An Interactive Clustering-Based Visualization Tool for Air Quality Data Analysis

Mahsa Ashouri1*, Frederick Kin Hing Phoa2, Chun-Houh Chen2, Galit Shmueli3

1 Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-1382, USA
2 Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan
3 Institute of Service Science, National Tsing Hua University, Hsinchu 30013, Taiwan

ABSTRACT

Examining PM$_{2.5}$ (atmospheric particulate matter with a maximum diameter of 2.5 micrometers), seasonal patterns is an important research area for environmental scientists. An improved understanding of PM$_{2.5}$ seasonal patterns can help environmental protection agencies (EPAs) make decisions and develop complex models for controlling the concentration of PM$_{2.5}$ in different regions. This work proposes an R Shiny App web-based interactive tool, namely a “model-based time series clustering” (MTSC) tool, for clustering PM$_{2.5}$ time series using spatial and population variables and their temporal features, like seasonality. Our tool allows stakeholders to visualize important characteristics of PM$_{2.5}$ time series, including temporal patterns and missing values, and cluster series by attribute groupings. We apply the MTSC tool to cluster Taiwan’s PM$_{2.5}$ time series based on air quality zones and types of monitoring stations. The tool clusters the series into four clusters that reveal several phenomena, including an improvement in Taiwan’s air quality since 2017 in all regions, although at varying rates, an increasing pattern of PM$_{2.5}$ concentration when moving from northern towards southern regions, winter/summer seasonal patterns that are more pronounced in certain types of areas (e.g., industrial), and unusual behavior in the southernmost region. The tool provides cluster-specific quantitative figures, like seasonal variations in PM$_{2.5}$ concentration in different air quality zones of Taiwan, and identifies, for example, an annual peak in early January and February (maximum value around 120 µg m$^{-3}$). Our analysis identifies a region in southernmost Taiwan as different from other zones that are currently grouped together with it by Taiwan EPA (TEPA), and a northern region that behaves differently from its TEPA grouping. All these cluster-based insights help EPA experts implement short-term zone-specific air quality policies (e.g., fireworks and traffic regulations, school closures) as well as longer-term decision-making (e.g., transport control stations, fuel permits, old vehicle replacement, fuel type).

Keywords: Time series, Clustering, Web-based tool, Air quality, Environmental protection agencies

1 INTRODUCTION

Air pollution, mainly caused by urban and industrial developments (Chen et al., 2019), has become a serious worldwide concern for experts and governments. Among all air pollution factors, atmospheric particulate matter (PM) with a diameter of less than 2.5 micrometers, known as PM$_{2.5}$, is especially important due to its effects on human health and the environment (Chen et al., 2020). Besides short-term health effects such as coughing and sneezing, PM$_{2.5}$ causes long-term health problems like asthma and lung cancers (Xing et al., 2016). PM$_{2.5}$ also leads to environmental issues, including reduced visibility (haze) and ecosystem diversity (U.S. EPA, 2020).

Analyzing PM$_{2.5}$ data helps environmental protection agencies (EPAs) tackle several decision-making problems. One important example is pollution control policy assessment. Peng et al. (2022) used kernel density estimation to describe the characteristics of the daily PM$_{2.5}$ data in China’s cities. The results helped in evaluating the effectiveness of pollution control policies implemented in major cities and estimated the heterogeneity of the PM$_{2.5}$ series among cities. Likewise, Yang et al. (2019)
evaluated China’s current air quality index system and showed the efficacy of pollution control policies in different regions.

Exploring seasonal variation across large areas also motivates researchers to analyze PM$_{2.5}$ datasets. Describing spatial and temporal variabilities in PM$_{2.5}$ datasets is the first step toward building sophisticated conceptual models of the causes of PM$_{2.5}$ concentrations. This approach can also differentiate the varying contributions of local and regional sources of PM$_{2.5}$ concentrations (Russell et al., 2004). Moreover, exploring the chemical composition of PM$_{2.5}$ during the seasons of the year is crucial for developing effective public pollution control strategies (Zhang et al., 2015). For example, Russell et al. (2004) examined the daily, seasonal, and spatial trends of PM$_{2.5}$ concentrations, compositions, and size distributions in Southeast Texas. Similarly, Zhao et al. (2018) investigated the weekly and yearly seasonal patterns of PM$_{2.5}$ concentrations over the US using the Prophet time series forecasting model. Identifying these temporal trends helps to accurately forecast PM$_{2.5}$ concentrations and mitigate human exposure.

