A Multi-variate Time Series clustering approach based on Intermediate Fusion: A case study in air pollution data imputation

Wedad Alahamade[a,b], Iain Lake[c], Claire E. Reeves[c], Beatriz De La Iglesia[a]

[a]*School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK*
[b]*School of Computing Sciences, Taibah University, Medina 42353, Saudi Arabia*
[c]*School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK*

Abstract

Multivariate Time Series Clustering (MVTS) is an essential task, especially for large and complex dataset, but it has received limited attention in the literature. We are motivated by a real-world problem: the need to cluster air pollution data to produce plausible imputations for missing measurements for some pollutants. Our main focus will be on the UK air quality assessments, the study uses data collected from automatic monitoring stations during four-year period (2015-2018).

In this work, we propose a MVTS clustering method followed by an imputation methods for the whole Time Series (TS). We compare two approaches to cluster the stations: univariate TS clustering using Shape-Based Distance (SBD) for individual pollutants, and MVTS clustering using the fused similarity that combines the SBD for all the pollutants. We run a k-means algorithm to produce clusters with each approach on the same dataset.

Our analysis shows that using MVTS clustering produces the best clusters as measured by various quality indexes and by the imputations they help to reduce the error average between imputed and real values based on the Root Mean Squared Error (RMSE) and its standard deviation (Std) .

Keywords: Time series clustering, Multivariate Time Series (MVTS), Fusion, Uncertainty, Air Quality, Imputation.

Email addresses: W.Alahamade@uea.ac.uk (Wedad Alahamade), i.lake@uea.ac.uk (Iain Lake), c.reeves@uea.ac.uk (Claire E. Reeves), B.Iglesia@uea.ac.uk (Beatriz De La Iglesia)

1. Introduction

Time Series(TS) is a sequence of observations that a variable takes over time, such as (t_1, v_1),...,(t_i, v_i),...(t_m, v_m), where t_i is the time step and v_i is the observation. The order in the time series data is important since the values are based on time. When several variables are observed and recorded simultaneously, this becomes a Multivariate Time Series (MVTS).

A large variety of real-world applications use time series analysis such as weather forecasting [1], earthquake prediction [2] or human activity recognition [3]. MVTS are becoming more prominent specially as part of large and complex datasets being produced [4]. In this study, we motivated by the need to generate modelling techniques for multivariate time series data, where air pollution is an example of such data. Our focus in this paper, will be on problems related to the air pollution in the UK, especially uncertainty resulting from missing data with air quality assessment. Hence, we encounter MVTS while looking at air pollution data, our proposed approach is based on the MVTS clustering and imputation.

Air pollution is one of the main risks to human health in several parts of the world. Sources of air pollution are varied and include anthropogenic sources such as combustion (e.g.in power plants, motor vehicles and residential heating), agriculture and industry as well as natural sources such as vegetation, soils, and lightning [5].

In the UK, the main four pollutants that are used to assess the quality of the air are ozone (O_3), nitrogen dioxide (NO_2) and particulate matter less than 2.5μm in diameter ($PM_{2.5}$) or less than 10μm in diameter (PM_{10}), so we focus on those four. These pollutants are measured at various monitoring stations and the measured concentrations of each pollutant become a time series (TS) requiring further transformation and analysis to produce air quality assessments.

One of the available resources to assess air quality in the UK is the Air Pollutants Monitoring Network. The network contains air pollution monitoring stations that record the air pollutant concentrations. There are 285 air quality monitoring sites across the UK, which are part of several types of networks with different objectives and coverage. Our focus will be on the automatic monitoring network called Automatic Urban and Rural Network (AURN). The instruments used in this network are automated and produce hourly pollutant concentrations. These data are collected and stored, then made directly available via the Internet [6]. Stations in this network are categorised by their environmental type into one of the following: rural, urban, suburban background, roadside, or industrial. The total number of these stations is 169 stations and the geographical distribution of the AURN monitoring stations is shown in Fig.1.

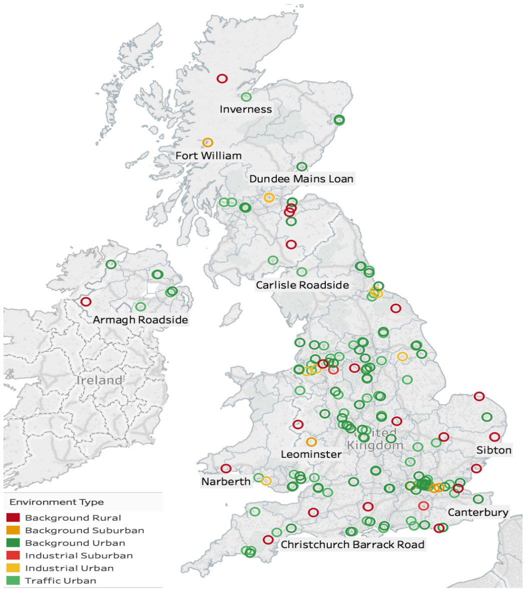


Figure 1: Geographical distribution of the air quality monitoring stations (AURN).

In the UK, air quality is quantified using the Daily Air Quality Index (DAQI) which is calculated using the concentrations of NO_2, O_3, $PM_{2.5}$, and 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 maximum value assigned to any of the pollutants. 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 see [6]).

The challenges associated with analysing air pollution data (i.e. the pollutants TS) are as follows. Not all the stations report all the pollutants and even if a station does, it may not measure a particular pollutant all the time due to instrument down-time. Together this results in high levels of missing data. Therefore current air quality assessments are based on high levels of uncertainty. As the DAQI is calculated based on the concentrations of measured pollutants only, which may not reflect the actual air pollution. This may lead to incorrect policy decisions, with further negative environmental and health consequences [7].

What makes the air pollution data analysis more complex is that pollutants have different behaviours and seasonal variation. Adding to that, pollutant can be emitted from various sources and be involved in different chemical reactions and so their concentrations exhibit different temporal and spatial distributions.

Particulate matter (PM) has lots of various sources, both primary (emitted directly into the atmosphere) and secondary (produced in the atmosphere via chemical and physical processes). Whilst PM concentrations are often greater at roadside [8], the particles can have lifetimes of several days in the atmosphere, meaning that they can be distributed widely. The larger particles are subject to greater loss via sedimentation, so $PM_{2.5}$ is more evenly distributed than PM_{10} [9].

The primary source of NO_2 comes from fuel burning such as cars, trucks and buses, power plants, and off-road equipment. This gives NO_2 a local pattern, concentrating where it is emitted in urban areas and near to the roadside [10].

Ozone is complex as it is not directly emitted into the air, but it is formed as a secondary pollutant by the reaction of nitrogen oxides (NO_x) and volatile organic compound (VOC) in the presence of sunlight [11]. So, the ozone formation depends on the VOC–NO_x ratio [12]. Ozone concentrations in urban areas have been found to be lower than those in rural areas [13], due to the presence of more NO_x in urban sites that can remove ozone via the reaction of NO with O_3 to give NO_2 and oxygen (O_2). O_3 and NO_2 are strongly anti-correlated, indicating that the O_3 is strongly depressed by high NO_x [14]. Furthermore, ozone can have a lifetime of days to weeks [15], meaning that ozone at a specific site may have been produced by NO_x and VOCs emitted from other distant locations. Ozone behaviour makes the seasonal variation with ozone concentrations, as ozone is lower in the winter due to scavenging by NO and higher in the summer due to photochemical ozone production [11]. While PM and NO_2 are at their lowest level during the summer [16].

Therefore, we aim is to investigate robust methods for estimating the missing values when there are no measurements of a particular pollutant at a site at all to reduce the uncertainty of the air quality assessment resulting from missing measurements which may be missing either partially or completely and to enhance the air quality data and provide a DAQI that is more realistic. As DAQI calculated from observed data only may give a false representation of the air quality, for example, if there were high concentrations of an air pollutant that was not being measured, the air quality may be worse than indicated by the DAQI.

To achieve our goal, we need to understand the relation between different pollutant concentrations and their geography. In particular, understanding such relations may enable us to impute missing data (including entire TS) where particular pollutants are not being measured. We postulate that in such cases pollutant measurements from other stations may act as a proxy measurement for the missing TS. Our approach to this starts with grouping stations with similar pollutant(s) behaviour in to groups using clustering algorithm. Once we have clusters, we can use those to impute various measurements for stations that may belong to a cluster with information from the cluster itself or stations within the cluster.

As known that clustering is an unsupervised learning method to group unlabelled objects into homogeneous groups [17]. Similarly for TS, we group together a set of time series with similar patterns. TS unique structure makes many traditional clustering methods unable to be applied directly [18]. One challenge for TS clustering is how to measure similarity, which is the core of any clustering algorithm. Some of the univariate TS similarity measures cannot handle missing data or TS of different lengths [19]. The problem becomes more challenging when more than one time series is involved (i.e. in a multivariate TS environment). In this work, we experiment with novel MVTS clustering approaches and evaluate them in the context of air pollution measurements, and particularly for the task of imputing missing pollutant TS. We deal with an observation-based MVTS dataset with a high level of missing data and uncertainty.

We proposed a MVTS clustering approach that starts by clustering stations based on all measured pollutants using a fusion approach that aggregates the similarity/dissimilarity of the univariate TS (pollutants) between every two MVTS (stations). This aggregated similarity represents the distance between MVTS in the k-means clustering algorithm. Then, based on the clustering results, we propose two methods to impute the whole time series for the missing pollutant at a given station.

Three experiments are carried out to demonstrate the validity of our approach. In these experiments, we compare the clustering and the imputation results obtained using MVTS clustering with imputation using the univariate TS clustering.

The structure of the paper is as follows: Section 2 discusses some of the existing TS clustering methods with their application and the limitations of the previously proposed time series clustering approaches to our case with air pollution data. Section 3 discusses in detail all the methods we used to measure the similarity between MVTS for the clustering algorithm, methods to impute the missing pollutants and evaluate our proposed solutions. Finally, in Section 5, we analyse and compare the results of our experiments, then we discuss these results in in Section 6. At the end of this paper, we conclude the work with some recommended future work in Section 7.

2. Related work

Due to the increasing availability of time series data and the demand to analyse them, clustering time series has attracted growing research interest in recent years [4, 20, 18, 21, 19, 22, 23]. However, most of the existing clustering methods are for univariate TS data, while clustering multivariate time series remains a challenging task [21].

