Clustering Imputation for Air Pollution Data

Wedad Alahamade[1,3], Iain Lake[1], Claire E. Reeves[2], and Beatriz De La Iglesia[1]

[1] University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
W.Alahamade, i.lake, c.reeves, B.Iglesia @uea.ac.uk
http://https://www.uea.ac.uk/
[2] Centre for Ocean and Atmospheric Sciences, School of Environmental Sciences,
University of East Anglia
[3] Taibah University, Medina, Saudi Arabia
https://www.taibahu.edu.sa/

Abstract. Air pollution is a global problem. The assessment of air pollution concentration data is important for evaluating human exposure and the associated risk to health. Unfortunately, air pollution monitoring stations often have periods of missing data or do not measure all pollutants. In this study, we experiment with different approaches to estimate the whole time series for a missing pollutant at a monitoring station as well as missing values within a time series. The main goal is to reduce the uncertainty in air quality assessment.
To develop our approach we combine single and multiple imputation, nearest neighbour geographical distance methods and a clustering algorithm for time series. For each station that measures ozone, we produce various imputations for this pollutant and measure the similarity/error between the imputed and the real values. Our results show that imputation by average based on clustering results combined with multiple imputation for missing values is the most reliable and is associated with lower average error and standard deviation.

Keywords: Air Quality · Uncertainty · Time Series Clustering · Imputation.

1 Introduction

Air is one of the essential natural resources not only for humans but for all life on this planet. With the development of economies throughout the world and population increases in cities, environmental problems involving air pollution have attracted increasing attention. Air pollution is defined as the contamination of the atmosphere by substances called *pollutants*. According to Kampa and Castanas [11], the air pollutants that negatively affect human health and the environment include carbon monoxide (CO), particulate matter ($PM_{2.5}$ and PM_{10}), ozone (O_3), nitrogen dioxide (NO_2) and sulfur dioxide (SO_2).

There are several detrimental effects of air pollution on health and the environment and its effect on human health in particular attracts considerable research effort. For example, several epidemiological studies have proven the associations between air pollutants and asthma e.g. [6], mortality e.g. [28] and morbidity e.g. [7]. The World Health Organization [26], estimated that 4.2 million premature deaths per year are linked to air pollution.

The air pollutant concentrations that are used to determine the air quality index in the UK are O_3, NO_2, SO_2, PM_{10}, and $PM_{2.5}$. These are measured at the Automatic Urban and Rural Network (AURN), which has 168 stations distributed around the UK. There are 165 stations that measure NO_2, 86 stations that measure $PM_{2.5}$, 82 stations that measure PM_{10}, and 83 stations that measure O_3. Stations are categorized by their environmental type to one of the following: rural, urban background, roadside, and industrial. Not all the AURN stations report all pollutants, as this mainly depends on the purpose of the monitoring station. Even though a station measures a particular air pollutant there are times them no data are reported, for example during periods of instrument failure or servicing. Together this results in high levels of missing data. Therefore current air quality assessments are based on high levels of uncertainty. This may lead to incorrect policy decisions, with further negative environmental and health consequences [8]. Our aim is therefore to investigate robust methods for estimating the missing observations.

In this study we focus on imputation of ozone (O_3), one of the main pollutants influencing pollution levels in the UK. We apply two different approaches to estimate the missing pollutant in a station: an imputation based on geographical distance, and one based on clustering. We then assess which results in more robust and accurate imputation. In the long term we hope that our imputed pollutant values, if accurate, will enable us to calculate new air quality indices and to show where more measurements may be beneficial.

2 Related work

In the context of environmental data various techniques have been proposed to impute missing values using single imputation such as mean [14,16]; Nearest Neighbor (NN) algorithms [29,3]; linear interpolation [2] and Expectation Maximization (EM) [10]. Other authors ([18,16]) have replaced each missing value of PM_{10} by the mean of the two data points before and after the missing value. In a similar study, Luong et al. [14] used the daily mean of each variable to replace the missing values. Their dataset contains temperature, PM_{10}, $PM_{2.5}$, NO_2, and SO_2.

Zheng et al. [29] used station spatial or temporal neighbours to fill the missing values in each monitoring station. Then, they built a model to forecast the readings of an air quality monitoring station over the next 48 hours for $PM_{2.5}$. Azid et al, [3] used NN based on distance to impute the missing data.