1.1 Clustering Methods for the Analysis of PM$_{2.5}$ Data

In the literature, various clustering approaches have been used to analyze air pollution, specifically PM$_{2.5}$. Blazquez and Montero (2020) performed a spatial and aspatial clustering analysis of PM$_{2.5}$ pollutants collected in a mobile campaign in Chile. Their analytical results imply that combining spatial and aspatial clustering methods produces high-quality partitions when considering spatial information. Zhang and Yang (2022) employed clustering methods on PM$_{2.5}$ pollution to define joint control regions for addressing severe regional air pollution in China. In sum, clustering approaches facilitate understanding and formulating PM$_{2.5}$ spatiotemporal aggregation methods for designing joint control measures among areas.

Clustering is also a practical exploratory tool for exploring the underlying patterns in PM$_{2.5}$ data (e.g., trend and seasonal patterns). Ab. Rahman et al. (2022) studied the overall trend of PM$_{2.5}$ from 65 monitoring stations in Malaysia using spatial classification cluster analysis based on the Agglomerative Hierarchical Cluster (AHC) method. Their clustering results explore the monthly and annual variations of PM$_{2.5}$ and their relationship with other air pollutants and meteorological factors. Additionally, Liu et al. (2016) applied two clustering methods (AHC and K-means) to one-year data of PM$_{2.5}$ in Beijing with a calendar visualization of PM$_{2.5}$ concentrations. These findings can play a role in formulating mitigation measures and policies.

1.2 The Air Pollution Problem in Taiwan

PM$_{2.5}$ is a major public concern in Taiwan, particularly due to haze events in different regions. Research by Wang et al. (2021) has also demonstrated a relationship between PM$_{2.5}$ and respiratory diseases in Taiwan. In response, Taiwan’s government has recently introduced several new rules to control PM$_{2.5}$ emissions, such as air pollution control devices and strict vehicle emission standards (Cheng and Hsu, 2019). Researchers are now actively investigating PM$_{2.5}$ emissions and pollution in Taiwan. For example, Cheng and Hsu (2019) examined variations in meteorological conditions caused by regional climate change and studied the implications of these variations on PM$_{2.5}$ in Taiwan through long-term trend analysis. Lee et al. (2020) applied a gradient-boosting-based machine learning approach to predict PM$_{2.5}$ concentration in Taiwan using the PM$_{2.5}$ data collected from the Taiwan air quality stations and the hourly weather data from the weather stations.

In addition, some studies rely on clustering analysis of PM$_{2.5}$ data in Taiwan. Su et al. (2020) analyzed the PM$_{2.5}$ index using clustering and studied its association with synoptic weather patterns. They applied a hierarchical clustering approach resulting in five clusters where three corresponded to severe air pollution. In Chuang et al. (2018), PM$_{2.5}$ data were clustered (using k-means clustering) into long-range transport of major and minor industrial emissions. They aimed to evaluate the associations between the bioreactivity of PM$_{2.5}$ in vitro and emission sources in the vicinity of a petrochemical complex in Taiwan.

In this work, we propose an interactive tool that clusters the PM$_{2.5}$ dataset in Taiwan using temporal patterns and other external variables, including spatial and population features. We explore the seasonal (monthly and daily) patterns of the series in each cluster. Taiwan EPA officials can apply this tool and leverage its results for improved decision-making in complex environmental problems both for short-term zone-specific interventions such as fireworks and traffic control.
during holidays, and for long-term planning involving fuel permits, fuel types change, aging motorcycle replacements, and instituting more transport controls (Executive Yuan, 2020). Further, our tool allows comparing data-driven groupings of zones with similar PM$_{2.5}$ behaviors to groupings by the EPA, potentially uncovering inconsistencies that can lead to updated EPA grouping. Our exploratory tool also contributes to the Taiwan EPA policy goal of improved monitoring data quality since it can be used on larger and longer PM$_{2.5}$ datasets collected by internet-of-things (IoT) devices.

The following sections explore the Taiwan PM$_{2.5}$ dataset and its cross-sectional features. Then we review the usefulness of interactive tools and web apps for statistical models, followed by a brief explanation of the model used for clustering PM$_{2.5}$ time series. Finally, we discuss the data visualization results.

1.3 Taiwan's Air Quality Zones and Stations

The Taiwan Environmental Protection Administration (TEPA) divides Taiwan into seven major air quality zones based on geographical characteristics and air quality conditions. Fig. 1, modified from Taiwan Environmental Protection Administration (2022), shows these seven major air quality zones, including northern Taiwan (NT—four districts: Keelung, Taipei, New Taipei, and Taoyuan), Chu—Miao (CM—two districts: Hsinchu and Miaoli), central Taiwan (CT—three districts: Taichung, Changhua, and Nantou), Yun—Chia—Nan (YCN—three districts: Yunlin, Chiayi, and Tainan), Kao—Ping (KP —two districts: Kaohsiung and Pingtung), Hua—Tung (HT —two districts: Hualien and Taitung), Yilan (YI—one district: Yilan) (Cheng and Hsu, 2019; Hwang et al., 2017). There are also three minor islands (Matsu (MT), Kinmen (KI), and Magong (MG)) air quality stations that we exclude from this study. We removed these islands from our analysis based on our expert Taiwan EPA collaborator’s suggestion, given that two of the islands are close to the Fujian coast in China, which results in their air quality being significantly influenced by China (Lai and Brimblecombe, 2021).
Table 1. Domain-relevant attributes in the air quality dataset and their categories/values.

