A spatial feature engineering algorithm for creating air pollution health datasets

Raja Sher Afgun Usmani a,b, Thulasyammal Ramiah Pillai a, Ibrahim Abaker Targio Hashem b, Noor Zaman Jhanji a, Anum Saeed c, Akibu Mahmoud Abdullahi a

a School of Computer Science and Engineering, Taylor’s University, Subang Jaya, Selangor, Malaysia
b College of Computing and Informatics, Department of Computer Science, University of Sharjah, Sharjah 27272, UAE
c Center for Advance Studies in Engineering, Islamabad, Pakistan

1. Introduction

Outdoor air pollution is seen as a major public health problem worldwide. About 4 million deaths occurred in 2016 due to air pollution, mainly from respiratory or cardiovascular diseases (Burns et al., 2019). Ambient air pollution is also associated with many other medical conditions, such as asthma. It is of great concern in countries with low and middle incomes, where air quality can deteriorate even more, and in high-income countries, ambient air pollution has declined in recent years (Burns et al., 2019). Researchers have recently focused on air pollution and its environmental effects. The researchers are looking into the techniques and policies that can minimize these associated health impacts.

The studies in this domain use different types of datasets, but in general, two datasets are necessary for conducting the health impact of air pollution study. The first dataset is the air quality dataset, and the second is the health dataset. Usually, the air pollution data is captured through the air quality monitoring (AQM) stations distributed across regions. The health data for the same region/state/county is obtained to associate with the air pollution data (Cox Jr et al., 2015). The residential address is one of the primary association parameter used to associate the person to the region (Huang and Batteman, 2000).

Feature engineering is used by researchers to extract meaningful information from these datasets using the domain knowledge (Wikipedia, 2020). Feature engineering is also used to create and optimize air quality and health features. Feature engineering gives researchers the flexibility to choose from different algorithms because less complex algorithms also offer good accuracy with useful features (Thanaki and Arora, 2017). On average, 80% of the time is spent on collection, processing, cleaning and organization of data (Ozdemir and Susarla, 2018). Furthermore, three-quarters of data scientists say that cleaning and engineering data is the least entertaining part of their job (Ozdemir and Susarla, 2018). Using residential address or any type of spatial parameter becomes a spatial problem when the AQM stations...
are concentrated in urban areas within the regions (Vitolo et al., 2018). Figure 1 shows example of concentrated AQM stations in Klang Valley, Malaysia. The distance between Petaling Jaya and Cheras AQM stations is 8.55 km. Fig. 1 shows the coverage area of the AQM stations with a radius of only 5000m (5km). It clearly shows an overlap of coverage at intersection area at Taman Desa and Brickfields regions in Kuala Lumpur, Malaysia. As the region/state/county is used as association parameter to merge these datasets, there is an obvious overlap in the AQM stations in urban areas, raising the question of how to associate the patients with the relevant AQM station (Usmani et al., 2020c).

To the best knowledge of the authors, there are very few studies that predict the health impact of air pollution (Dewi et al., 2019; Li et al., 2017; Vitolo et al., 2018; Yang et al., 2020) and authors found no studies that provide an efficient and automated way to generate air pollution health dataset. Most studies in this domain focus on the impact and associations of ambient air pollution and health. The most significant hurdle in conducting a study in this domain is the availability of datasets and the lack of any appropriate spatial feature engineering method that can efficiently combine these datasets and provide the combined dataset in a usable format. Keeping these conditions in mind, we propose a spatial feature engineering algorithm to combine air pollution and health datasets. The proposed algorithm will generate a new dataset in a simple, usable format. The main contributions of this research work are as below:

- A function to find coordinates of patients automatically.
- A function to calculate the distance between AQM stations and patient records.
- An algorithm to associate the patient record with the closest station based on the calculated distance. Hence, eliminating the limitation of concentrated AQM stations in urban areas, with the additional radius facility to exclude spatially distant records.