Arroyo et al. [2] used multiple regression techniques (linear and nonlinear) and artificial neural network models to impute the missing values in the daily data averages of ozone based on the concentrations of five other pollutants: NO, NO_2, PM_{10}, SO_2, and CO. Jhun et al. [9] estimated the missing data for an O_3 hourly trend dataset using criteria in the dataset such as the hour of the day, season at each region, and the seasonal pattern of the trend. The Expectation Maximization (EM) algorithm is often used to fill the missing data using available data in the cases when the missing data has a linear relation with the available data [10]. Some other studies deleted any incomplete data and only considered data that are captured between 50% to 75% of the time [9,27].

3 Problem definition

3.1 Air Quality

The quality of air is negatively affected by particles and gases which can be influenced by several factors including location, time, and other variables [12]. In the UK, air quality is quantified using the Daily Air Quality Index (DAQI) which is calculated using the concentrations of five air pollutants namely nitrogen dioxide (NO_2), sulphur dioxide (SO_2), ozone (O_3), particles < 2.5 ($PM_{2.5}$), and particles < 10 (PM_{10}). This index is numbered from 1 to 10, and divided into four bands: 'low' (1–3), 'moderate' (4–6), 'high' (7–9) and 'very high' (10). An index value is initially assigned for each pollutant depending on its measured concentration. Then the DAQI is taken to be the highest value assigned to a pollutant. Periods of poor air quality can be identified using this index. Air quality is negatively correlated with the DAQI index, meaning that a higher DAQI index represents worse air quality (for more details `https://uk-air.defra.gov.uk/air-pollution/`).

There are DAQI values calculated for different stations and geographical areas. However, sufficient data must be available for this. For example, to calculate the particles(PM_{10} , $PM_{2.5}$) daily mean contributing to the index, 75% of the daily observations must be captured; otherwise, the pollutant is considered as missing that day. Moreover if there is no measurement for a pollutant, then the DAQI is based on the concentrations of just those pollutants measured. This means that if, for example, the PM10 concentration was such that it had the highest index for an individual pollutant, but it's concentration was not measured, then the DAQI, which would be determined by the measured pollutant with highest index, would give an unrealistically low value. This would give the impression that the air quality is better than it actually is.

3.2 Data Imputation Methods

To impute the missing data, there are two main methods available: single imputation and multiple imputation. In single imputation each missing value is imputed by only one estimated value. An easy though naive imputation is to replace with the mean or most commonly occurring value [1]. The main drawback of this method is that does not reflect the uncertainty inherent in missing data [21].

Multiple imputation is a statistical technique, that replaces each missing value with a set of plausible (n) values. The results of the multiple imputation methods are (n) datasets [22]. The differences between these datasets reflect the uncertainty of the missing values [25].

One of the most effective multiple imputation methods is Multivariate Imputation via Chained Equations (MICE), also known as Sequential Regression multiple imputation [19]. It is based on Fully Conditional Specification (FCS). Each incomplete variable is imputed by a separate model on a variable-by-variable basis so each variable can be modeled according to its distribution. For example, continuous variables can be modeled using linear regression, binary variables modeled using logistic regression and categorical data using polytomous regression [25]. For a time series (TS), predictive mean matching (pmm) can be used in the imputation process [4].

3.3 Imputation of Air Quality Measurements

For any given station, j, and pollutant i we can approximate the pollutant concentration P_i^j over time using a number of methods. For example, since geographical distance or similarity in the type of station may be relevant we could construct a nearest-neighbour approach based on similarity or distance measures. Alternatively, we can use a form of clustering or grouping of stations to obtain values from other stations in the same cluster which appear to be most similar to the j station.

3.4 Evaluation

If real values are known, we can compare our imputation to those real values in order to evaluate which imputation method works best. Hence, for our experimental set up we take each existing Time Series (TS) for a given pollutant and station, P_i^j in turn, and impute it by the various methods to obtain an imputed TS, PI_i^j. This enables evaluation with a 'ground truth'.

We can compare the real values to the imputed values by a number of measures including distance and regression error measures. We used the Root mean squared error (RMSE), which measures the average magnitude of the errors between the actual and the imputed data. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2} \tag{1}$$

where in our case x_i represent the observed data points and $\hat{x}_i$ represent the imputed values.

The method that gives the lowest error on average for all stations (i.e. imputed TS) will be the considered the best method. Note that the best methods may change from one pollutant to another and may be affected by other factors such as station type (e.g. urban background, rural and roadside) or frequency of data measurement (e.g. hourly, daily).