The main problem with multivariate time series is dimensionality, and the majority of the existing researchers have proposed methods for dimensionality reduction to measure the similarity between multivariate time series (MVTS), such as Principal Component Analysis (PCA) *similarity factor* [24, 23], feature extraction methods that transform TS data into set of features [19, 20], and statistical [22] or an ensemble model to aggregate the similarity of multivariate TS components (i.e. univariate TS) [25]

There has been some research into similarity within MVTS. For example, Fontes et al. [24] proposed a MVTS clustering method based on extracted features from the univariate TS. Principal Component Analysis (PCA) is used to measure the similarity between MVTS, then fuzzy k-means is used to cluster these TS. This clustering approach was used for fault detection in a gas turbine. Li Hailin [23] proposed a multivariate time series clustering based on common principal component analysis (CPCA) to construct a projection coordinate space and to lower the dimension of the data for the clustering process. The proposed clustering approach has two main stages, the first is assigning every MVTS to a cluster based on its the similarity to the projection coordinate space (i.e. cluster prototype) and the other is to construct a new prototype of the clusters based on CPCA.

Wu et al. [26] transformed MVTS data into Independent Components (IC) using Independent Component Analysis (ICA) to find the independent patterns for each TS. Then they proposed a new clustering algorithm called ICACLUS to cluster these patterns according to the extracted ICs instead of the traditional k-means. In this algorithm, the similarity between TS is measured based on the number of matching ICs. Recently, Li et al. [27] transformed the MVTS into a network that called component relationship network (CRN) to reflect the relationship of the MVTS data, then an improved version of Dynamic Time Warping (DTW) is used to measure the similarity for each component to cluster the MVTS data.

Zhou et al. [22] developed a model-based multivariate time series clustering algorithm that first discovers the temporal patterns in each TS using confidence value to represent the relationship between different variables. Their algorithm is based on the k-means and aims to group MVTS based on the degree of patterns discovering into the same cluster.

D'Urso et al. [4] proposed robust fuzzy clustering models for MVTS based on an exponential transformation of the dissimilarities. This algorithm was applied to real-world data on the concentrations of three pollutants (NO, NO_2, and PM_{10}) in the Metropolitan City of Rome for the problem of detecting pollution alarms. Recently, Li et al. [20] proposed a multivariate time series clustering of weighted fuzzy features based on two distance measurement methods Dynamic Time Warping (DTW) and shape-based distance (SBD). They first picked initial cluster centers by fast search and find of density peaks (DPC), then a fuzzy membership matrix is generated by performing DTW on each diminution (i.e. univariate time series), then SBD is utilised to measure distances within each dimension and generate fuzzy membership matrices which is used with the fuzzy c-means clustering algorithm.

Ensemble have been applied to time series clustering, Mikalsen et al. [25] proposed a method called Time series Cluster Kernel (TCK) to learn the similarities between multivariate time series (MVTS) with missing data without using any imputation methods. This method uses an ensemble learning procedure that combined the clustering results of several Gaussian mixture models (GMM) from the final kernel to deal with uncertainty. The proposed approach achieved good results, however the main drawback of this method is that it works only on datasets of equal length, also it needs ensemble learning with numerous learning datasets that are not available in our case with air pollution data since we only have one source of this data.

Our approach is a raw-data-based approach, while all other proposed MVTS clustering techniques are either model-based or feature-based approach, that aim to transform the MVTS data into another type of data such as features, network, independent component, ..., etc to prepare the data for the machine learning algorithms. The MVTS objects (stations) in our dataset (i.e. the UK's air pollution data) do not have equal dimensions as not all pollutants TS recorded in the stations.

3. Methods

3.1. Time series analysis

As previously mentioned, that the UK air pollution data that is used in this work, has high level of missing data either partially or completely. As a pre-processing step, we impute partial missing values within the TS to create a complete dataset. Imputing the missing observations in an early stage enable us to measure the similarity between TS using univariate time series similarity measures that cannot handle missing data (i.e. Dynamic Time Warping (DTW) and Shape-Based Distance (SBD)).

In this step, we use a multiple imputation technique called (MICE) [28]. The process of multiple imputations starts with an incomplete dataset, then it imputes every missing value n times creating n completed datasets, in our experiment, we set n=5, then we averaged the imputed values for each missing value to create a final completed version of the dataset.

MICE is selected based on our initial exploratory experiments [29]. From this work, we found that using MICE to impute the TS missing observations is better than using some single imputation methods such as Simple Moving Average (SMA) for the purpose of clustering and imputation of the univariate TS.

Also in this work, we compared different time series distance measures and imputation techniques to impute the missing observations and missing pollutants (TS). We experimented with two distance metrics that are suitable for TS data, Dynamic Time Warping (DTW) [30] and Shape-Based Distance (SBD) [31]. Our analysis showed, that SBD gives better separated cluster than DTW to cluster stations based on univariate TS clustering using the k-medoids (PAM) clustering algorithm [32] to cluster stations and impute the missing pollutants using the cluster average imputation method (CA).

In the current work, we continue using the SBD to measure the similarity between two time series (i.e. individual pollutant concentrations) with the k-means clustering algorithm [17] since the Cluster Average (CA) is more useful than using the cluster medoid for the purpose of the pollutant imputation. However, this work includes more pollutants, which make MVTS clustering.

To measure the similarity between stations with all measured pollutants (i.e. MVTS), we combine the distance matrices of all the univariate TS into one matrix that represents the similarity between the MVTS. Then, we attempt an intermediate fusion approach to cluster our dataset/stations based on four TS, in our case concentrations of the air pollutants O_3, PM_{10}, $PM_{2.5}$, and NO_2.

In fusion clustering, intermediary fusion refers to algorithms that somehow use fusion to operate (e.g. by fusing distances) [33]. The intermediate fusion approach we use is an adaptation of work by Mojahed et al. [34] who used a k-medoids clustering algorithm to cluster objects that were represented by different data types (e.g. text, images and TSs). We adapt this as our objects are all TS, hence we use SBD to measure distance. From each pollutant measured in two stations we can generate a distance matrix, which is then fused in the fused matrix.

In our dataset, the distance between two stations A and B is the distance between hourly pollutant concentrations (TS) from these stations using SBD. Since we only focus on four air pollutants, we will have four distance matrices (DM) each represents the similarity between stations of each pollutant. We define the entries of DM for pollutant P_i, DM_{Pi}, as follows.

$$DM_{P_i}(A, B) = SBD(A_{P_i}, B_{P_i}) \tag{1}$$

where A and B are two stations, and SBD is use to measure the distance between concentrations of pollutant P_i in stations A and B.

Then, Fused Distance Matrix (FDM) is calculated for each pair of stations. The aggregated distance is calculated as the simple average distance of all pollutants (O_3, PM_{10}, $PM_{2.5}$, and NO_2) when they are measured or of only those pollutants that are measured at the stations. For example for stations A and B, and supposing these stations measure all the pollutants, the fused distance between these stations is:

$$FDM(A, B) = \frac{\sum_{i=1}^{p} SBD(A_{P_i}, B_{P_i})}{p} \tag{2}$$

where p is the number of pollutants. In our case $p = 4$ if all the pollutants are measured at station A and B. $SBD(A_{P_i}, B_{P_i})$ is extracted from DMs for each pollutant.

3.2. K-means MVTS Clustering

The k-means [17] is one of the most widely used clustering algorithms, that is based on the distance of objects to cluster centroids. For our MVTS data which consist of 4 pollutants, a cluster centroid is also a MVTS. A cluster centroid is calculated for each pollutant using the average TS of all the stations within the cluster. Associated with that, a new FDM is calculated by measuring the distance between all the objects (stations in our application) and the centroids for each cluster, then we fuse these distances using the simple average using Equation. 2.

We applied the basic k-means to cluster the objects (stations) based on the fused distance matrix. We start the process using randomly selected stations (this would be a medoid in clustering), but after the first iteration we compute proper centroids. The processes of running the basic k-means on the fused distance matrix (FDM) is as follows:

Initialisation:

1. Randomly select k objects/stations as the initial centroids to start with.
2. Assign all objects to the nearest centroids based on the initial fused distance matrix (FDM).

Repeat:

1. Calculate the centroid of each cluster. The centroid will now be the average of the TS from all the stations within the cluster. Since we have 4 pollutants, every centroid will have 4 time series, one for each pollutant.
2. Calculate distances between all the stations and the new centroids, therefore creating a new FDM to represent distance between the new centroids and all the data objects.
3. Re-assign the objects to the closest centroid based on the new calculated FDM (from the previous step).

Until: No change in the cluster centroids.

In our work, as we lack a reference or ground truth clustering solution, to measure the cluster quality we use the clustering internal quality measures. These measures are based on intrinsic properties of the clustering solution such as compactness, separation, and connectedness of the cluster partitions, so they are based on measurable aspects of a clustering solution [35]. The compactness is a cluster homogeneity measure that reflects how close are the objects within the cluster by measuring the within-cluster variation, while the separation is the degree of separation between clusters. It measures how well separated a cluster is from other clusters by measuring the between-cluster variation [36]. The connectedness is the connectivity between objects in the dataset. It is the degree to which neighbouring objects have been placed in the same cluster [37]. We selected Silhouette Width (ASW) [38] and Dunn index (DI) [39] as they measure the cluster compactness and separation and the connectivity (Conn) measure that reflects how connected objects are within the clusters [40]. Whereas Silhouette Width and Dunn Indices are to be maximised, connectivity should be minimised. We also compare the Within cluster sum of squares (WCSS), which measures the variability of the objects within each cluster, and between cluster sum of squares (BCSS), that measure variability between the centroids of the clusters. In general, the clustering process tries to minimise within-cluster distance and maximise between-cluster distance [41].

3.3. Imputation Methods of Missing Pollutant TS

Once a clustering of the stations is obtained, we use the clustering solution to impute missing TS (pollutants). If station j belongs to cluster C_x, ($1 \leq x \leq k$, where k is the number of clusters) given the measured pollutants over time, then, to impute pollutant P_i based on the clustering results, we use two methods:

1. **Cluster Average (CA)**: in this method, we impute the missing pollutant (P_i) using the average of that pollutant (P_i) from all stations in the cluster (c_i). Which is the hourly average concentrations of pollutant P_i in all the stations that fall in this cluster.

2. **(CA+ENV)**: in this method, we impute the average of pollutant P_i in cluster c_i, but using only stations that have the same environment

type to station j within the cluster, such as Background Rural, Background Urban, Traffic, or Industrial. This is in recognition that the type of station may be important and result in closer measurements of pollutant concentrations.

Fig. 2 visualises the clustering average imputation (CA) method using the MVTS clustering approach, CA+ENV imputation method is the same with considering the station environment type.