<table>
<thead>
<tr>
<th>Domain-relevant attributes</th>
<th>Categories/Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEPA zones</td>
<td>northern Taiwan (NT), Chu-Miao (CM), central Taiwan (CT), Yun-Chia-Nan (YCN),</td>
</tr>
<tr>
<td></td>
<td>Kao-Ping (KP), Hua-Tung (HT), Yilan (YI)</td>
</tr>
<tr>
<td>Station type</td>
<td>ambient (A), background (B), traffic (T), industrial (I), park (P), other (O)</td>
</tr>
<tr>
<td>Metro (underground transportation)</td>
<td>yes (there is a metro in the county), no (there is no metro in the county)</td>
</tr>
<tr>
<td>Population in county</td>
<td>Numerical value (between 13089 and 4018696)</td>
</tr>
<tr>
<td>Administrative Level</td>
<td>city, county, municipality</td>
</tr>
</tbody>
</table>

In addition, TEPA classifies the air quality stations into six categories based on their locations, namely, ambient (A—surface stations), background (B), traffic (T), industrial (I), park (P), and others (O). We use these spatial and station categories in our clustering and visualization process.

1.4 Taiwan PM$_{2.5}$ Dataset

Our dataset, collected from Environmental Protection Administration Executive Yuan (2022), includes the measured PM$_{2.5}$ in different monitoring stations in Taiwan. It consists of 71 daily series (one series per monitoring station) from January 2015 to November 2019 (the length of each series is 1826 periods). Cross-sectional variables are five location-related variables, called domain-relevant attributes, including TEPA zones (7 categories), station type (6 categories), metro (2 categories), population (numeric), and administrative level (3 categories). See Table 1. Note that the population is the estimated population collected at the end of 2019 from the City Population (2021). Each time series in the collection belongs to a subgroup specified by a combination of the above five domain-relevant attributes. We decided to use these domain-relevant attributes and organize the data based on inputs from an EPA decision-maker/expert.

2 CLUSTERING PM$_{2.5}$ TIME SERIES DATA WITH DOMAIN-RELEVANT ATTRIBUTES

The amount of time series data has dramatically increased in the last decade due to the proliferation of technologies for capturing time series data, thus making automated clustering methods extremely important. While existing clustering methods (see Liao (2005)) are typically applied to time series only, data collected by devices often contain not only time series but also cross-sectional attributes about each time series—i.e., domain-relevant attributes.

Devices monitoring information such as road traffic, heart rate, or room temperature usually collect data in the form of time series, which can then be enriched with labels or tags such as the device’s hardware version, operating time zone, installation location, and much more, see Cloud Architecture Center of Google Cloud (2019). For example, stations and devices that collect time series of air quality index measurements can be coupled with information on the location and context of each time series. The combined temporal and cross-sectional information is richer than the individual components and helps analysts and end-users make better decisions.

2.1 Using Web Apps for Visualizing Complex Underlying Models

Many statistical models, machine learning algorithms, and data analyses performed by statisticians and data scientists are implemented in programming software such as R and Python. However, end-users such as managers, who need to understand decision-making results, are not necessarily familiar with the programming or software employed. One way to empower these end-users is by using web-based tools and apps, which let users visually interact with results by manipulating different parameters. Web apps improve the communication of complex scientific methods with management agencies, industry practitioners, and stakeholders, as well as increase the willingness of users to make use of academic research results (Wszola et al., 2017). R Shiny (RStudio Inc., 2013) is a package from RStudio (RStudio Team, 2020) that simplifies the process of building web-based interactive visualization apps.
Interactive tools are commonly used to visualize the results of time series clustering. For example, Cachuclo et al. (2016) created a Shiny app for multivariate time series data that helps users explore and visualize multivariate time series using a biclustering approach. Another example of visualizing time series clustering comes from Andrienko and Andrienko (2015), who introduced an interactive visual embedding of the partition-based clustering of multidimensional spatial time series using an open-source Weka library.