The rest of the paper is organized as follows: In Section 2, First, we discuss the background and the motivation to study air pollution and its health impacts. At the end of the section, we explore about the motivation to develop our feature engineering algorithm. Section 3 presents the details of the case study, the datasets we are using, and the feature engineering algorithm in detail. Section 4 includes the results, discussions, and future work. Lastly, the conclusions are presented in Section 5.

2. Related work

Air pollution is among the leading causes of mortality year after year. According to the World Health Organization (WHO) annual mortality statistics, the use of tobacco causes 7 million deaths; AIDS is the cause of 1.2 million deaths; tuberculosis is the cause of 1.1 million deaths, and malaria causes 0.7 million deaths (WHO, 2017). In the same year, 6.4 million deaths were due to air pollution globally, with 4.2 million deaths attributed to outdoor air pollution and 2.8 million due to indoor air pollution. If not aggressively controlled, The projection of ambient air pollution fatalities in the year 2060 is between 6 and 9 million (OCDE, 2016). Also, exposure to air pollution is listed as a significant risk factor for neurodegenerative diseases in adults and neurodevelopmental disorders in children. (WHO, 2017).

In recent years, many researchers have focused their attention on the associations of air pollution and health (Usmani et al., 2020d). Several leading medical articles have continually identified the link between particulate matter in ambient air pollution and its negative impact on respiratory and cardiovascular diseases and a sharp increase in fatalities among senior citizens. It is reported that these associations are unlikely to be elaborated by any confounding factor; they mainly represent cause and effect. The nature of the urban particulate cloud, however, is where the explanation lies (Seaton et al., 1995). It also indicates that these ultra-fine particles can cause alveolar inflammation by releasing mediators capable of causing aggravation of increasing the formation of blood clots and lung disease. This indicates the rising number of cardiovascular mortality due to urban air pollution.

Since the production of sulfur dioxide has increasingly declined over the globe (EPA) and in Asia (CAI-Asia, 2010), the focus has turned to particulate matter, nitrogen dioxide and ozone. Globally, millions of people live in rural areas in developing countries consume biomass fuels at rates that are, by percentage, higher than recently found in developed countries. About 2 million children suffer fatalities from acute respiratory infections caused by these exposures (Moraga et al., 2017).

According to European Union (EU), 80% of premature mortalities are due to coronary heart disease (CHD) and stroke, also attributed to air pollution, making it the most common cause of premature death from air pollution in Europe (EU, 2016). The factors by which cardiovascular disease is aggravated by air pollution are found to be similar to those that cause respiratory disease.
Cardiovascular diseases are attributed for 60-80% of air pollution-related deaths globally (Bourdrel et al., 2017). Therefore, the relation of air pollution to heart disease is reported in the 2016 European air quality report (Logstrup, 2017). The report reported 41 European countries. The European air quality report shows that air pollution caused 444,000 premature deaths from CHD and stroke. According to European Union statistics from 2013, air pollution was the cause of 416,000 premature deaths itepLXKBB037.