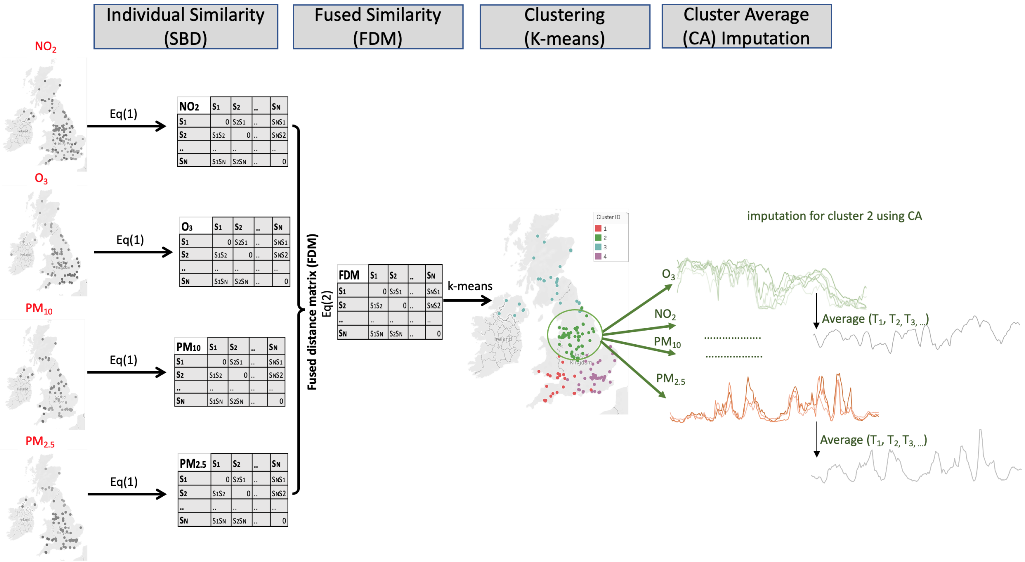

Individual Similarity (SBD)
Fused Similarity (FDM)
Clustering (K-means)
Cluster Average (CA) Imputation
NO2
O3
PM10
PM2.5
Eq(1)
Eq(1)
Eq(1)
Eq(1)
NO2
O3
PM10
PM2.5
Fused distance matrix (FDM)
Eq(2)
FDM
k-means
Cluster ID
imputation for cluster 2 using CA
O3
NO2
PM10
PM2.5
Average (T1, T2, T3, ...)
Average (T1, T2, T3, ...)

Figure 2: Visual representation of clustering imputation through the MVTS clustering process using cluster average (CA) method.

3.3.1. Imputation methods evaluation

To evaluate how plausible the imputation is using different methods, we can compare truth values to imputed values. We can do this by taking each existing TS for which we have values, one at a time, and consider them to be missing. We impute the whole TS by various methods and compare to the ground truth. We can then average the behaviour of the different imputation methods to establish the one that provides imputed values closest to the real values.

Hence, for our experimental set up we take each existing 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. We compare the real values to the imputed values using 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:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2} \tag{3}$$

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 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).

4. Experimental set up

Clustering algorithm and imputation methods were implemented in R, Version (3.5.2). 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 (2015-2017) as a training set to build the imputations, including the clustering results, and then impute on the next year (2018) of the TS to evaluate the goodness of fit.

To fully evaluate the advantages of the MVTS clustering over the univariate TS clustering using SBD for individual pollutants for the purpose of pollutant imputation, we compare the results of these approaches in our experiments.

Our experimentation design contains three main stages, as shown in Fig. 3: the first stage is the pre-processing stage that includes missing observations imputation to create a complete dataset. Then the second stage is to group/cluster stations based on their temporal similarity, which means using the stations similarity in pollutant concentrations through clustering algorithm. The third stage is to impute and evaluate whole missing pollutant TS using clustering information from the previous stage. In the following sections, we will describe each stage in details.

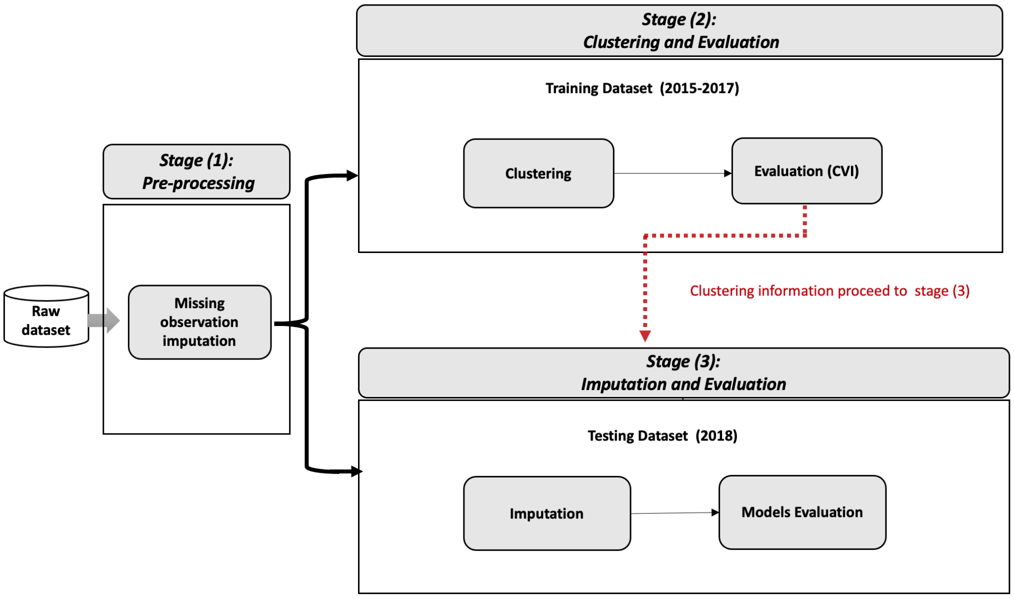

Stage (2):
Clustering and Evaluation
Training Dataset (2015-2017)
Clustering
Evaluation (CVI)
Stage (1):
Pre-processing
Raw dataset
Missing observation imputation
Clustering information proceed to stage (3)
Stage (3):
Imputation and Evaluation
Testing Dataset (2018)
Imputation
Models Evaluation

Figure 3: The overall proposed experimentation design represents the main three stages.

4.0.1. Stage (1): Pre-processing

This stage includes imputing missing observations within a TS using MICE for the dataset (2015-2018) to create a dataset with complete TS. Then, we divided the complete dataset into training set that includes the data of years 2015-2017, and test set that includes data of year 2018.

4.0.2. Stage (2): Time Series Clustering and Evaluation

In this stage, we used the training set (2015-2017) to cluster the stations based on their temporal similarity. We apply the k-means clustering algorithm with SBD to cluster the stations based on their hourly pollutant concentrations.

We experimented with two approaches: univariate and multivariate time series clustering. In the first approach, we used the basic k-means clustering algorithm to cluster the stations based on each pollutant independently using SBD to measure similarity between TS. So in this case, each pollutant is used to derive its own clusters and then imputation is based on that clustering solution so independent for each pollutant. These clustering results are fed to the next stage (stage 3) to impute pollutant concentrations using the proposed imputation methods.

In the second approach, we applied the intermediate fusion approach to cluster stations based on the four pollutants using the k-means clustering as explained in Sec.3.2 with the calculated fused distance matrix as in Equation (2). In this matrix (FDM), 12 stations out of 169 stations are removed because they measure no common pollutant, so they cannot deliver any information to the fused matrix. After removing these stations there are 157 stations left to construct the FDM. The 12 stations are allocated to the final cluster based on their similarity to the cluster centroids.

At the end of this stage, we evaluated the clustering results from these approaches to select the best clustering solution using Clustering Validity Indices (CVI).

4.0.3. Stage (3): Imputation and Evaluation Models for Missing Pollutants

In this stage, we impute the missing pollutant, which is a whole TS in the test dataset (data of year 2018) by applying all proposed imputation methods in Sec.3.3. For each pollutant at each site, we imputed the pollutant using the proposed imputation methods with our previously obtained clustering solutions. At the end of this stage, we compare and evaluate the clustering solutions in terms of the imputation methods using the RMSE (as described in Sec.3).

To fully evaluate how well our MVTS approach works for our imputation problem, we conducted three experiments. In the first experiment, we used univariate TS clustering approach. In the second and third experiments, we used MVTS clustering approach.

In the second experiment we include every station, however in the third experiment, we excluded stations that measure only one pollutant, which is always NO_2. We found that these stations are difficult to cluster using the fused similarity, because the fused distance of these stations is the distance of the only measured pollutant. As an alternative solution, we allocate these stations to the clusters based on their similarity to the cluster centroids.

5. Results

5.1. Experiment 1: Univariate TS Clustering

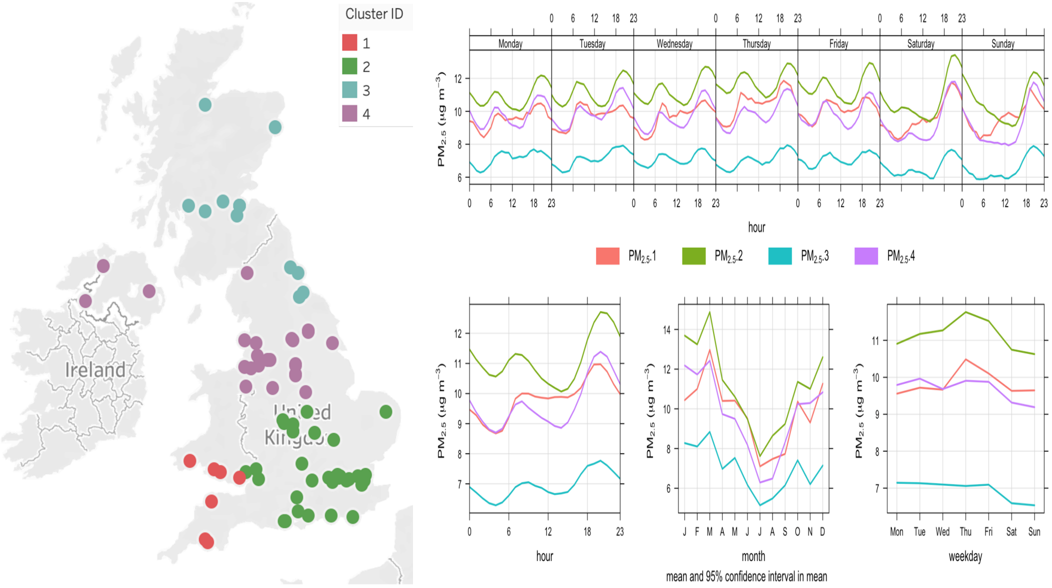
In this experiment, we applied the first approach to impute the pollutant concentrations using our proposed imputation methods through clustering individual pollutant. We analyse the clustering results of each pollutant, in the following sections.