Additionally, interactive tools are considered very efficient in visualizing the air pollution index (PM2.5). For example, Lu et al. (2017) introduced two visualization tools that focus on air quality monitoring data for all major cities in China. Another example is Upadhyya et al. (2020), where the authors proposed an R Shiny-based package that analyzes, visualizes, and produces a spatial map of air quality data collected by specific devices installed on a moving platform. This research explores Taiwan’s PM2.5 seasonal pattern in spatial and population categories.

Using the tree-based clustering algorithm developed by Ashouri et al. (2019), we developed an interactive R Shiny app for visualizing and exploring clustering results of PM2.5 time series data with domain-relevant attributes. Our interactive tool allows end-users and Taiwan EPA experts to explore clustering results by changing parameters affecting the clustering algorithm. Our tool also lets end-users select the number of desired clusters and specify the preferred domain-relevant attributes used in the clustering process. We call this tool the “Model-based Time Series Clustering (MTSC) tool”. Based on our literature search and according to our expert Taiwan EPA collaborator, no such tool yet exists for Taiwan, thus illustrating the importance of our research.

### 2.2 Clustering Time Series with Domain-relevant Attributes Using MOB Trees

In the literature, researchers often use domain-relevant attributes in a post-hoc fashion for interpreting the results of clustering purely time series data (e.g., Dasu et al., 2005 and Jank and Shmueli, 2010, p. 63). Instead, we describe here, in brief, the method proposed by Ashouri et al. (2019, 2022) and validated on several real and simulated datasets, which uses domain-relevant attributes (in addition to the temporal information) to cluster the time series directly. This method is implemented in our web-based tool.

Ashouri et al. (2019)’s clustering algorithm is based on model-based (MOB) trees (Zeileis et al., 2008). The MOB algorithm includes four main steps. First, we fit a parametric model to the data such as a linear regression model, and estimate the model parameters. Second, different variables (“splitting variables”) are used to split the sample into nonoverlapping sub-samples. The model is then re-estimated for each sub-sample created by a splitting variable, and a parameter instability test, such as the Chow test (Chow, 1960), is used to compare the model parameters across the sub-samples from a splitting variable. This is repeated for each splitting variable. The splitting variable with the most instability across sub-samples is selected (or none are selected and the sample is not split). Third, for the selected splitting variable, the algorithm finds the split point (value) that locally optimizes the objective function (e.g., minimizes the sum of squared errors). Finally, these three steps are repeated for each resulting sub-sample. Using MOB therefore requires selecting the parametric model to fit, the splitting variables, and the objective function.

In the Ashouri et al. (2019) approach, the parametric model used is a flexible linear regression model that can capture time series temporal components (trend, seasonality, and autocorrelation); Domain-relevant attributes are used as candidates for splitting the tree nodes (splitting variables).

In other words, each time series is modeled as a linear regression with the following formula:

\[ Y = Trend + Season + AR(p) + error, \]  

where \( AR(p) \) is a weighted average of the time series lags \( \{y_{t-1}, y_{t-2}, ..., y_{t-p}\} \). The order \( p \) can be either specified or equal to the seasonality order based on the data type and domain knowledge. The linear model in Eq. (1) captures the time series trend, seasonality, and autocorrelation by incorporating each of these components into the equation. To be more explicit and using MOB notation (Zeileis et al., 2008), this formulation can be written as:

\[ y_t = \alpha_0 + \alpha_1 t + \beta_1 Season_{t1} + \beta_2 Season_{t2} + \cdots + \beta_{m-1} Season_{(m-1)t} + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + Z_t, \]  

where \( Z_t \) is a vector of domain-relevant attributes.
where $y_t (t=1, 2, ..., T)$ is the value of the series at time $t$; $f(t)$ — the first component in Eq. (1) — is a function of the time index that captures the time series trend (e.g., linear or $f(t) = t$, quadratic or $f(t) = t + t^2$; $\text{Season}_t$ — the second component in Eq. (1) — is a dummy variable taking the value 1 if time $t$ is in season $j$; and $m$ is the number of seasons (e.g., for a monthly time series with the month-of-year seasonality, $m = 12$ and 11 dummy variables). Further, $y_t - j$ — the third component in Eq. (1) — is the $j$th lagged value and domain-based attributes are denoted by $Z_t$, and $Z_1, ..., Z_q$ represent $q$ splitting variables, or domain-relevant attributes.

Three criteria for determining the final number of clusters for the different number of splits of the MOB tree are (1) the mean squared error (MSE) used by MOB to prune the tree, (2) the Bayesian information criterion (BIC) (Schwarz, 1978) or AIC, and (3) the complexity of the tree (the number of terminal nodes). The tree is pruned to the level with the largest decrease in MSE (MSE jump). Final clusters (represented by a regression model in each terminal node) are compared via coefficient plots, which visualize and compare coefficients for each predictor of the linear model within each cluster. Finally, clusters are defined and interpreted using domain-relevant attributes.