To provide a more robust testing scenario we separate the 'model building' stage for the imputation from the testing stage. We use an initial data period of three years as a training set to build the imputations, and then impute on the next year of the TS to evaluate goodness of fit.

4 Proposed Methods

We have two levels of missing data in our TS: partial and total. The first corresponds to missing observations within the TS for a given pollutant. The second corresponds to a pollutant not being measured at all for the station.

4.1 Imputing Missing Observations

We imputed the missing observations of a measured pollutant in each station using single and multiple imputation methods; then we applied a TS clustering algorithm to each complete dataset. For single imputation, we used a Simple Moving Average (SMA) method. This method replaces each missing value using a weighted moving average. The mean value in this method is calculated from an equal number of observations on either side of a central missing value; the user can identify the length of that window [15]. In our experiment, we set the window length to 30, so the missing value is replaced by the monthly moving average before and after the missing value.

For multiple imputations, we used MICE to impute the missing value with n different values. In our experiment, we set n=5.

4.2 Imputing Missing Pollutant Time Series

4.2.1 Nearest (geographical) Neighbours imputation To impute the missing pollutant P_i at station j, we first looked at geographically close stations. For this, we measured the geographic distance between station j and all other stations that measure pollutant P_i using the Harvison metric which calculates geographic distance on earth based on longitude and latitude as follows:

$$\begin{aligned} a &= \sin^2(\Delta\varphi/2) + \cos\varphi_1.\cos\varphi_2.\sin^2(\Delta\lambda/2) \\ c &= 2.(\sqrt{a}, \sqrt{1-a}) \\ d &= R.c \end{aligned} \tag{2}$$

where φ represents latitude, λ represents longitude, and R is earth's radius (mean radius = 6,371km).

Then to impute pollutant P_i for station j we use:

- The nearest neighbour (**1NN**) using the Harvison distance to station j.
- The average of the two nearest neighbours (**2NN**) to station j.

4.2.2 Clustering for imputation Clustering requires a measure of similarity or distance between objects, points, groups, or TS. In this exercise, we experimented with two distance metrics suited to TS.

Dynamic Time Warping (DTW) is a distance measure that is used to find the optimal alignment (shortest path) that minimizes the sum of distances between two TS. It was proposed by Sakoe and Chiba [23]. It is an extension of the Euclidean Distance measure (ED) that offers a non-linear alignment between two series.

Shape-Based Distance (SBD) is a faster alternative to DTW, and is based on the cross-correlation with coefficient normalization (NCCc) sequence between two series. It was proposed as part of the k-Shape clustering algorithm by Paparrizos et al. [17] and for its application the TS data should have appropriate amplitudes, or be z-normalized in order to get better clustering results using SBD metric. The SBD distance is calculated by the following formula:

$$SBD(X,Y) = 1 - \frac{max(NCC_c(x,y))}{\|x\|_2 \|y\|_2} \tag{3}$$

where $\|.\|_2$ is the l_2 norm of the series calculated as the square root of the sum of the squared vector values. SBD range lies between 0 and 2, with 0 indicating perfect similarity [24].

In our experiment, SBD gave well separated clusters that were more compact than those obtained using DTW, so we report it in our results.

Once we have defined a distance measure a clustering algorithm will group objects according to their distance/similarity. Partition clustering algorithms divide the data points into non-overlapping subsets/clusters. The best-known partitioning algorithm is the k-medoids, also called Partitioning Around Medoids (PAM). It was proposed by Kaufman and Rousseeuw [13]. The cluster medoids act as the cluster 'centers', which are the most representative objects of a cluster. The average dissimilarity between medoids and all data points in the cluster is minimised. The concept of cluster medoids is similar to cluster centroids, but medoids are always members of the data set and may not be located at the center of the cluster, whereas centroids may not correspond to real objects.

PAM requires us to identify the number of the cluster (K) before running the algorithm. To do that, we used Silhouette index (Sil), which is a well-known measurement for estimating the number of clusters in a dataset proposed by Rousseeuw et al. [20].

Hence we used PAM as a clustering algorithm to produce a clustering of the stations. If station j belongs to cluster C_i, given the measured pollutant over time, then, to impute pollutant P_i based on the clustering results, we use:

- The cluster medoid, (**CM**), to impute the missing pollutants at station j, P_i^j.
- The average of pollutant P_i in cluster C_i, (**CA**) which is the daily average of pollutant P_i in all the stations that fall in this cluster.