5.1.1. $PM_{2.5}$ Clustering Results

The total number of stations that measure $PM_{2.5}$ is 77. The result of applying the basic k-means clustering algorithm to this set of stations is shown in Fig.4(a), which represents a geographical map with the stations colour coded according to the clustering results (i.e. (cluster 1, (red), (cluster 2, green), (cluster 3, light blue), and (cluster 4, purple)). There are four clusters located in four geographical locations (North, Center, South East, South West). These clusters geographically look compact and well separated even though the clustering is based on pollutant concentration values. This means that there appears to be a geographic pattern to the concentrations of $PM_{2.5}$.

To further analyse results, we show the time variation between the clusters' centroids (and we will do similarly for the other pollutants). This enables us to further understand how $PM_{2.5}$ concentrations are distributed in the UK and if there are specific time effects. The cluster centroids represent the average concentrations in the clusters for a particular pollutant. Fig.4(b), represents the variations in pollutant concentrations on each day of the week in the top graphs. Then the variation is broken down into hourly, monthly and weekday variations in the bottom graphs. From this figure, we observe the variation of $PM_{2.5}$ concentrations at each cluster centroid, it is noticed that centroid's of cluster 2 (green) is the highest, while concentrations at centroid's of cluster 3 (blue) is the lowest. While the concentrations of reaming two centroids of cluster 1 and 4 (red and purple) have similar behaviours. This variation can also be seen with the monthly, weekly and hourly analysis.

Fig. 4(b), shows that there is a graduation of the concentrations of $PM_{2.5}$ at the cluster centroids from the South (highest) to the North (lowest). That is from cluster 3 (light blue) to cluster 2 (green). This can be observed in all the graphs. The concentrations of cluster 2 (green) which represents the South East are the highest among all other cluster centroids. On the other hand, cluster 3 (light blue) located in the North has the lowest concentrations. The concentrations on clusters in the Centre of the UK (cluster 4, Purple) and South West (cluster1, Red) are very similar to one another. Concentrations of $PM_{2.5}$ are slightly lower in the weekend for all clusters and appear highest at peak hours, particularly during the evening (as shown in the bottom right plot). As we can see there are low concentrations during the summer (June, July, and August) compared to the rest of the year (as shown in the middle bottom plot).



((a)) $PM_{2.5}$ clusters.

((b)) Time variations of cluster's centroids.

Figure 4: (a) Geographical distribution of stations that measure $PM_{2.5}$ with the colour coded clusters obtained using the basic k-means algorithm, (b) Time variations of the 4 cluster's $PM_{2.5}$ centroids.

Fig.5 presents the monthly average concentration for each cluster. Historically, peak values can be seen in January and particularly during 2017 with the South being markedly higher during those peaks than the North. These geographical variations are consistent with understanding of the sources of $PM_{2.5}$ which tend to be greatest in the south of the UK and the influence of sources in continental Europe [42].

In general, there is a seasonal variation with $PM_{2.5}$ concentrations during these years (as shown in Fig.5), the concentrations tend to be higher in the winter and lower in the summer. Trend cannot be seen with the monthly concentrations in this plot, it could be noticed with the yearly mean concentrations.

5.1.2. PM_{10} Clustering Results

The total number of stations that measure PM_{10} is 75 stations. The result of clustering this set of stations is shown in Fig.6(a). There are three clusters (i.e. (cluster 1, (red), (cluster 2, green), and (cluster 3, light blue)), two large

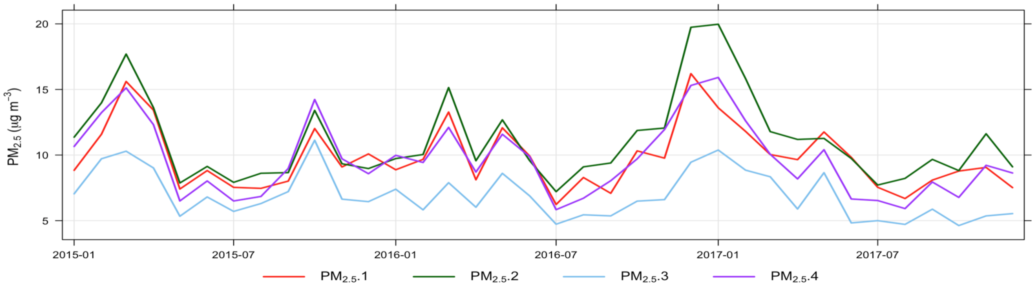


Figure 5: Monthly average concentrations of the 4 cluster's $PM_{2.5}$ centroids.

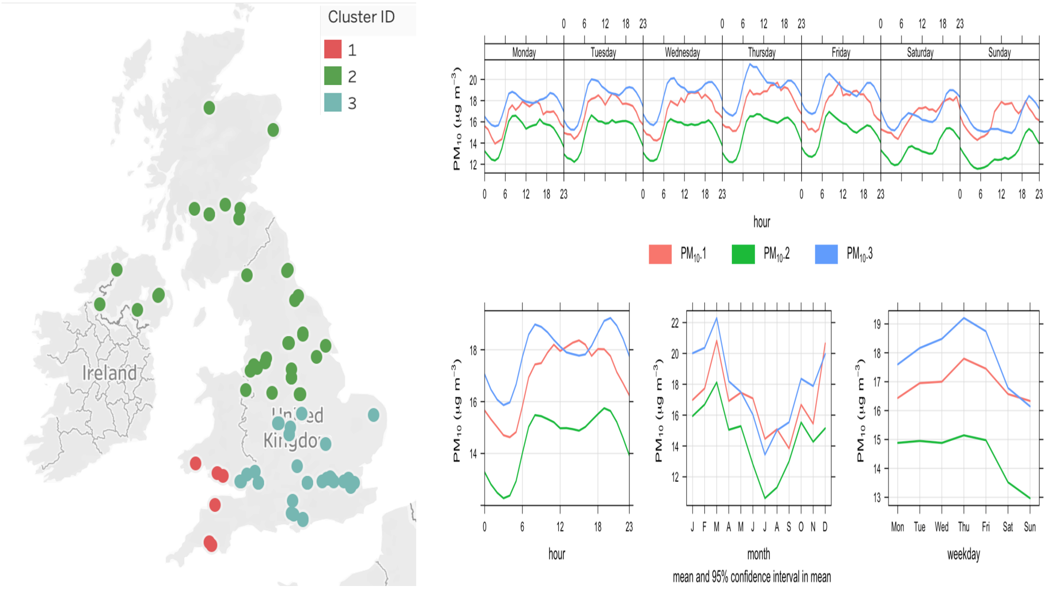
clusters located in the North and the South and a small cluster that contains six station on the South West. Fig.6(b), shows the time variation analysis of clusters centroids, the centroids of cluster 2 (green), which is located in the North, and cluster 3 (blue), which is located in the South, exhibit very similar behaviour. However, cluster 2 in the North has lower concentrations of PM_{10}. In terms of time variation, for the two main clusters, there is still some effect of day of the week with lower values at the weekend (as shown in bottom right plot), and higher peak hourly values although the variations are much less than for $PM_{2.5}$. The summer months (June, July, August) also register the lowest concentrations (as shown in bottom middle plot). The centroid of cluster 1 (red) has average concentrations compared to the other two.

Fig.7 show the monthly average concentrations of each cluster's centroid. We see similar patterns and trends in cluster 2 and 3, while cluster 1 has a higher peak during 2016. At other times, the South cluster is generally higher. The seasonal variation for PM_{10} is very similar to $PM_{2.5}$.

5.1.3. O_3 Clustering Results

The total number of stations that measure O_3 is 71 stations. The clusters obtained for these stations are shown in Fig.8(a). In this map, there are three clusters (i.e. (cluster 1, (red), (cluster 2, green), and (cluster 3, light blue)) located roughly across North/West, Center, South/East though for O_3 the geographical separation of the clusters is less clear. The clusters are separated geographically except for some stations that are light blue (cluster 3, mostly covering the North) which are mixed within the green points (cluster 2, mostly covering the Center).

Fig.8(b) shows the time variations of the cluster centroids. In general, based on the clustering results there are two levels of O_3 concentrations, high



((a)) PM_{10} clusters.

((b)) Time variations of cluster's centroids.

Figure 6: (a) Geographical distribution of stations that measure PM_{10} with the colour coded clusters obtained using the basic k-means algorithm, (b) Time variations of the 3 cluster's PM_{10} centroids.

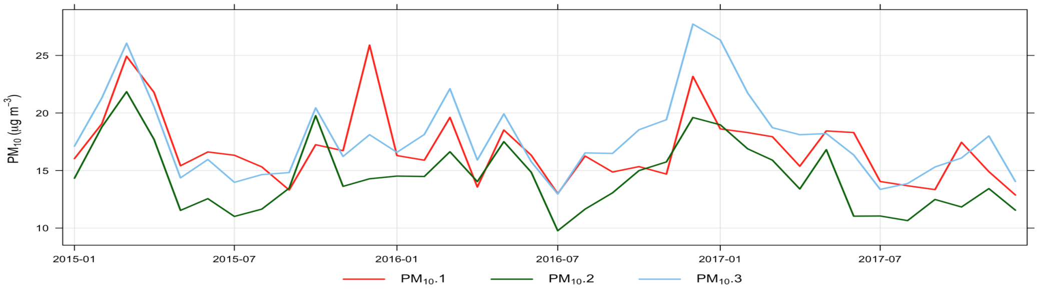
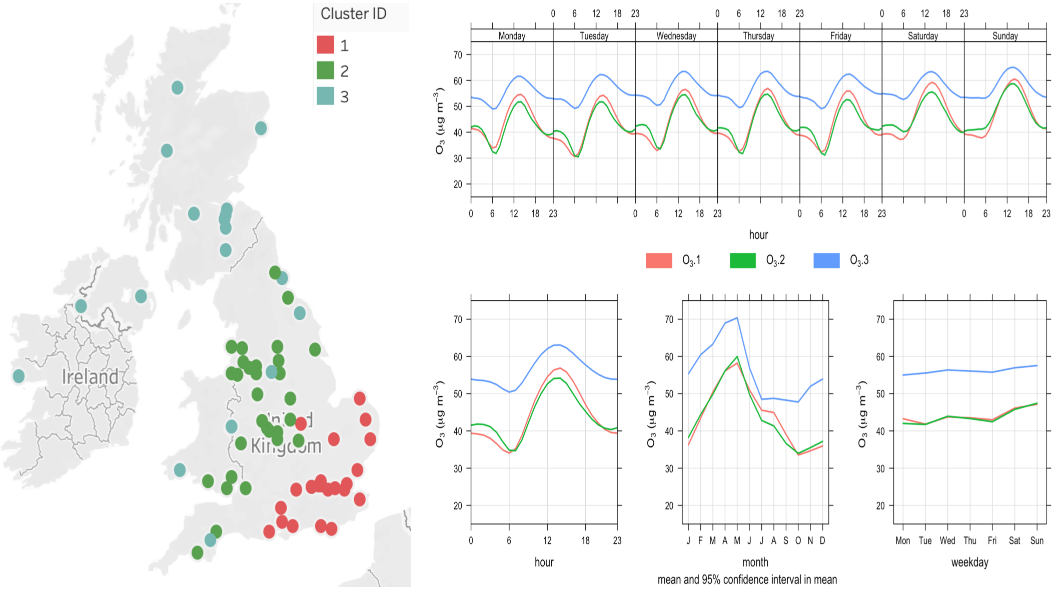


Figure 7: Monthly average concentrations of the 3 cluster's PM_{10} centroids.

in the North/West and lower in the Center and the South/East (as shown in the top plot). We can see that the centroid of cluster 3 (light blue) that is located in the North/West has the highest concentrations among all other clusters centroids. However, cluster 1 and 2 are almost identical and have low concentrations compared to cluster 3. Concentrations appear higher at the end of the week (Friday-Sunday) and during the early afternoon (as shown in the bottom right plot). We can also see there are higher concentrations during March, April, and May compared to the rest of the year, hence O_3 shows dissimilar behaviour to the particulate matter (as shown the bottom middle plot). Fig.9 shows the monthly average concentrations for each cluster centroid. According to this figure there is some seasonality during these years with peaks occurring in late Spring.