3 CLUSTERING PM$_{2.5}$ TIME SERIES WITH DOMAIN-RELEVANT ATTRIBUTES

To compare two sets of parameter settings, our MTSC tool creates two sets of results presented in two tabs. Fig. 2 displays an overview screenshot of our web-based MTSC tool for the Taiwan PM$_{2.5}$ time series. It corresponds to one depth (depth = 3), and different depth choices appear in separate tabs. In Figs. 3–7 we displayed each part of the MTSC tool in separate figures for better visualization. Please note that in order to enhance the visibility of the heatmap, line, and coefficient plots, we have transformed the figures into vertical plots (Figs. 5–8). This adjustment allows for a larger display, ensuring a more detailed and comprehensive visualization. Our MTSC tool shows the mean squared error (MSE), Akaike information criterion (AIC) (Akaike, 1998), and Bayesian information criterion (BIC) (Schwarz, 1978) as well as four charts that provide users with detailed information on the dataset and clustering results (Figs. 3–7).

The values displayed at the top are the MSE, AIC, and BIC of the MOB tree depth. Fig. 2 shows two splits (MOB depth = 3). The first three plots are heatmaps showing the resulting time series patterns; their corresponding MOB trees are displayed on the right side. In these heatmaps, each row represents one series. The top heatmap displays time series in their original time scale (daily). The second and third heatmaps combine time periods into a useful (seasonal) aggregation to highlight potential weekly and monthly seasonal effects. Since we scale the series, values are centered around zero. There are two sets of heatmap annotations. The first set at the bottom displays the time features of the series, including weekday, month, and year. The second set on the left indicates location-related information, latitude, and longitude. Finally, the green-labeled boxes next to the heatmaps include the cluster number and the number of series in each cluster.

The selection of an appropriate color spectrum is very important for displaying all the information and projecting the numerical properties of the data (Wu et al., 2010). Here, we use sequential pallets with different color options. The default is green-white-red for low-to-high values, and the user can switch to blue-white-orange. For both heatmaps, we order series by comparing their values at each time point. On the right side of each heatmap, we display the MOB tree that creates clusters. The MOB tree makes it easy to see what domain-relevant attributes are selected in each cluster so that one can distinguish between clusters easily.

The fourth-from-top chart is a line chart showing time plots of the individual and average series. In this chart, we display all series of all clusters in gray and overlay the average across all series in red. While selecting a specific point along the time series line, the corresponding value of the series is shown in a box at the bottom of the chart.

The bottom-most plot is used to compare linear models in different clusters. This coefficient plot compares regression coefficients for all selected predictors in the linear models. Clicking on a line generates a box at the bottom of the chart, displaying the corresponding coefficient value. Appendix A describes the different operations that can be achieved from the user interactive menu.
Fig. 2. Clustering PM$_{2.5}$ series in Taiwan - Overview.
Fig. 3. Clustering PM$_{2.5}$ series in Taiwan - Performance statistics.

Model-Based Time Series Clustering Tool

Choose file to upload

Browse... Taiwan-PM2.5-Daily.2015-2019.csv

Uploaded complete

Frequency

- Weekly
- Monthly

First Set

MOB depth

3

Prune options

- AIC

Split variables

- Administrative
- Population
- Metro
- TEPA
- Station

Heatmap Color - High

red

Heatmap Color - Low

green

Second Set

MOB depth

2

Prune options

- AIC

Split variables

- Administrative
- Population
- Metro
- TEPA
- Station

Heatmap Color - High

red

Heatmap Color - Low

green

Take a screenshot

Fig. 4. Clustering PM$_{2.5}$ series in Taiwan - User interactive menu.
Fig. 5. Clustering PM$_{2.5}$ series in Taiwan - Overall heatmap.
Fig. 6. Clustering PM$_{2.5}$ series in Taiwan – Day of week heatmap.
Fig. 7. Clustering PM$_{2.5}$ series in Taiwan - Month of year heatmap.
3.1 Interpreting the PM$_{2.5}$ Clustering Results

To obtain the results in Fig. 2, we used MOB depths of 3 (two levels of tree splitting). We chose AIC as the pruning criterion and selected all domain-relevant options (TEPA air quality zones, station type, metro, population, and administrative level) as splitting variables. We used a green-white-red heatmap color pallet for these figures. This selection results in four clusters in Fig. 2. In Fig. 2, the first splitting variable is the TEPA air quality zone that separates KP from other zones, and its second split is on the station type (A, B, I, T vs. ~A/P). Among all other TEPA air quality zones (NT, CM, YI, CT, YCN, and HT), NT is further separated. The attribute profiles of the four clusters are given in Table 2. This table shows all the domain-relevant attributes for each of the clusters displayed in Fig. 2.