Table 1 presents a comprehensive account of studies regarding air pollution and health in the last 10 years. We also present the major ailments and diseases associated with ambient air pollution and the major air pollutants linked with these diseases. These studies used two types of data, i.e., air pollution data and health data. In order to merge these datasets residential address (Alexeef et al., 2018; Burte et al., 2018; Cesaroni et al., 2012; Cox Jr et al., 2015; Deng et al., 2016; Dono et al., 2014; Lin et al., 2002; Meier-Girard et al., 2019; Middleton et al., 2010; Rich et al., 2019; Turner et al., 2019; Wu et al., 2009), post/zip code Brauer et al. (2008); Kheirbek et al. (2016); Lavigne et al. (2017); Ribeiro et al. (2019), community/county/block/city Atalabi et al. (2019); Barnett et al. (2005); Beelen et al. (2007); Choi et al. (2020); Feng and Yang (2012); Ghosh et al. (2015); Goudarzi et al. (2016); Guo et al. (2016); Hwang et al. (2006); Lee et al. (2002); Liu et al. (2018); Luong et al. (2018); Perez et al. (2012); Requia et al. (2016); Rich et al., 2019; Wu et al., 2009), post/zip code Brauer et al. (2008); Kheirbek et al. (2016); Lavigne et al. (2017); Ribeiro et al. (2019), community/county/block/city Atalabi et al. (2019); Barnett et al. (2005); Beelen et al. (2007); Choi et al. (2020); Feng and Yang (2012); Ghosh et al. (2015); Goudarzi et al. (2016); Guo et al. (2016); Hwang et al. (2006); Lee et al. (2002); Liu et al. (2018); Luong et al. (2018); Perez et al. (2012); Requia et al. (2016), hospital/school address (Buchdahl et al., 2000; Dono et al., 2014; Tajudin et al., 2019; Thompson et al., 2001) was used as association parameter. The association parameter these studies used to link the air pollution and health dataset are presented in Table 1 as well. As discussed in Section 1 and demonstrated in Fig. 1, using these parameters for the association of air pollution and health datasets is insufficient as there is a concentration of AQM stations in urban regions. Also, in order to use community/county/block/city parameters with multiple AQM stations researchers rely on average value of air pollutant from multiple station (Choi et al., 2020). In most of the studies, the distance of patients to the AQM stations is also not taken into account. To cater to this problem, we propose a spatial feature engineering algorithm, which automatically finds the appropriate AQM station and associate the patient with it.

3. Methodology

In this section, we explain the datasets used in this study and the spatial feature engineering algorithm in detail. Figure 2 represents the flow of the methodology. Three types of datasets are utilized in this methodology, i) Health dataset, ii) Air pollution dataset, and iii) AQM stations’ information. This section also presents the proposed algorithm in detail in Section 3.2. Although the case study is based in Klang Valley, Malaysia, the algorithm is not specific to the case study. The proposed algorithm can be used to associate air quality and health datasets in any location around the world. The MapQuest API used to find coordinates of address is also freely available with data for all countries (MapQuest, 2020).

3.1. Data

This study is conducted using the Klang Valley, Malaysia, as a case study. Klang Valley is an urban region in Malaysia that is centered in Kuala Lumpur. Klang Valley includes Kuala Lumpur’s adjoining cities and towns in the state of Selangor. In this case study, we used two datasets, i.e., air pollution dataset from the AQM stations provided by the Department of Environment (DOE), Malaysia, and hospitalizations dataset, provided by the Ministry of Health (MOH), Malaysia. The details of the datasets are provided below.

3.1.1. Air pollution dataset

In Malaysia the monitoring of air quality monitoring was conducted via Alam Sekitar Malaysia Sdn Bhd (ASMA) until 2017. ASMA is a private company, which was selected by the Malaysian Meteorological Department (METMalaysia) and DOE. ASMA was responsible for gathering, processing, analyzing and circulating measurements of air pollutants. There are 66 AQM station throughout Malaysia, 14 Manual sampling (High Volume Sampler) stations operated by METMalaysia, and 52
Table 2

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{10}$</td>
<td>51.15864</td>
<td>45.70833</td>
<td>40.37500</td>
<td>10.70588</td>
<td>580.87500</td>
</tr>
<tr>
<td>CO</td>
<td>0.86181</td>
<td>0.79044</td>
<td>0.72870</td>
<td>0.04000</td>
<td>5.96762</td>
</tr>
<tr>
<td>$O_3$</td>
<td>0.01912</td>
<td>0.01830</td>
<td>0.01983</td>
<td>0.00005</td>
<td>0.01882</td>
</tr>
<tr>
<td>$NO_2$</td>
<td>0.03593</td>
<td>0.03230</td>
<td>0.03100</td>
<td>0.00023</td>
<td>0.15462</td>
</tr>
<tr>
<td>$NO$</td>
<td>0.01970</td>
<td>0.01882</td>
<td>0.01700</td>
<td>0.00005</td>
<td>0.06562</td>
</tr>
<tr>
<td>$NO_2$</td>
<td>0.01622</td>
<td>0.01300</td>
<td>0.00800</td>
<td>0.00000</td>
<td>0.011291</td>
</tr>
<tr>
<td>$SO_2$</td>
<td>0.00310</td>
<td>0.00274</td>
<td>0.00200</td>
<td>0.00000</td>
<td>0.03415</td>
</tr>
</tbody>
</table>