These spatial and temporal distributions are consistent with the UK being a net sink of surface O_3 due to emissions of NO_x and dry deposition to the surface [43]. Tropospheric background O_3 peaks in the spring due to photochemical production and exchange with stratosphere. This is mainly imported into the UK in the prevailing westerly air flow. Greater NO emissions in the south east and during week days and rush hours reduce the surface concentrations of O_3.



((a)) O_3 clusters.

((b)) Time variations of cluster's centroids.

Figure 8: (a) Geographical distribution of stations that measure O_3 with the colour coded clusters obtained using the basic k-means algorithm, (b) Time variations of the 3 cluster's O_3 centroids.

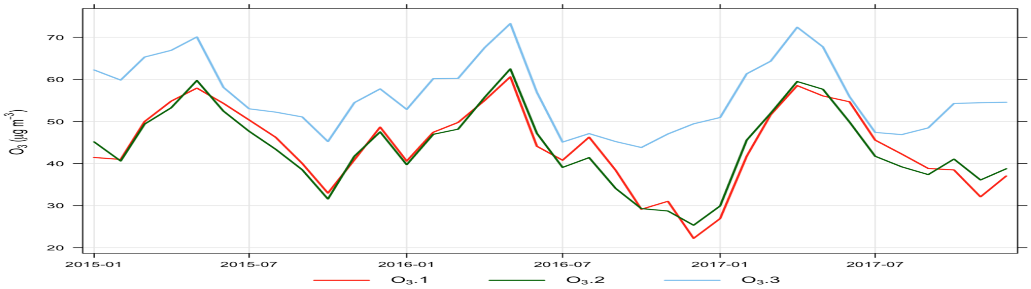
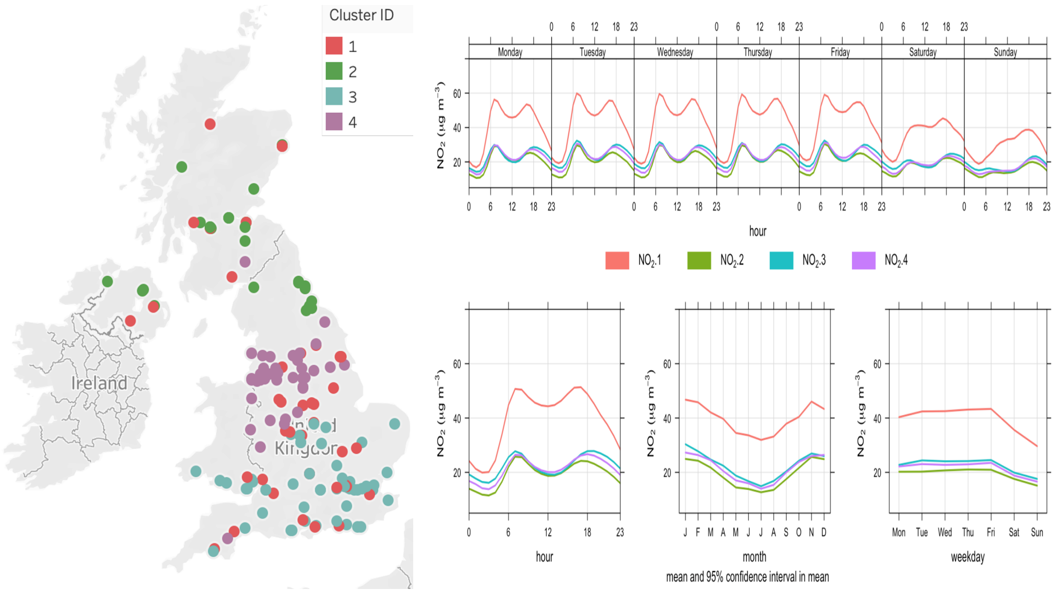


Figure 9: Monthly average concentrations of the 3 cluster's O_3 centroids.

5.1.4. NO_2 Clustering Results

The total number of stations that measure NO_2 is 157 stations, so this is the most measured pollutant. The map in Fig.10(a), shows the clustering of these stations. There are three clusters located roughly around the North (cluster 2, Green), Center (cluster 4, Purple) and South (cluster 3, light blue) and the fourth cluster (cluster 1, red) that is spread all over the other clusters hence NO_2 does not show the same neat geographical division as other pollutants. The red cluster has the highest concentrations among all other centroids as shown in Fig.10(b). This cluster includes 95% traffic urban stations, that are located near to traffic (roads, motorways, highways), and the pollution level at these stations is determined predominantly by the emissions from nearby traffic, and 5% background urban stations that are located in the big and crowded cites such as Greater London, Nottingham, etc. Since NO_2 is the main traffic related air pollutant and it has a lifetime of just minutes to hours, it is not unsurprising that these sites form a cluster and also that the centroid concentrations are higher than those of the other clusters.

The centroids of the other 3 clusters are however very similar. Cluster 3 (light blue) located in the South has slightly higher NO_2 concentrations followed by cluster 4 (purple) at the Center, then the cluster at the North (green) has the lowest concentrations. There are lower concentrations for all clusters on weekend days and there are peaks during the rush-hours (around 7 am and 6 pm) as shown on the left bottom plot. There are also lower concentrations for all clusters during the summer months (June, July, August) as shown on the middle bottom plot. Fig.11 shows the monthly average concentrations of each cluster centroid again exemplifying how the red cluster is very different to the others although it shows similar seasonality.



((a)) NO_2 clusters.

((b)) Time variations of cluster's centroids.

Figure 10: (a) Geographical distribution of stations that measure NO_2 with the colour coded clusters obtained using the basic k-means algorithm, (b) Time variations of the 4 cluster's NO_2 centroids.

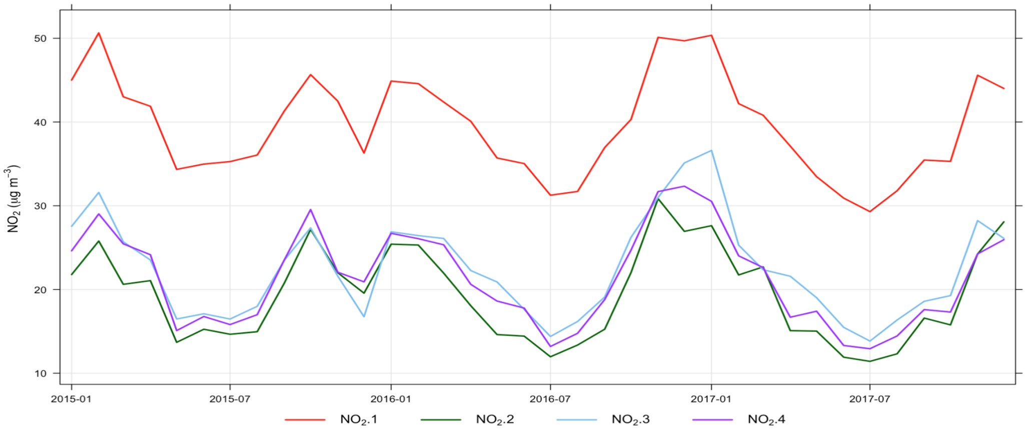


Figure 11: Monthly average concentrations of the 4 cluster's NO_2 centroids.

5.1.5. Univariate TS clustering evaluation

We evaluated the clustering solutions produced for each pollutant based on CVIs for the clusters first. Then we evaluated how good these clusters are for the purposes of imputing the pollutants with our proposed imputation methods.

Table.1 shows the comparison of these clusters using the CVIs. We used the Dunn Index (DI) to measure cluster compactness and separation, Average Silhouette Width (ASW) to measures how close each point in one cluster is to points in the neighbouring clusters, and the connectivity measure (Conn) to reflect how connected objects are within the clusters.

Table 1: Comparing the k-means clusters for each pollutant using the Cluster Validity Indexes (CVI) in experiment 1.

Measure	Criteria	$PM_{2.5}$	PM_{10}	O_3	NO_2
Optimal number of cluster (k)		4	3	3	4
Average Silhouette Width (ASW)	Maximised	**0.235**	0.183	0.227	0.129
Dunn index (DI)	Maximised	0.911	**0.990**	0.761	0.871
Connectivity ($Conn$)	Minimised	36.033	**27.812**	28.962	102.088

From this table, we find that PM10 produces the best clustering solution based on DI and the best connectivity compared to the other pollutants, so the stations clustered together are similar to one another yet dissimilar from stations in other clusters. However, the clustering solution obtained for PM2.5 has the maximum ASW because the number of the cluster is higher. It has similar DI index.

On the other hand, to evaluate how good these clusters are for imputing the pollutants, we compare the RMSE of our imputation methods using the clustering solutions for individual pollutants. In Table.2, the method that gives the lowest RMSE is (CA+ENV) for NO_2 and O_3, and (CA) for $PM_{2.5}$ and PM_{10}. That indicates that NO_2 and O_3 concentrations change from a location to another based on the environmental type, for example the stations that are located at the roadside have higher concentrations of NO_2 than those at the rural background. However, $PM_{2.5}$ and PM_{10} have more regional patterns, and wider distribution. If we compared these values, we can see that the lowest error average is associated with the imputation of the PM_{10} and $PM_{2.5}$, this support our previous evaluation of the clustering quality.