The upper heatmaps in Fig. 2 (or Fig. 5) display all the series in each cluster. Each row (vertical line) in Fig. 2 or each column (horizontal line) in Fig. 5 is one series, and series are sorted by leaf nodes in a dendrogram obtained by hierarchical clustering (Hahsler et al., 2008). In Fig. 2, we set the color range of these heatmaps from green to red (from low to high), and missing values appear in gray. Based on the MSE result, the best number of splits is two (depth = 3), as shown in Fig. 2 and Fig. 3.
Table 2. Cluster categories for the Taiwan PM$_{2.5}$ time series by choosing the MOB depth of three, AIC as pruning option, and TEPA air quality zones, station type, metro, population, and administrative level as domain-relevant attributes - based on Fig. 2.

<table>
<thead>
<tr>
<th>Cluster Categories</th>
<th># series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>32</td>
</tr>
<tr>
<td>TEPA zones: CM, YI, CT, YCN, HT</td>
<td></td>
</tr>
<tr>
<td>Cluster 2</td>
<td>24</td>
</tr>
<tr>
<td>TEPA zones: NT</td>
<td></td>
</tr>
<tr>
<td>Cluster 3</td>
<td>14</td>
</tr>
<tr>
<td>TEPA zones: KP</td>
<td></td>
</tr>
<tr>
<td>Station type: A, B, I, T</td>
<td></td>
</tr>
<tr>
<td>Cluster 4</td>
<td>1</td>
</tr>
<tr>
<td>TEPA zones: KP</td>
<td></td>
</tr>
<tr>
<td>Station type: A/P</td>
<td></td>
</tr>
</tbody>
</table>

Furthermore, the size of each cluster box differs based on the number of series in the cluster. For example, in Fig. 2, the series shows slightly different patterns when comparing different clusters. The number of series in each cluster is labeled in the green boxes next to the heatmaps (also in Table 2).

Here clusters 1, 2, and 3 show obvious seasonality, with a high concentration of PM$_{2.5}$ during cold seasons and a low concentration during hot seasons. Cluster 3 is formed by time series obtained from ambient, background, industrial, and traffic type of stations in the KP TEPA air quality zones, and it experienced heavier seasonality from 2015 to 2017, where the amount of PM$_{2.5}$ is higher (darker red) during winters and lower during summers (darker green) compared with the other two clusters. The KP zone usually exhibits a relatively higher PM$_{2.5}$ concentration (darker red) since it is situated on the leeside of the mountain, and the prevailing northeasterly wind flow can become obstructed (Cheng and Hsu, 2019). This means stricter air quality control policies (e.g., more transport control stations, improving fuel type) are required in this air quality zone. Clusters 1 and 2 also show high winter PM$_{2.5}$ concentrations between the years 2015 and 2016, although they begin to decrease from 2017 to 2019. Cluster 4 is formed by one A/P station in the KP zone and does not indicate an obvious seasonality pattern. Among all clusters, the NT zone shows the greatest reduction in PM$_{2.5}$ from 2015 to 2019.

We can also see the similarity of series inside each cluster. Based on our results and literature review (Hsu et al., 2019), PM$_{2.5}$ concentrations increase from south to north across Taiwan. This can be spotted easily from the first, second, and third clusters. Missing values appear in different clusters in different time periods, a phenomenon reflecting measurement problems at the stations or in data collection processes. For instance, in addition to the year 2015 in the first cluster, there are missing values in 2017 and 2018 in the third cluster. Tackling the missing values problem is essential in improving the PM$_{2.5}$ monitoring process and should be considered a priority for environmental scientists.

The middle heatmap (Fig. 6) represents a series aggregated by day of the week. This means all points within a day are grouped, and the color represents the average PM$_{2.5}$ concentration. With this plot, we can easily identify aggregated seasonal variations of the series on each day of the week. In Fig. 6, we see that the pattern is heavier in the third cluster, and this pattern is related to monthly and yearly seasonalities. In all the clusters, we do not see any specific pattern related to the changes in the days of the week, meaning there is no significant weekend effect on the PM$_{2.5}$ concentration in Taiwan.

The lower heatmap (Fig. 7) illustrates a series aggregated by month of the year, grouping all points within a month. Fig. 7 shows the seasonal pattern that decreases from May to August (summer in Taiwan) and increases from January to April (winter in Taiwan), which is common for all clusters. The beginning of January and February shows a high peak in the PM$_{2.5}$ concentration, which can be caused by using fireworks during the new year and Chinese new year celebrations (Lai and Brimblecombe, 2017). For other months, except for December, the PM$_{2.5}$ amount starts high (dark red) and ends low (dark green). Here, we also see a higher concentration of PM$_{2.5}$ in southern Taiwan during cold seasons. This plot can help guide changes to air quality policies in different months in different regions.