Continuous air quality monitoring (CAQM) stations operated by ASMA (DOE, 2020a). The DOE has recently increased the number of CAQM station from 66 stations to 68 stations (DOE, 2020b). In this study, we are using the data from eight CAQM stations in Klang Valley. The geo-coded details of these CAQM stations are provided in Fig. 3.

The air pollution dataset contained daily readings for the major air pollutants, i.e., PM$_{10}$, CO, $O_3$, $NO_2$, $NO$, and $SO_2$. The air pollution dataset contained data for 6 years, i.e., 2011-2016. The statistics of the dataset are provided in the Table 2. This dataset is cleaned and engineered by using the novel feature engineering algorithm for air pollution datasets (Usmani et al., 2020a).

3.1.2. Hospitalizations dataset

MOH, Malaysia provide the hospitalizations dataset for this study. The dataset spans the same duration as the air pollution dataset, i.e., 2006-2016. The hospitalizations dataset contains the daily cardio-respiratory hospitalization records of patients in Klang Valley. Table 3 lists down the sample hospitalization data from the hospitalization dataset. In this dataset, diagnosis is provided in the form of the International Classification of Diseases (ICD) codes. The ICD is maintained by the WHO and used worldwide for mortality and morbidity statistics. P-ID denotes the patient identification number, which is unique to a hospital. The hospitalization dataset does not contain any redundant data and all the patients present in the dataset are included in the case study. If the patients without proper residential address exists in the dataset, they will be excluded from the study as residential address is the association parameter used to merge the datasets.

3.2. Proposed algorithm

The proposed spatial engineering algorithm consists of four major functions. The first part, FindCoordinates function takes the Health
dataset as the input and finds the coordinates using the spatial information of the patient, i.e., the residential address, and this information, along with the Health dataset is then used in \textit{FindDistance} function. The \textit{FindDistance} function passes the coordinates from Health dataset and the AQM stations information into \textit{CalculateDistanceInMeters} function to find the distance between patients and the AQM stations. These distances, along with the Health dataset, are used in the function \textit{AssociateDatasets}, which spatially combines the air pollution dataset and the health dataset. The AQM station-wise Air Pollution Health dataset is produced at the end of the proposed algorithm. The Algorithm 1 presents the complete algorithm, with \textit{Start} function as a entry point of the algorithm. In the coming sections, we present and discuss the functions of the proposed algorithm.

3.2.1. \textit{FindCoordinates}

The first part of the proposed algorithm is the \textit{FindCoordinates} function, which performs a crucial task known as forward geocoding. Forward geocoding or address geocoding is the process of locating the coordinates (latitude and longitude) for an address. The \textit{FindCoordinates} function performs the forward geocoding using the address of the patient. For forward geocoding, we utilize the MapQuest free Geocoding API (MapQuest, 2020). The \textit{FindCoordinates} function loads all the hospitalization records and creates a MapQuest web request to find the coordinates for the address. The Geocoding API returns the coordinates of the given address, and these coordinates are saved as the patient’s coordinates.

3.2.2. \textit{CalculateDistanceInMeters}

The \textit{CalculateDistanceInMeters} function uses the library \textit{System.Device.Location} from the.NET Framework (Microsoft, 2020), which is an implementation of Haversine’s formula (Sinnott, 1984). The \textit{CalculateDistanceInMeters} function finds the distance in meters between two sets of coordinates. The mathematical representation of the \textit{CalculateDistanceInMeters} function is shown in Eq. (1). In Eq. (1), the \textit{Cos} function returns the cosine of the specified angle, the \textit{Sin} function returns the sine of the specified angle, and the \textit{Atan2} function returns the angle whose tangent is the quotient of two specified numbers. The variables \textit{latitude}_1, \textit{longitude}_1 represent the first coordinate’s latitude and longitude, and the variables \textit{latitude}_2, \textit{longitude}_2 represent the second coordinate’s latitude and longitude. The variable \textit{r} is the radius of the earth. The mean radius of earth value used in \textit{System.Device.Location} is 6,376,500 m (Microsoft, 2020).