5.2. Experiment 2 and 3: MVTS clustering based on the fused similarity

In these experiments, we use the basic k-means algorithm to cluster the fused distance matrix we calculated in Sec.3. This is to investigate if a clustering that takes into account all of the pollutants measured at a particular

Table 2: The average RMSE and its standard deviation (Std) using the basic k-means clustering algorithm in experiment 1.

Imputation Method	NO_2		O_3		$PM_{2.5}$		PM_{10}	
	RMSE	Std	RMSE	Std	RMSE	Std	RMSE	Std
CA	15.037	7.260	14.877	4.169	**5.728**	1.524	**8.312**	2.750
CA+ENV	**14.095**	7.051	**14.500**	3.885	6.147	1.779	8.367	2.739

station and their similarity may give a better understanding of the patterns of concentration and also better imputation results.

Fig.12 shows the geographical distribution of clustering the stations based on the basic k-mean in experiment 2 where we include all the stations in the dataset. From the geographical distribution of this clustering solution (Fig.12), there are three clusters located in three geographical areas (cluster 2/North (green), cluster 1/South East (red), and cluster 3/South West (light blue)). These clusters are well separated, however there are some stations from cluster 1 (red), that appear within other cluster's geographical areas.

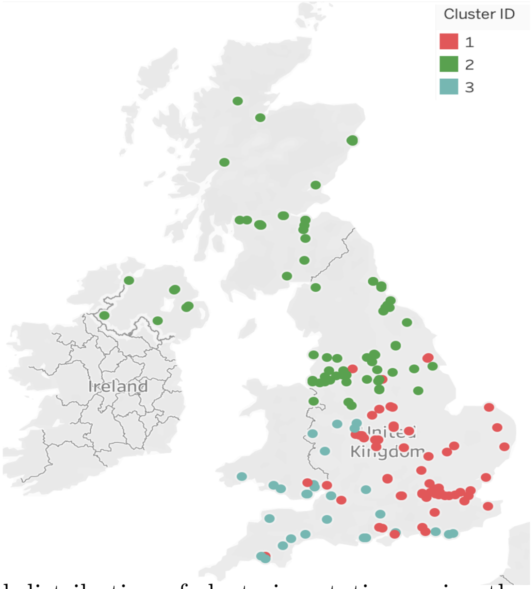


Figure 12: Geographical distribution of clustering stations using the basic k-means algorithm experiment 2.

In experiment 3, we exclude 43 stations from the clustering process as these stations measure NO_2 only and including them in the FDM may negatively affect the clustering process. What we do this time is to allocated these stations to clusters after clusters are constructed. The allocation is based on their partial similarity of NO_2 to the cluster centroids.

Fig.13 shows the geographical distribution of stations for experiment 3, noting that this time we have four clusters (i.e. (cluster 1, (red), (cluster 2, green), and (cluster 3, light blue)). As we can see there is good geographical distribution with the four clusters located in the North, central, South East, and South West. These clusters are well separated; in fact better than those in experiment 2, because stations that only measure NO_2 disrupt the geographical connectivity of the clusters. This is because the sites that only measure NO_2 are usually sited in areas where there are concerns about compliance with NO_2 air quality standards which is normally close to sources such as roads.

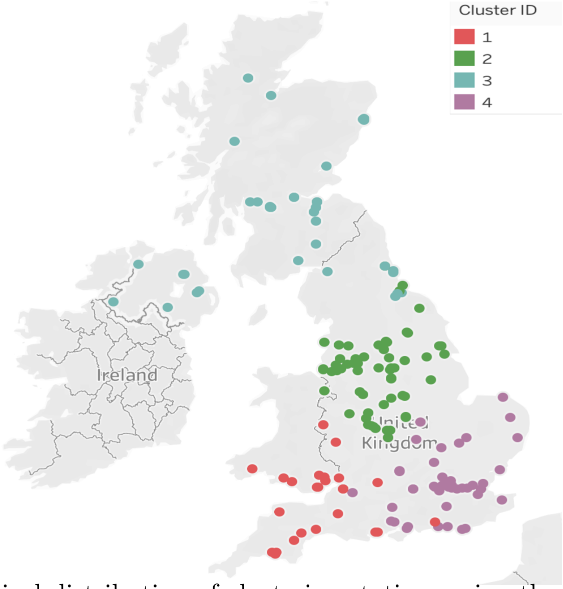


Figure 13: Geographical distribution of clustering stations using the basic k-means algorithm experiment 3.

5.2.1. Clustering Evaluation

We compared the basic k-means clustering solutions from experiment 2 and 3 based on the CVI to select the best clustering solution, i.e. one that is more compact and well separated. Table.3 shows the comparison of these indices. Then, we imputed the missing pollutants using our imputation methods and compared the imputed with the real TS using RMSE and its standard deviation (Std) as shown in Table.4.

Table.3 shows that the clustering solution from the third experiment is better than the one from experiment 2 with a number of these indices including: Within Clusters Sum of Squares (WCSS) measuring the variability within each cluster, Between Clusters Sum of Squares (BCSS) measuring variability between the centroids of the clusters; and it also achieved the highest value with the Silhouette Width (ASW) measuring how close each point in one cluster is to points in the neighbouring clusters.

On the other hand, we evaluated these clustering solutions in terms of missing pollutants imputation using the RMSE, as shown in Table.4, and we compared the clustering imputation methods based on the basic k-means clustering algorithm derived from these experiments. Looking at the second experiment only, at the top of Table.4, we find that using CA+ENV gives the lowest RMSE for NO_2 and O_3 with (13.947, 14.733, respectively). However, using CA method gives the lowest RMSE (5.234, 8.247) for $PM_{2.5}$ and PM_{10} respectively. This is consistent with the single pollutant imputation from the first experiment (i.e. experiment 1 in Sec. 5.1).

For the third experiment, at the bottom of Table.4, the result for the best imputation methods for each pollutant agreed with experiment 2. Importantly, using this clustering solution to impute the pollutants gave the lowest average error for all the pollutants except NO_2. This indicates that it is a good clustering solution and helpful for imputation. In the next section, we discuss and compare these results to the results from other experiments in more details.

Table 3: Cluster Validity Indexes for clustering solutions in experiment 2 and 3 (highlighted cells represent better results in the comparison between experimenter 2 and 3).

Measure	Criteria	Basic k-means (Exp. 2)	Basic k-means (Exp. 3)
Optimal number of cluster (k)		3	4
Within Clusters Sum of Squares ($WCSS$)	Minimised	0.3409	**0.3251**
Between Clusters Sum of Squares ($BCSS$)	Maximised	0.4248	**0.4263**
Average Silhouette Width (ASW)	Maximised	0.1240	**0.1351**
Dunn index (DI)	Maximised	**0.1291**	0.1037
Connectivity ($Conn$)	Minimised	**69.1968**	91.8095

5.2.2. Analysis of pollutant concentrations in each cluster

Since the third experiment gives the best clustering solution in terms of the clustering quality and also achieves the lowest RMSE in the missing pollutants imputation, we analyse the time variation for all the pollutants in this solution in order to compared the cluster centroids. Fig.14 and Fig.15 show the time variation for PM_{10}, $PM_{2.5}$, NO_2, and O_3 respectively, colour

Table 4: The average RMSE and its standard deviation using the basic k-means clustering algorithm in experiments 2 and 3.

Imputation Method	NO_2		O_3		$PM_{2.5}$		PM_{10}	
	RMSE	Std	RMSE	Std	RMSE	Std	RMSE	Std
			Experiment 2					
CA	15.805	8.107	15.211	3.526	5.234	1.253	8.247	2.720
CA+ENV	**13.947**	7.299	14.733	3.469	5.427	1.226	8.283	2.650
			Experiment 3					
CA	16.024	8.443	14.963	4.387	**4.986**	1.155	**7.943**	2.775
CA+ENV	13.965	7.355	**14.265**	3.882	5.332	1.197	8.284	2.751

coded in these figures according to the clustering results in Fig. 13 (i.e. (cluster 1, (red), (cluster 2, green), (cluster 3, light blue), and (cluster 4, purple)).

We will analyse the time variations from these figures based on the four cluster centroids obtained from experiment 3 results, as shown in Fig. 13.

The centroid of cluster 1 (red), located in the South West, has high concentrations of PM_{10}, $PM_{2.5}$ (as in Fig.14) and the highest concentrations of O_3 among all other cluster centroids, but has the lowest concentrations of NO_2 (as in Fig.15).

The centroid of cluster 2 (green), located in the center of the UK, has an average concentration of all the pollutants compared to other centroids, as clearly shown in these figures.

The centroid of cluster 3 (light blue), located in the North has the lowest concentrations of PM_{10}, $PM_{2.5}$ as shown in Fig.14. Very high concentrations of O_3 as shown in the top plot of Fig.15, while it has a low to average NO_2 concentrations comparing to other clusters.

Finally, the centroid of cluster 4 (purple), located in the South East, has the highest concentrations of PM_{10}, $PM_{2.5}$ as shown in Fig.14. Also, it has the highest concentrations of NO_2, but the lowest concentrations of O_3 as shown in Fig.15.

From these figures, if we compare the time variation of pollutants concentration based on the location of the clusters from the MVTS clustering (i.e. experiment 3) with individual pollutants clustering (Sec. 5.1), we can see that the pollutants concentration in these locations similar to one another. For example, the UK North region has the lowest concentrations of PM_{10}, $PM_{2.5}$, NO_2 but highest concentrations of O_3, while the opposite is true for South regions. Which confirmed the ability of the MVTS clustering to reflect and understand multi pollutants behaviour.