As we examine the line chart (displaying the average series in red, overlaid on the original series in gray) in Fig. 8, Cluster 1 (Chu-Miao [CM], Yilan [YI], Hua-Tung [HT], central Taiwan [CT],
and Yun-Chia-Nan [YCN] TEPA air quality zones) has the largest number of series (e.g., monitoring stations), and the average line shows the most pronounced decline in the PM$_{2.5}$ concentration.

Finally, from the coefficient plot in Fig. 8, coefficients in all clusters differ mostly in terms of autocorrelation (first lag). Table 3 summarizes the main MTSC tool results and presents decision-making insights.

In summary, our results show the PM$_{2.5}$ seasonal pattern in Taiwan air quality zones for various types of stations. We find the following patterns: There is an increasing pattern in PM$_{2.5}$ concentration from Taiwan’s northern to southern regions and a higher concentration of PM$_{2.5}$ (maximum value around 120 µg m$^{-3}$) during the cold season compared to the hot season (maximum value around 60 µg m$^{-3}$). Our results do not show any seasonal variation by day of the week (no significant weekend effect), but there is a seasonal pattern each year. In all the clusters, there is a peak in early January and February, which is likely caused by using fireworks during new year celebrations (Lai and Brimblecombe, 2017). An obvious seasonality pattern is also observed in the CM, CT, HT, YCN, and YI zones, and a slight pattern in the NT zone. In the KP air quality zone, the winter/summer seasonal pattern is more pronounced in ambient, background, traffic, and industrial stations. The fourth cluster is formed by one series (ambient/park - Pingtung/Hengchun station) located in the southernmost region of Taiwan. This series behaves differently from other clusters with no apparent seasonal pattern. We note that this region is currently grouped by TEPA together with other southern regions, but we suggest that it should be separate.

The PM$_{2.5}$ concentration is generally higher in southern Taiwan than in northern Taiwan. The NT zone shows the highest reduction of PM$_{2.5}$ concentration in all seasons among other air quality zones. In the KP zone, the PM$_{2.5}$ concentration is high due to its geographical situation and northeasterly wind flow. In the literature and based on the EPA expert’s insight, the NT and CM zones (northern region) are grouped together based on their PM$_{2.5}$ concentration pattern (Cheng and Hsu, 2019). Yet, in our results, the NT zone is separated as one cluster. This might be due to the NT zone’s noticeable improvement in air quality compared with the other zones. Another difference between our results and the EPA expert’s insight is the third cluster, where KP is a separate cluster. Typically, however, experts expect to see all of southern Taiwan together. In some studies, such as Cheng and Hsu (2019), the KP air quality zone is analyzed as a separate category. An advantage of our analysis is the use of a long observation period (five years of daily data), which reveals long-term variations in PM$_{2.5}$ seasonal patterns.

We also show whether the results confirm existing knowledge in the literature or from EPA experts. For example, EPA officials expect to see higher concentrations of PM$_{2.5}$ during the Chinese New Year season. However, our results show a peak in early January and February, which is likely caused by using fireworks during new year celebrations.

Table 3. Results from the MTSC tool and decision-making insights.

<table>
<thead>
<tr>
<th>MTSC Tool Results</th>
<th>Insights</th>
<th>Decision-making Insights</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{2.5}$ concentrations increase from south to north</td>
<td>needs stricter policies in southern Taiwan</td>
<td></td>
</tr>
<tr>
<td>Missing values appear in different clusters</td>
<td>needs improvement in publicly available datasets</td>
<td></td>
</tr>
<tr>
<td>Heavy seasonality in all clusters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General improvement in PM$_{2.5}$ from 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT as one cluster</td>
<td>northern Taiwan together less polluting industry in north</td>
<td>✓</td>
</tr>
<tr>
<td>KP as one cluster</td>
<td>southern Taiwan together KP is home to many manufacturing centers</td>
<td>✓</td>
</tr>
<tr>
<td>NT greatest reduction in PM$_{2.5}$ from 2015 to 2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KP shows higher concentration of PM$_{2.5}$</td>
<td>no difference between weekdays and weekend</td>
<td></td>
</tr>
<tr>
<td>No daily (weekend) effect</td>
<td>no need to change policies on weekends</td>
<td></td>
</tr>
<tr>
<td>Long holidays effect (early February and January)</td>
<td>Chinese new year (new year) effect changing fireworks habits increasing highway traffic control policies</td>
<td>✓</td>
</tr>
<tr>
<td>Monthly effect</td>
<td>cold seasons: higher - hot seasons: lower winter winds blowing in from China</td>
<td>✓</td>
</tr>
</tbody>
</table>
new year holidays. Our tool reveals a similar effect. Decreasing fireworks usage and increasing highway traffic control policies could be two decision-making insights for controlling air pollution during consecutive holidays.