\[
\begin{align*}
    d_1 &= \text{latitude}_1 \times \left( \frac{\pi}{180} \right), \text{num}_1 = \text{longitude}_1 \times \left( \frac{\pi}{180} \right) \\
    d_2 &= \text{latitude}_2 \times \left( \frac{\pi}{180} \right), \text{num}_2 = \text{longitude}_2 \times \left( \frac{\pi}{180} \right) - \text{num}_1 \\
    d_3 &= \sin\left(\frac{d_2-d_1}{2}\right)^2 + \cos(d_1) \times \cos(d_2) \times \sin\left(\frac{\text{num}_2}{2}\right)^2 \\
    \text{distance} &= 2 \times r \times \text{Atan2}\left(\sqrt{d_3}, \sqrt{1-d_3}\right)
\end{align*}
\]

3.2.3. \textit{FindDistance}

The third part of the algorithm reads all the hospitalization records and AQM stations’ information and is responsible for finding patients’ distance from the air quality monitoring stations. The most important feature of this part of the algorithm is the radius parameter. A researcher can specify the maximum distance between the AQM station and the patient, ensuring that irrelevant patients are not associated with the dataset. It iterates through all the hospitalization records and uses the \textit{CalculateDistanceInMeters} function to find the distance in meters between the record and the AQM stations. In the end, it calculates the AQM station with the smallest distance to the patient and associates the specific AQM station with the patient.

3.2.4. \textit{AssociateDatasets}

The last part of the algorithm generates a new dataset by combining the two datasets. The \textit{AssociateDatasets} function reads all AQM stations information, air pollution dataset and updated hospitalizations dataset. It iterates over the stations and gets all the dates in the datasets for that station. The next step iterates through these dates and combines the air pollution reading of that date with the number of hospitalizations on the same date for the station. It saves the new record, and the new AQM station-wise dataset is generated with the combination of health and air pollution datasets.

4. Results & discussions

The results indicate that the proposed algorithm generates an air pollution health dataset efficiently. The proposed algorithm allows the researchers to automatically find coordinates of patients records, calculate the distance between AQM stations and patient’s residential address, and generate a new air pollution health dataset by associating the closest station to the patient. Hence, using the proposed algorithm will reduce feature engineering time and also reduce the efforts of researchers indubitably. The major motivation behind the proposed algorithm was to solve the overlap due to the congestion of AQM stations in urban areas. The research trends show that the residential address (Alexeiff et al., 2018; Burte et al., 2018; Cesaroni et al., 2012; Cox Jr et al., 2015; Deng et al., 2016; Dons et al., 2014; Lin et al., 2002; Meier-Girard et al., 2019; Middleton et al., 2010; Rich et al., 2019; Turner et al., 2019; Wu et al., 2009), post/zip code (Brauer et al., 2008; Kheibek et al., 2016; Lavigne et al., 2017; Ribeiro et al., 2019), community/county/block/city (Alotaibi et al., 2019; Barnett et al., 2005; Beelen et al., 2007; Feng and Yang, 2012; Ghosh et al., 2015; Goudarzi et al., 2016; Guo et al., 2016; Hwang et al., 2006; Lee et al., 2002; Liu et al., 2018; Luong et al., 2018; Perez et al., 2012; Requia et al., 2016), hospital/school address (Buchdahl et al., 2000; Dons et al., 2014; Tajudin et al., 2019; Thompson et al., 2001) are utilized to associate the patients/subjects with the AQM stations or air pollution datasets. Using these parameters in a congested urban area with multiple AQM stations create an overlap of coverage area, as shown in Figure 1. The proposed algorithm was able to resolve this problem by finding the distance between every patient with the AQM stations and picking the closest one to generate the air pollution health dataset.

