set overall is of the order of 800 readings.

Knowing that data samples were taken from five different locations, the value of $K$ was set to 5 for the fixed-$K$ methods, and the range 2-5 was offered to those two methods that determine $K$.

All methods had difficulty separating the data for the locations “at door” and “outdoors with noise”, with K-Means Bisecting and K-Mediods mixing the data across the two clusters and Hierarchical and X-Means essentially ignoring the “at door” cluster; X-Means also determines that there are only 4 clusters. K-Means performs poorly with fairly low values for most performance measures and Cascading K-Means cannot cope at all, determining there are only 2 clusters.

Looking at the data in Figure 2, it is not really surprising that the methods have difficulty separating data between the “at door” and “outdoors + noise” since there is a lot of similarity. Hence another set of experiments was undertaken with the same data sets, but this time with the fixed value $K$ methods using a value $K = 4$, the value determined by X-Means. These results are also shown in Table 2. K-Means Bisecting and K-Mediods have similar results to those previously for the “outdoors” cluster and K-Means is similarly as bad as the earlier results. However, the real surprise is Hierarchical that does not put any data in the outdoors cluster.

The conclusion from these tests is that applying these methods to this data set could lead to unpredictable results, although X-Means does determine the number of clusters correctly (given how similar two of them are) and gets good results for Precision and Recall of the “outdoors” cluster (the main focus) and the two indoors sets.
Table 2. Results of clustering across all data sets

| Method             | Indoors | Indoors+noise | At door | Outdoors+noise | Outdoors | Specified | Found |
|--------------------|---------|---------------|---------|----------------|----------|-----------|-------|
| K-Means Bisecting  | 92%     | 100%          | 100%    | 100%           | 57%      | 63%       | 67%   |
|                    | 100%    | 100%          | 100%    | 54%            | 99%      | 100%      | 5     |
| The data at the door and the noise data outside the door are mixed into two clusters |
| K-Medoids          | 93%     | 100%          | 100%    | 100%           | 61%      | 75%       | 76%   |
|                    | 100%    | 100%          | 100%    | 53%            | 100%     | 100%      | 5     |
| The data at the door and the noise data outside the door are mixed into two clusters |
| K-Means            | 0%      | 0%            | 51%     | 100%           | 59%      | 64%       | 63%   |
|                    | 0%      | 0%            | 100%    | 56%            | 100%     | 57%       | 5     |
| Hierarchical       | 100%    | 99%           | 100%    | 100%           | 0%       | 0%        | 51%   |
|                    | 100%    | 100%          | 100%    | 100%           | 0%       | 0%        | 51%   |
| Data at the door and the noise data outside the door are put into the same cluster |
| X-Means            | 91%     | 100%          | 100%    | 100%           | 0%       | 0%        | 53%   |
|                    | 100%    | 100%          | 100%    | 99%            | 99%      | 100%      | 2-5   |
| Data at the door and the noise data outside the door are put into the same cluster |
| Cascading K-Means  | 0%      | 0%            | 49%     | 100%           | 0%       | 0%        | 35%   |
|                    | 0%      | 0%            | 100%    | 0%             | 0%       | 0%        | 2-5   |
| Fails very badly |

Table 3. Results of clustering across temperature, CO₂ and humidity

| Method             | Indoors | Indoors+noise | At door | Outdoors+noise | Outdoors | Specified | Found |
|--------------------|---------|---------------|---------|----------------|----------|-----------|-------|
| K-Means Bisecting  | 100%    | 100%          | 100%    | 100%           | 74%      | 83%       | 81%   |
|                    | 100%    | 100%          | 100%    | 81%            | 71%      | 100%      | 100%  | 5     |
| Fails very badly |

It is clear from the above discussion that the clustering might be considered a bit “hit and miss” and that while some methods give good results, a small change in configuration may cause that method to stop working. In our scenario, 5 of the 8 types of data did not change as much between location as CO₂, Temperature and Humidity. These sets of similar data reduce the effective distance between clusters, so another set of similar experiments was carried out, but this time using only those three parameters. The results of these are shown in Table 3.

Table 3. Results of clustering across temperature, CO₂ and humidity

| Method             | Indoors | Indoors+noise | At door | Outdoors+noise | Outdoors | Specified | Found |
|--------------------|---------|---------------|---------|----------------|----------|-----------|-------|
| K-Means Bisecting  | 100%    | 100%          | 100%    | 100%           | 77%      | 97%       | 96%   |
|                    | 100%    | 100%          | 100%    | 96%            | 71%      | 100%      | 100%  | 5     |
| Fails very badly |

Clearly these results are much better without the dilution caused by the sets of parameters that were very similar. K-Means still produces poor Recall on the “outdoors” set and misses the “indoors” cluster completely when K is set to 4. All the methods fail to separate the “at door” and “outdoors + noise” clusters, but as before this is to be expected since the parameters are very similar.