4 CONCLUSION

We present a web-based visualization tool, namely the model-based time series clustering (MTSC) tool, to study Taiwan PM$_{2.5}$ seasonal variations in clusters made by spatial and population attributes. This tool helps to categorize the PM$_{2.5}$ dataset using cross-sectional features and explore seasonal patterns of the series by years, months, and weekdays. These preliminary results can be used to inform the decision-making of EPA experts in complex environmental problems and commonly encountered planning interventions, such as closures of schools and new transportation conditions. Our tool can also help to adjust air quality control policies by years, months, weekdays, or air quality zones. For example, rigid rules are suggested for the cold season and the KP zone in our demonstration. To the best of our knowledge, few user-friendly tools analyze PM$_{2.5}$ concentrations in Taiwan. This emphasizes the importance of conducting more research in this area.

Our clustering and visualization tool provides a simple interface for users and EPA experts to explore and decide on clustering setups using a variety of plots. Moreover, users do not need to know R programming or the specific details of the underlying clustering algorithm. Consequently, our web-based app is also a highly user-friendly data exploration tool, suitable for a wide range of users—from those with no knowledge of the underlying algorithm to those with a high degree of knowledge. While other datasets might be much larger than our example (e.g., Taiwan PM$_{2.5}$ from Internet-of-Things (IoT) monitoring devices), our web-based tool easily scales up in terms of the length of time series, the number of series, and the number of domain-relevant attributes.

The visualization tool includes MSE, AIC, and BIC values, three heatmaps with their MOB trees, a line chart, and a coefficient plot. To allow easy comparison of two different sets of parameters, it can create two tabs, each with one set of results and charts. A ‘screenshot’ button allows users to save all the results. Heatmaps show time series seasonal patterns in each cluster, and the corresponding MOB trees show the categories of each cluster based on the automatically selected domain-relevant attributes. Users can also apply different sets of colors for low and high values. These heatmaps further display critical information, such as missing values, which is very useful for troubleshooting. Line charts of the original series, along with the average line, display the overall patterns and trends of the series in each cluster. The coefficient plot can be used for comparing linear models across clusters. There is also an option to check the numerical value of the coefficients by clicking on the coefficient points on the line chart. The web app combines and summarizes all this rich information into a user-friendly interface. Annotations at the bottom and left side of the heatmaps are visually helpful in checking the series’ time (weekday, month, and year) and the location of the stations.

Although this app is designed to visualize Taiwan PM$_{2.5}$, it can be easily applied to other time series datasets associated with domain-relevant attributes. To apply this web-based tool to a different dataset, we only need a dataset with ‘Series’, ‘Date’, ‘cat.col’ (determines the series), ‘latitude’, ‘longitude’, and domain-relevant attribute (with arbitrary names) columns. By changing the dataset, the domain-relevant attributes list is updated automatically. We also need to consider one limitation of the MOB clustering approach regarding the number of domain-relevant attribute categories. When one or more attributes contain many categories, the MOB tree algorithm implementation in R encounters memory limitations. To handle this problem, as suggested in Ashouri et al. (2019), we can group categories of the splitting variable into a smaller set using domain knowledge.

ACKNOWLEDGMENTS

The authors would like to express their special appreciation and thanks to Dr. Charles C.-K. Chou for his expert ideas and feedback. The authors also would like to thank Dr. Travis Greene
for providing English editing services for this paper. This work is partially supported by the Academia Sinica grant number AS-TP-109-M07 (Phoa, Ashouri), and the National Science and Technology Council (Taiwan) grant numbers 111-2410-H-007-030-MY3 (Shmueli), 109-2118-M-001-003-MY2 (Chen), 107-2118-M-001-011-MY3 (Phoa) and 111-2118-M-001-007-MY2 (Phoa).

SUPPLEMENTARY MATERIALS

Codes and datasets are available: https://github.com/mahsaashouri/model-based-time-series-clustering-tool. Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.230124

REFERENCES


Wang, F., Chen, T., Chang, Q., Kao, Y.W., Li, J., Chen, M., Li, Y., Shia, B.C. (2021). Respiratory diseases are positively associated with PM$_{2.5}$ concentrations in different areas of Taiwan. PLoS One 16, e0249694. https://doi.org/10.1371/journal.pone.0249694