4.3 Experimental Framework

All our proposed methods were implemented in R. We divide our experiment into two phases: the first phase is imputation of missing observations and clustering process based on the training set as shown in Fig. 1. In general, the clustering results obtained from each individual dataset created by MICE are slightly different hence we merged them into one final clustering result using the majority voting. Majority voting [5] is a simple ensemble technique which chooses the cluster for a station chosen by the majority of the clustering results. The first stage results in a set of complete TS and clustering results.

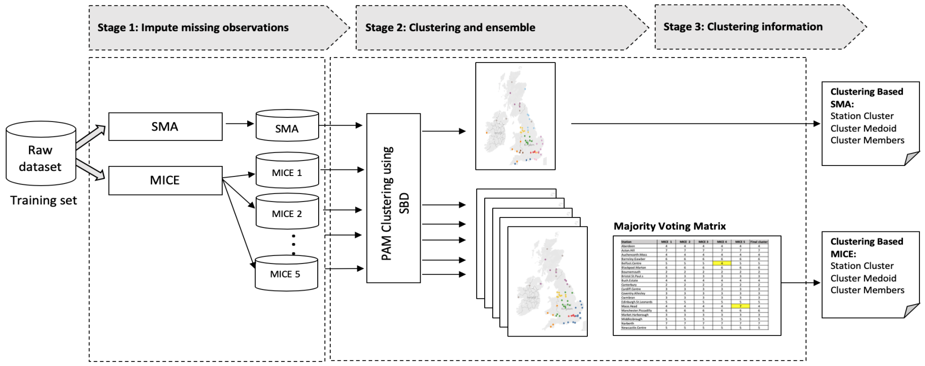


Fig. 1: Phase 1: TS missing observations imputation and clustering process.

The second phase is our proposed imputation method for TS in the test period, as shown in Fig. 2. The imputation of missing observations takes place for the test data as it did for the train data, however in the test set we combined the MICE datasets into one by averaging the n imputed values for each individual observation creating one value. Then, based on the clustering results from the first phase, we assigned a cluster number and cluster medoids for each station. Then the clustering results and NN imputation (1NN and 2NN) are used to produce whole imputed TS for each station.

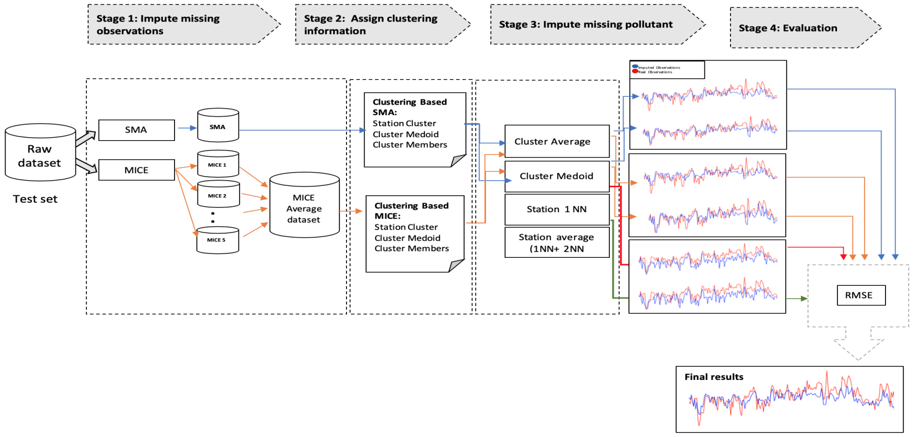


Fig. 2: phase 2 : Air pollutant imputation methods.

5 Air pollutants concentrations Dataset

The dataset generated at AURN stations include multivariate TS data that show the hourly concentrations of different air pollutants. In this study we only focused on stations that measure ozone. The data can be obtained from `https://uk-air.defra.gov.uk/data/data_selector`. The observations we download from each station included date, time, hourly pollutant concentration for each pollutant measured at the station, and Status (R =Ratified, P=Provisional, P*=As supplied).

In total, there are 83 stations around the UK, in which the ozone (O_3)is measured. We removed 18 stations that have more than 25% of missing data. In total we included 65 stations in our analysis. Fig. 3, shows the geographical distribution of these stations. We divided the dataset into two parts: the training set including observations for a period of three years (2015-2017); and the test set including the observations of the following year (2018).