Overall, X-Means, not only determines a suitable value of K, it also produces Precision and Recall values that as good as any other method and performs well in the clustering, for all the clusters (assuming the lumping together “at door” and “outdoors + noise” is accepted), not just “outdoors”. In this case, Cascading K-Means also finds an appropriate value of K.
In Winter, the temperature will be reversed with higher temperatures inside the building, although the humidity will still be lower indoors (shown in Figure 3). As mentioned earlier CO\textsubscript{2} levels generally depend on the occupancy so will vary enormously. Overall though, there is expected to be sufficient variation for clustering to identify successfully the “outdoors” and “indoors” cluster at least. The issue is how well this approach would work in Spring and Autumn when the temperatures indoors and out may be very similar, as may be humidity if air conditioning is not used. This is an area of further work.

From Figure 4 it is clear that the temperature outdoors varies a lot with the season, but not so much the average humidity, although it was particularly low when the results for this paper were taken.

Figure 4. Average values of temperature (highest and lowest) and humidity in Macao for 1981-2010

5. Conclusions

This work has shown that it is possible to separate out noise clusters from the outdoor results required, something that is necessary in a densely populated area like Macao. However, care does need to be taken in identifying the parameters that do vary and can be used for the clustering, avoiding dilution of the data by removing parameters that are similar across all clusters. Of all the methods investigated, X-Means was clearly the best as it determined correctly the number of clusters, assuming that it is reasonable to merge “at door” with “outdoors+noise”, and performed as well as others in terms of Precision and Recall. Putting that value as K in the fixed-K methods produced no noticeable improvement.

Performing clustering over the three parameters that vary most is only used to identify where that’s et of values belongs; the values of PM samples and other pollutants can than be tied to that cluster for analysis of the pollution map.

Further work will investigate the performance in different seasons in Macao and locations showing different physical characteristics.

6. References

[1] Meteorological and Geophysical Bureau M S 2018 Guangdong-Hong Kong-Macao Pearl River Delta Regional Air Quality Monitoring Network - January to March 2018 - Statistical Summary of the First Quarter Monitoring Results
[2] Fisk W J, David F and Douglas S P 2008 Accuracy of CO2 Sensors Lawrence Berkeley Natl. Lab. 9
[3] Orr K 1998 Data quality and systems theory Commun. ACM 41 66–71
Redman T C 1998 The impact of poor data quality on the typical enterprise Commun. ACM 41 79–82

Xiong H, Pandey G, Steinbach M and Kumar V 2006 Enhancing data analysis with noise removal IEEE Trans. Knowl. Data Eng. 18 304–19

Larsen B and Aone C 1999 Fast and effective text mining using linear-time document clustering Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’99 KDD ’99 (New York, NY, USA: ACM) pp 16–22

Zhang T, Ramakrishnan R and Livny M 1996 BIRCH Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD ’96 (New York, NY, USA: ACM) pp 103–14

Portnoy L, Eskin E and Stolfo S 2001 Intrusion Detection with Unlabeled Data Using Clustering Proceedings of ACM CSS Workshop on Data Mining Applied to Security pp 5–8

Aggarwal C C and Reddy C K 2013 Data clustering: algorithms and applications (CRC press)

MacQueen J B 1967 Some methods for classification and analysis of multivariate observations Proceedings of the 5th Berkeley Symposium on Mathematics Statistics and Probability (Berkeley, Calif.: University of California Press) pp 281–97

Steinbach M 2000 A Comparison of Document Clustering Techniques KDD Work. 400 1–2

Mirkin B G 2005 Clustering for data mining: a data recovery approach (Chapman and Hall/CRC)

Pelleg D, Pelleg D, Moore A W and Moore A W 2000 X-means: Extending K-means with efficient estimation of the number of clusters Proc. Seventeenth Int. Conf. Mach. Learn. table contents 1 727–734

Schwarz G 1978 Estimating the Dimension of a Model Ann. Stat. 6 461–4

Akaike H 1974 A new look at the statistical model identification IEEE Trans. Autom. Control 19 716–23

Milligan G W and Cooper M C 1985 An Examination of Procedures for Determining the Number of Clusters in a Dataset Psychometrika 50 159–79

Calinski T and Harabasz J 1974 A dendrite method for cluster analysis Commun. Stat. - Theory Methods 3 1–27

Visalakshi N K and Thangavel K 2009 Impact of normalization in distributed k-means clustering Int. J. Soft Comput. 4 168–72

Saranya C and Manikandan G 2013 A study on normalization techniques for privacy preserving data mining Int. J. Eng. Technol. 5 2701–4

Witten I H, Frank E and Hall M A 2011 Data Mining: Practical Machine Learning Tools and Techniques (Google eBook) (Morgan Kaufmann)

Schubert E, Koos A, Emrich T, Züfel A and Klaus Arthur Schmid A Z 2015 [ELKI 0.7] A Framework for Clustering Uncertain Data Proc. VLDB Endow. 8 1976–89

Jain A K 2010 Data Clustering: 50 Years Beyond K-Means1 Pattern Recognit. Lett. 31 651–66

Jain A K and Dubes R C 1988 Algorithms for clustering data (Prentice-Hall, Inc.)

Rendón E, Abundez I, Arizmendi A and Quiroz E M 2011 Internal versus external cluster validation indexes Int. J. Comput. Commun. 5 27–34

Chormunge S and Jena S 2015 Efficiency and Effectiveness of Clustering Algorithms for High Dimensional Data Int. J. Comput. Appl. 125