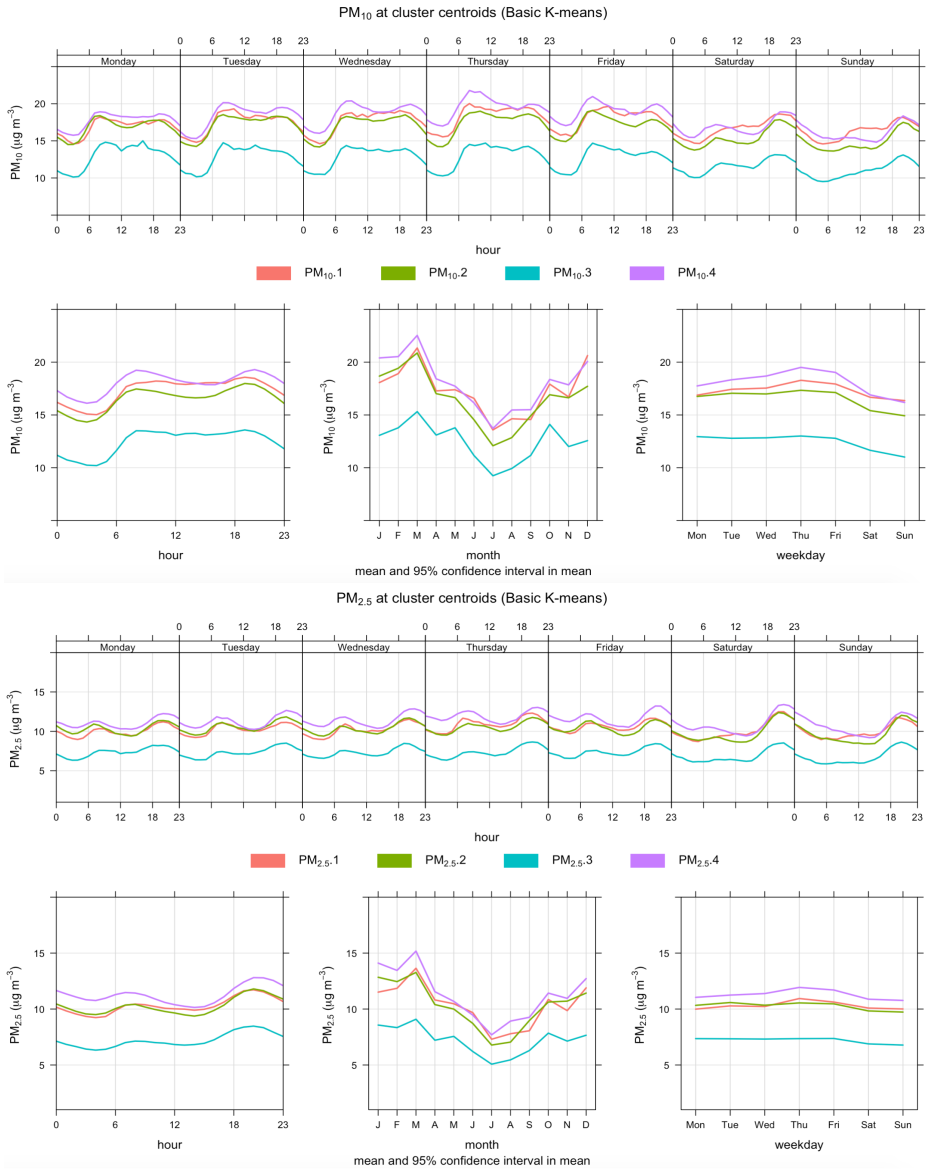


Figure 14: Time variation of basic k-means cluster centroids of PM_{10} (top) and $PM_{2.5}$ (bottom) concentrations.

Next, we show an example of our imputed TS compared to the real TS for each pollutant using the selected imputation methods. For this comparison,

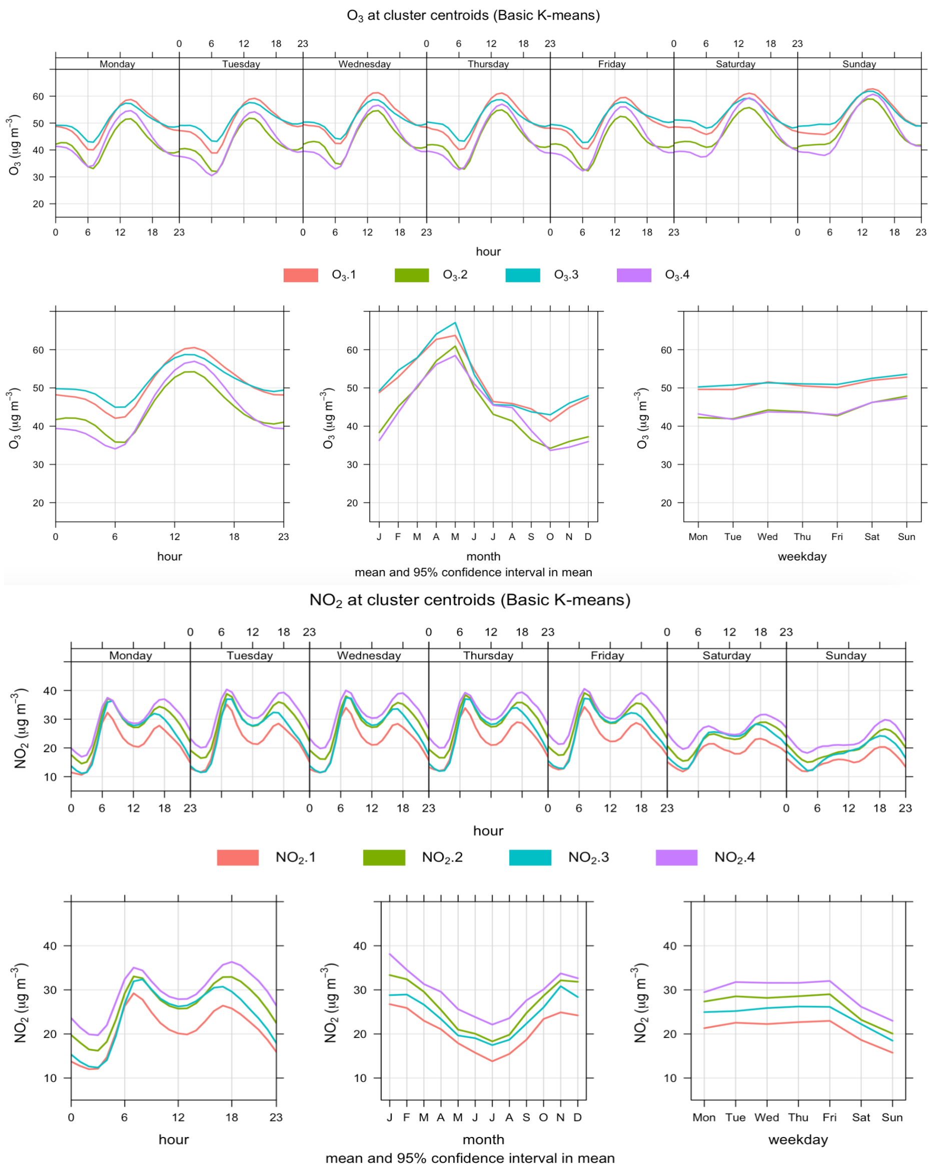
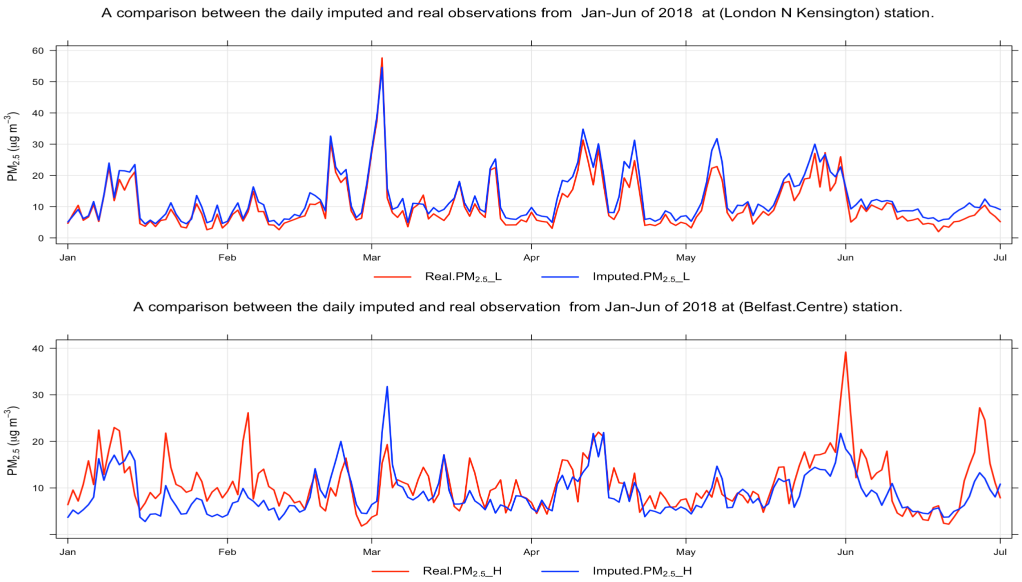


Figure 15: Time variation of basic k-means cluster centroids of O_3 (top) and NO_2 (bottom) concentrations.

we selected two stations that associated with the highest and lowest RMSE for each pollutant. We compare the daily mean of the imputed and the real TS for the period of six months (Jan-Jun) of the year of 2018. We only include an examples of $PM_{2.5}$ and O_3 due to the limited space in this paper.

Fig.16 is an example of the $PM_{2.5}$ imputed using the CA method and real TS at 'London N Kensington' (lowest RMSE) and 'Belfast Centre' stations (highest RMSE). The imputed TS for London N. Kensington has slightly higher values than the real TS, while the opposite is true for Belfast Center, where the imputed TS is slightly lower than the real TS. Again, the trends are very similar and they represent valid imputations.

Fig.17 shows a comparison between the imputed O_3 TS using the CA+ENV method and real TS at 'London N Kensington' in the top (lowest RMSE), and 'London Hillington' in the bottom (highest RMSE). We can see that in the second plot for 'London Hillington' site the variation between the imputed and the real TS is slightly higher compared to the previous example although again the trend is good.

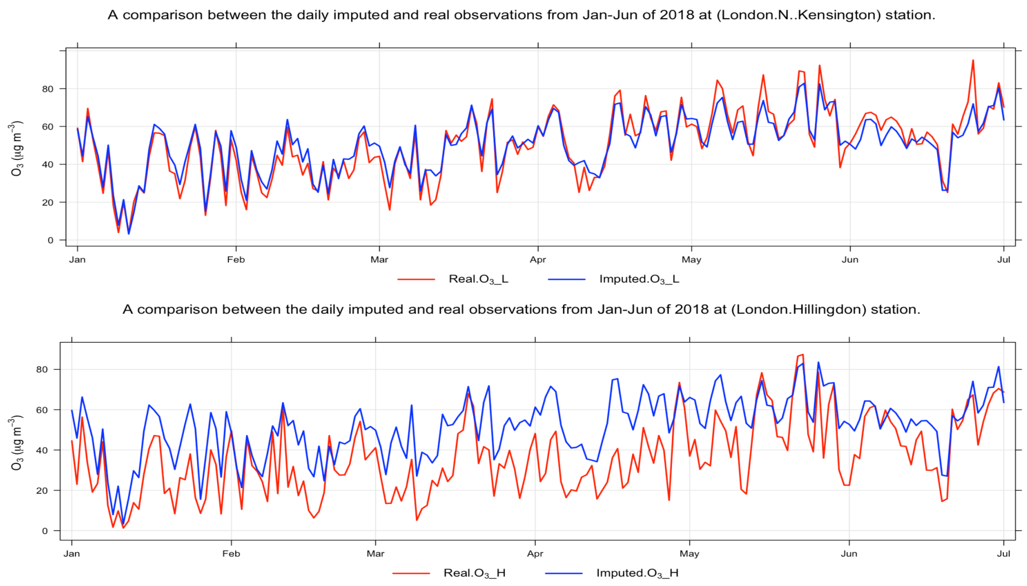

A comparison between the daily imputed and real observations from Jan-Jun of 2018 at (London N Kensington) station.