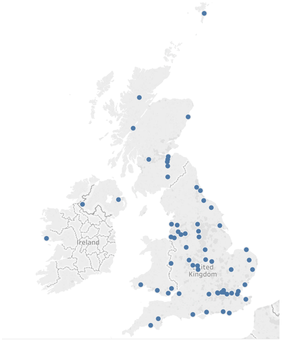
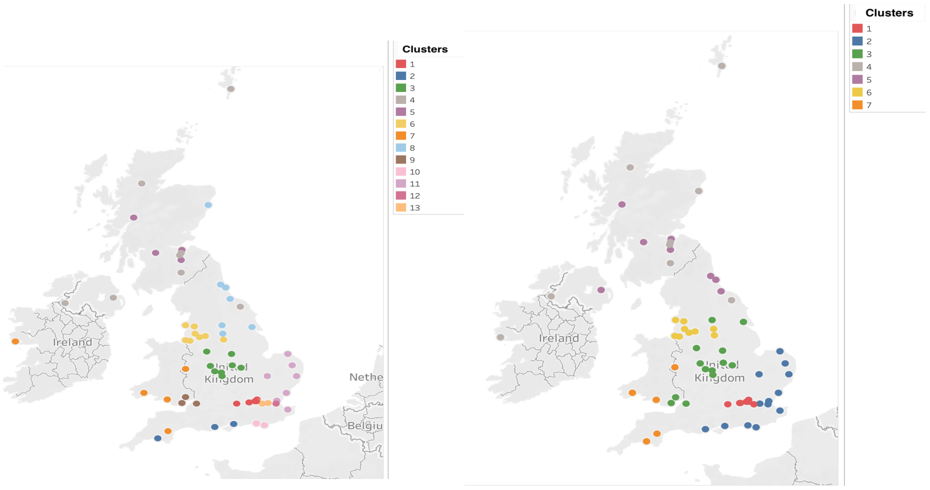


Fig. 3: Geographical distribution of ozone monitoring stations in the UK used in the experiment.

6 Results

Although the optimal number was always 2, this did not provide us with a sufficiently granular clustering result. We used in each case the second best number of clusters which was 13 using SMA and 7 using MICE. For MICE, the optimal number of clusters did not vary for the 5 datasets. The visualisation of the clusters obtained are shown in Fig. 4. As can be observed, even though we only used the pollutant concentrations as TS to cluster the stations using the temporal similarity as measured by SBD, the results of the clustering show that there is a spatial (geographical) correlation between stations in each cluster.

According to our 4 described methods, using the Cluster Medoid (**CM**), the Cluster average (**CA**), the 1NN (**1NN**) and the average of the 2NN (**2NN**), we created 8 different imputed TS, 4 for the SMA dataset and 4 for the combined MICE datasets. We evaluate those by calculating the RMSE to measure the dif-



(A) Clustering results of SMA dataset.

(B) Clustering results of the combination of MICE datasets clustering.

Fig. 4: Clustering results of training datasets using SMA and MICE imputation methods.

ference between the imputed and the real data for each station. For each station, we ranked our imputation methods for each dataset based on the value of RMSE from smallest to largest, hence the best imputation method for SMA will have a rank of 1, etc., and similarly for the MICE combined dataset. We then compare these methods based on the average ranks to select the best imputation method. Table. 1 shows the comparison of the average of RMSE and the average rank for all methods from all stations for both the MICE and SMA datasets, respectively, using the Cluster Average (**CA**) is associated with the minimum average rank $(2.25, 2.37)$, the minimum average of error $(10.003, 10.315)$, and the minimum standard deviation $(3.901, 4.218)$. This is followed by **2NN** and then **CM** with **1NN** providing the worst results.

An example of the 4 imputed TS for one station (Glasgow Townhead) for the period of six months (Jan-Jul of 2018) using SMA (top) and MICE (bottom) datasets is shown in Fig. 5. It shows that all the imputations reproduce the trend well, though they may generate slightly higher values. Some periods, early in the year appear to show more deviation and this may be due to temperatures having an effect. MICE with CA appears to produce the closest results.

Fig. 6 shows a different station, "Glazebury". In this figure, the red TS represents the real observations at the stations with some missing values in the middle of the TS. The green TS represents imputed missing observations using SMA (Top), and MICE (bottom). On the other hand, the blue TS is the result of imputing the whole pollutant TS using the Cluster Average **CA**.

