Application of k-means clustering and calendar view Visualisation for air pollution index analysis

ABSTRACT

Two years of diurnal concentration of particulate matter (PM10) and nitrogen dioxide with the addition of relative humidity measurement, collected from Putrajaya, Malaysia's ground-based measurement station from January 2014 to December 2015, were analysed. Kmeans clustering was employed and optimal clusters of four were identified for each year based on the most suggested number of clusters from internal cluster validation measures of the total within sum of square, silhouette index and gap statistics. Each cluster was then profiled where each mean pollutant sub-indices were calculated and the contributing pollutant to the air pollution index (API) was determined by looking at the maximum value from all subindices. This mechanism closely follows the Recommended Malaysian Air Quality Guidelines (RMG) for determining API. Particulate matter was found to be the dominant sub-index in all clusters and then paired with the mean relative humidity for visualisation. A calendar view was selected to show the temporal patterns and we observed a consistent cluster profile with the actual mean values of the selected parameters for most months. The calendar view also suggested that overall, the API (based on particulate matter) in 2014 was much better as compared to 2015.