A comparison between the daily imputed and real observation from Jan-Jun of 2018 at (Belfast.Centre) station.

Figure 16: Imputed and real TS comparison for $PM_{2.5}$ with lowest RMSE (top) and the highest RMSE (bottom).

6. Discussion

Our analysis showed that a basic k-means algorithm with fused distances results in geographical patterns that are consistent with our understanding of sources and lifetimes of these pollutants, as explained in Sec. 5.2.


A comparison between the daily imputed and real observations from Jan-Jun of 2018 at (London.N..Kensington) station.

A comparison between the daily imputed and real observations from Jan-Jun of 2018 at (London.Hillingdon) station.

Figure 17: Imputed and real TS comparison for O_3 with lowest RMSE (top) and the highest RMSE (bottom).

We found that using the basic k-means with the MVTS clustering and fused similarity in the second and third experiments gave a clear geographical correlation between the stations. Our results of analysing the centroids of the clusters identify similar pollutant concentrations levels and geographical distribution to the results in one of the most recent reports from Centreforcities [10]. This report focuses on analysing the concentrations level across the UK for NO_2 and $PM_{2.5}$, and explores the fact that NO_2 and O_3 have an anti-correlation [14], hence their concentrations in a region are at opposite ends of the scale. This is corroborated in our clustering results.

In terms of imputation, using the basic k-means with the defined imputation methods helped to impute/estimate plausible concentrations of multiple pollutants at a station. Although the best imputation method with lowest error average may be different from one pollutant to another, all experiments agreed that using CA+ENV to impute NO_2 and O_3 gave the lowest error average (RMSE), and using CA is better for the imputation of $PM_{2.5}$, and PM_{10} concentrations due to the behaviour of each pollutant.

We also observe that univariate and MVTS clustering analysis lead to different clustering results. Comparing the error average of these methods from the univariate TS clustering (experiment 1) to the MVTS clustering (experiment 2 and 3) showed that the error average using CA+ENV for NO_2 imputation decreased by (0.15, 0.13) in the second and third experiments using MVTS clustering compared to using the univariate TS clustering. Even though, the error averages increased for pollutant imputation in the second experiment for O_3 and NO_2, they decreased in the third experiment by 0.2, 0.7, 0.4 for O_3, $PM_{2.5}$, and PM_{10} respectively. This indicates that using NO_2 data from NO_2 only sites has a detrimental effect in the imputation of O_3 and PM.

Furthermore, MVTS clustering enables imputation even when no measurement is available for a given pollutant since the station can be allocated to a cluster based on the value of the other pollutants measured as we demonstrated with the stations that measure only NO_2 in the third experiment.

7. Conclusion

In this work, we proposed a model to impute missing pollutant (whole TS) through a MVTS clustering approach. We conducted multiple experiments to evaluate the effectiveness of our approach. We compared the proposed approach (i.e. the MVTS clustering using the fused similarity that combines the SBD for all the pollutants) with the univariate TS clustering using SBD for individual pollutants. These two approaches are compared in term of the clustering and the imputation quality.

We found that using the basic k-means with the fused distance performs better than other clustering algorithms for imputation and gives very compact geographical clustering. This indicates that using the fused distance to measure the similarity between the pollutants helped us to solve some of the uncertainty problems associated with missing pollutant values and enabled us to discover multiple patterns of pollutant behaviour that are manifested in different areas around the UK. This knowledge can then be used to understand the behaviour of the pollutants that indicate the air pollution level.

In future work, we will apply imputation to evaluate the AURN network and to help identify where the calculated DAQI might have differed if more data (measured or imputed) were available. We can also improve imputation methods by considering environmental type of the stations further, and information about the stations like station altitude and locations. Also, we may consider the correlation between the pollutants.

We intend to compare our applied MVTS clustering technique with some ensemble clustering methods that are based on univariate TS clustering, then apply our imputation methods to see the performance of our MVTS clustering technique compared to traditional ones.

References

[1] J. J. Carbajal-Hernández, L. P. Sánchez-Fernández, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, Assessment and prediction of air quality using fuzzy logic and autoregressive models, Atmospheric Environment 60 (2012) 37–50.

[2] G. Di Bello, V. Lapenna, M. Macchiato, C. Satriano, C. Serio, V. Tramutoli, et al., Parametric time series analysis of geoelectrical signals: an application to earthquake forecasting in southern italy (1996).

[3] S. Seto, W. Zhang, Y. Zhou, Multivariate time series classification using dynamic time warping template selection for human activity recognition, in: 2015 IEEE Symposium Series on Computational Intelligence, IEEE, 2015, pp. 1399–1406.

[4] P. D'Urso, L. De Giovanni, R. Massari, Robust fuzzy clustering of multivariate time trajectories, International Journal of Approximate Reasoning 99 (2018) 12–38.

[5] A. Kurt, A. B. Oktay, Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks, Expert Systems with Applications 37 (12) (2010) 7986–7992.

[6] DEFRA air information resource, http://uk-air.defra.gov.uk.

[7] P. Holnicki, Z. Nahorski, Emission data uncertainty in urban air quality modeling—case study, Environmental Modeling & Assessment 20 (6) (2015) 583–597.

[8] Public Health sources and effects of pm2.5, https://laqm.defra.gov.uk/public-health/pm25.html.

[9] National Statistics concentrations of particulate matter pm_{10} and pm_{25}, https://www.gov.uk/government/publications/air-quality-statistics/concentrations-of-particulate-matter-pm10-and-pm25.

[10] Centreforcities cities outlook 2020, https://www.centreforcities.org/publication/cities-outlook-2020/.

[11] F. M. Diaz, M. A. H. Khan, B. Shallcross, E. D. Shallcross, U. Vogt, D. E. Shallcross, Ozone trends in the united kingdom over the last 30 years, Atmosphere 11 (5) (2020) 534.

[12] G. M. Mazzuca, X. Ren, C. P. Loughner, M. Estes, J. H. Crawford, K. E. Pickering, A. J. Weinheimer, R. R. Dickerson, Ozone production and its sensitivity to nox and vocs: Results from the discover-aq field experiment, houston 2013 (2016).

[13] M. A. H. Khan, W. C. Morris, M. Galloway, B. M. A. Shallcross, C. J. Percival, D. E. Shallcross, An estimation of the levels of stabilized criegee intermediates in the uk urban and rural atmosphere using the steady-state approximation and the potential effects of these intermediates on tropospheric oxidation cycles, International journal of chemical kinetics 49 (8) (2017) 611–621.

[14] C. Lin, X. Feng, M. R. Heal, Temporal persistence of intra-urban spatial contrasts in ambient no2, o3 and ox in edinburgh, uk, Atmospheric Pollution Research 7 (4) (2016) 734–741.

[15] S. Lee, G. Wolters, L. Grant, T. Schneider, Atmospheric ozone research and its policy implications, Elsevier, 1989.

[16] UK-AIR ozone in the united kingdom, https://uk-air.defra.gov.uk/library/assets/documents/reports/aqeg/aqeg-ozone-report.pdf (2009).

[17] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA, 1967, pp. 281–297.

[18] C. Guo, H. Jia, N. Zhang, Time series clustering based on ica for stock data analysis, in: 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, IEEE, 2008, pp. 1–4.

[19] X. Wang, K. Smith, R. Hyndman, Characteristic-based clustering for time series data, Data mining and knowledge Discovery 13 (3) (2006) 335–364.

[20] H. Li, M. Wei, Fuzzy clustering based on feature weights for multivariate time series, Knowledge-Based Systems 197 (2020) 105907.

[21] T. W. Liao, Clustering of time series data—a survey, Pattern recognition 38 (11) (2005) 1857–1874.

[22] P.-Y. Zhou, K. C. Chan, A model-based multivariate time series clustering algorithm, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2014, pp. 805–817.

[23] H. Li, Multivariate time series clustering based on common principal component analysis, Neurocomputing 349 (2019) 239–247.

[24] C. H. Fontes, H. Budman, A hybrid clustering approach for multivariate time series–a case study applied to failure analysis in a gas turbine, ISA transactions 71 (2017) 513–529.

[25] K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, R. Jenssen, Time series cluster kernel for learning similarities between multivariate time series with missing data, Pattern Recognition 76 (2018) 569–581.

[26] E. H. Wu, L. Philip, Independent component analysis for clustering multivariate time series data, in: International Conference on Advanced Data Mining and Applications, Springer, 2005, pp. 474–482.

[27] H. Li, T. Du, Multivariate time-series clustering based on component relationship networks, Expert Systems with Applications 173 (2021) 114649.

[28] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, P. Solenberger, et al., A multivariate technique for multiply imputing missing values using a sequence of regression models, Survey methodology 27 (1) (2001) 85–96.

[29] W. Alahamade, I. Lake, C. E. Reeves, B. De La Iglesia, Clustering imputation for air pollution data, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2020, pp. 585–597.

[30] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE transactions on acoustics, speech, and signal processing 26 (1) (1978) 43–49.

[31] J. Paparrizos, L. Gravano, k-shape: Efficient and accurate clustering of time series, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 1855–1870.

[32] L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, Vol. 344, John Wiley & Sons, 2009.

[33] M. Žitnik, B. Zupan, Data fusion by matrix factorization, IEEE transactions on pattern analysis and machine intelligence 37 (1) (2014) 41–53.

[34] A. Mojahed, B. de la Iglesia, An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach, Knowledge and Information Systems 50 (1) (2017) 27–52.

[35] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. PéRez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognition 46 (1) (2013) 243–256.

[36] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence (2) (1979) 224–227.

[37] J. Handl, J. Knowles, An evolutionary approach to multiobjective clustering, IEEE transactions on Evolutionary Computation 11 (1) (2007) 56–76.

[38] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.

[39] J. C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters (1973).

[40] E. Chen, F. Wang, Dynamic clustering using multi-objective evolutionary algorithm, in: International Conference on Computational and Information Science, Springer, 2005, pp. 73–80.

[41] Q. H. Nguyen, V. J. Rayward-Smith, Internal quality measures for clustering in metric spaces, International Journal of Business Intelligence and Data Mining 3 (1) (2008) 4–29.

[42] A. GROUP, et al., Fine particulate matter (pm 2.5) in the united kingdom, Department for Environment, Food and Rural Affairs, London (2012).

[43] A. GROUP, Aqeg: Ozone in the united kingdom. fifth report of the air quality expert group, Department for Environment, Food and Rural Affairs, London (2009).