It is worth noting that using the Cluster Medoid (**CM**) to impute the missing pollutant is not possible in some cases. If the station we are going to impute is itself the medoid of the cluster, or if the cluster has only one station then we

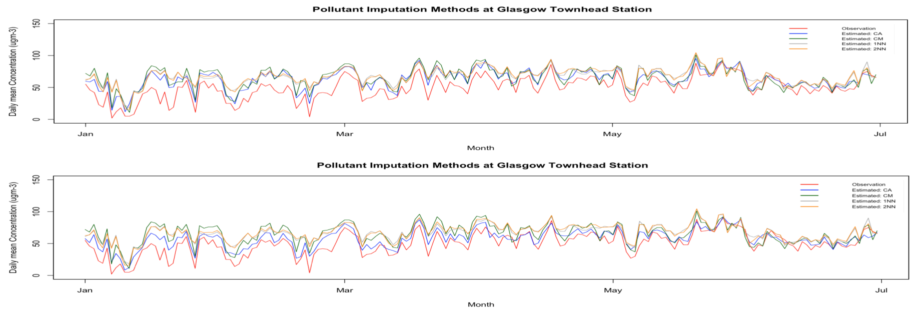


Fig. 5: Pollutant imputation methods using SMA dataset (top) and MICE (bottom) at Glasgow Townhead station.

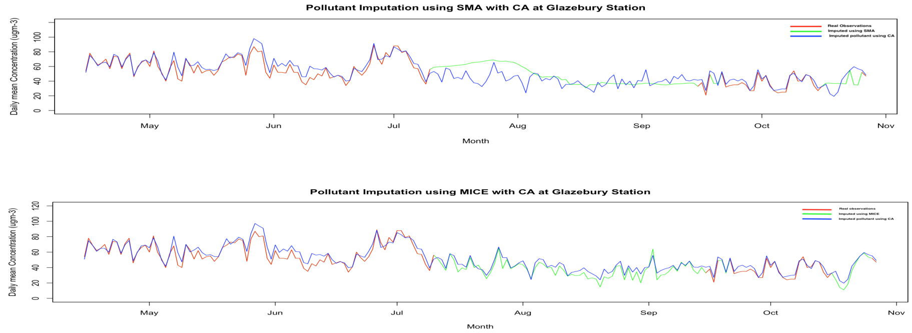


Fig. 6: Observations and **CA** imputed TS using SMA (top) and MICE (bottom) at Glazebury station.

have no feasible imputation hence we record how many stations the imputation was possible for as the last column of Table 1. As we can see from the table, it is not possible to create a cluster average for the SMA dataset at "Rochester Stoke" station, because the cluster has only that station. In this case we cannot use the Cluster Average and the Cluster Medoid in the imputation process.

7 Conclusions and Future Work

We have proposed and compared a number of techniques to impute ozone values for a station when missing partially or completely. We found that using the clustering average as obtained by clustering the stations on the pollutant values over time using SDB and PAM to impute the missing pollutants gave better results compared to other techniques. This was true regardless of the method used to impute missing observations (partial imputation). However, the combination of multiple imputation for partial missing values and cluster average for pollutant imputation gave the best results.

Our future work is to apply this method to all air pollutants that contribute to the DAQI, then calculate the DAQI from the imputed pollutants and compare it with the historical reported DAQI based on datasets with missing data. From

Table 1: The average RMSE and rank comparison for each methods in the two datasets.

Methods	Average Rank	Average Errors (RMSE)	Standard deviation (std)	Station contributing
		MICE Dataset		
Cluster Average (CA)	**2.25**	**10.003**	**3.901**	65
Cluster Medoid (CM)	2.78	11.410	4.544	58
First neighbor (1NN)	2.86	12.116	5.417	65
Average of 2NN	2.41	10.784	5.272	65
		SMA Dataset		
Cluster Average (CA)	2.37	10.315	4.218	64
Cluster Medoid (CM)	2.61	11.692	4.404	52
First neighbor (1NN)	3.05	12.641	5.263	65
Average of 2NN	2.53	11.266	5.121	65

that we can identify any deviations between DAQI values calculated with more information (i.e. the imputed information) and the historical reported DAQI, and this may highlight stations where more measurements will be beneficial, for example where inclusion of the measurement of another pollutant at an AURN station will likely lead to a more accurate DAQI.

References

1. Allison, P.D.: Missing data, vol. 136. Sage publications (2001)
2. Arroyo, Á., Herrero, Á., Tricio, V., Corchado, E., Woźniak, M.: Neural models for imputation of missing ozone data in air-quality datasets. Complexity **2018** (2018)
3. Azid, A., Juahir, H., Toriman, M.E., Kamarudin, M.K.A., Saudi, A.S.M., Hasnam, C.N.C., Aziz, N.A.A., Azaman, F., Latif, M.T., Zainuddin, S.F.M., et al.: Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in malaysia. Water, Air, & Soil Pollution **225**(8), 2063 (2014)
4. Buuren, S.v., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of statistical software pp. 1–68 (2010)
5. Dimitriadou, E., Weingessel, A., Hornik, K.: Voting-merging: An ensemble method for clustering. In: International Conference on Artificial Neural Networks. pp. 217–224. Springer (2001)
6. Gass, K., Klein, M., Chang, H.H., Flanders, W.D., Strickland, M.J.: Classification and regression trees for epidemiologic research: an air pollution example. Environmental Health **13**(1), 17 (Mar 2014)
7. Gore, R.W., Deshpande, D.S.: An approach for classification of health risks based on air quality levels. In: Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on. pp. 58–61. IEEE (2017)
8. Holnicki, P., Nahorski, Z.: Emission data uncertainty in urban air quality modeling—case study. Environmental Modeling & Assessment **20**(6), 583–597 (2015)
9. Jhun, I., Coull, B.A., Schwartz, J., Hubbell, B., Koutrakis, P.: The impact of weather changes on air quality and health in the united states in 1994–2012. Environmental Research Letters **10**(8), 084009 (2015)

10. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M.: Methods for imputation of missing values in air quality data sets. Atmospheric Environment **38**(18), 2895–2907 (2004)
11. Kampa, M., Castanas, E.: Human health effects of air pollution. Environmental pollution **151**(2), 362–367 (2008)
12. Kang, G.K., Gao, J.Z., Chiao, S., Lu, S., Xie, G.: Air quality prediction: Big data and machine learning approaches. International Journal of Environmental Science and Development **9**(1), 8–16 (2018)
13. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons (2009)
14. Luong, L.M., Phung, D., Sly, P.D., Morawska, L., Thai, P.K.: The association between particulate air pollution and respiratory admissions among young children in hanoi, vietnam. Science of the Total Environment **578**, 249–255 (2017)
15. Moritz, S., Bartz-Beielstein, T.: imputets: time series missing value imputation in r. The R Journal **9**(1), 207–218 (2017)
16. Norazian, M.N., Shukri, Y.A., Azam, R.N., Al Bakri, A.M.M.: Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia **34**(3), 341–345 (2008)
17. Paparrizos, J., Gravano, L.: k-shape: Efficient and accurate clustering of time series. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1855–1870. ACM (2015)
18. Plaia, A., Bondi, A.: Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment **40**(38), 7316–7330 (2006)
19. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P., et al.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology **27**(1), 85–96 (2001)
20. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
21. Rubin, D.B.: An overview of multiple imputation. In: Proceedings of the survey research methods section of the American statistical association. Citeseer (1988)
22. Rubin, D.B.: Multiple imputation for nonresponse in surveys, vol. 81. John Wiley & Sons (2004)
23. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing **26**(1), 43–49 (1978)
24. Sardá-Espinosa, A.: Comparing time-series clustering algorithms in r using the dtwclust package. Vienna: R Development Core Team (2017)
25. Van Buuren, S., Oudshoorn, K.: Flexible multivariate imputation by MICE. Leiden: TNO (1999)
26. WHO: Ambient air pollution: Health impacts. `https://www.who.int/airpollution/ambient/health-impacts/en/` (2019)
27. Wong, C.M., Vichit-Vadakan, N., Kan, H., Qian, Z.: Public health and air pollution in asia (papa): a multicity study of short-term effects of air pollution on mortality. Environmental health perspectives **116**(9), 1195–1202 (2008)
28. Yang, Y., Li, R., Li, W., Wang, M., Cao, Y., Wu, Z., Xu, Q.: The association between ambient air pollution and daily mortality in beijing after the 2008 olympics: a time series study. PloS one **8**(10), e76759 (2013)
29. Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., Li, T.: Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2267–2276. ACM (2015)