In the upcoming sections, the results for the functions \textit{FindCoordinates}, \textit{FindDistance} are discussed, and then the dataset generated by \textit{AssociateDatasets} function is presented. In the end, we discuss the impact
Algorithm 1: Spatial feature engineering algorithm

Data: Air pollution dataset, Hospitalization dataset and AQM station information
Result: Air pollution health dataset

Function FindCoordinates()
allRecords = Load all hospitalization records
foreach record in allRecords do
  if record does not have proper address then
    continue
  end
  webrequest = Create MapQuest webrequest
  completeAddress = record.Address + record.postcode
  webrequest.Address = completeAddress
  jsonResult = webrequest.getResponse()
  record.Latitude = jsonResult.Latitude
  record.Longitude = jsonResult.Longitude
  Save record
end

Function CalculateDistanceInMeters(lat1, long1, lat2, long2)
PI = 3.141592653589793

d1 = lat1*(PI/180.0)
n1 = long1*(PI/180.0)
d2 = lat2*(PI/180.0)
n2 = long2*(PI/180.0) - n1

d3 = Power(Sin((d2 - d1) / 2.0), 2.0) + Cos(d1) * Cos(d2) * Power(Sin(n2 / 2.0), 2.0)

return 6376500.0 * (2.0 * Atan2(Sqrt(d3), Sqrt(1.0 - d3)));}

Function FindDistance
allRecords = Load all hospitalization records
allStations = Load all stations
radius = Set maximum distance from station
foreach record in allRecords do
  allStationDistances = reset allStationDistances
  foreach station in allStations do
    distance = CalculateDistanceInMeters(record.Latitude, record.Longitude, station.Latitude, station.Longitude)
    Add distance, station to allStationDistances
  end
  smDistSt = Get station with smallest distance from allStationDistances
  if smDistSt.distance <= radius then
    record.station = smallestDistanceStation
    Save record
  end
end

Function AssociateDatasets
allHospRecords = Load all hospitalization records
allApRecords = Load all air pollution records
allStations = Load all stations
foreach station in allStations do
  allDates = Get all dates from datasets for the station
  foreach date in allDates do
    apRecord = Get apRecord on that date for the station from allApRecords
    hospCountRecord = Get hospitalizations count on that date for the station from allHospRecords
    aphealthRecord = create single row object with date, apRecord, hospCountRecord
    Save aphealthRecord
  end
end

Function Start()
FindCoordinates()
FindDistance()
AssociateDatasets()
of the parameter radius and how the output changes when the radius parameter is changed.

4.1. FindCoordinates()

The FindCoordinates function is responsible for the crucial task of forward geocoding, i.e., the process of locating the coordinates (latitude and longitude) for an address. The MapQuest free Geocoding API (MapQuest, 2020) to find the coordinates for all the patients’ addresses. The Table 4 presents the output of the FindCoordinates function. The input used for this output is the Table 3. For the sake of privacy, we have used sample data to demonstrate the FindCoordinates function’s output. The columns Latitude and Longitude will be used as coordinates in the functions CalculateDistanceInMeters and FindDistance.

4.2. FindDistance()

The second part of the algorithm reads all the hospitalization records and AQM stations’ information and is responsible for finding patients’ distance from the air quality monitoring stations. The most important feature of this part of the algorithm is the radius parameter. A researcher can specify the maximum distance between the AQM station and the patient, ensuring that irrelevant patients are not associated with the dataset. It iterates through all the hospitalization records and uses the CalculateDistanceInMeters function to find the distance in meters between the record and the AQM stations. In the end, it calculates the AQM station with the smallest distance to the patient and associates the specific AQM station with the patient.

The Table 5 presents the sample output of the FindDistance function. The output is generated with the data from Table 4 as input, along with the AQM stations’ information, which includes the coordinates of the AQM stations. The output of the FindDistance function contains the distance of the patient with each AQM monitoring station via the CalculateDistanceInMeters function. Also, the smallest distance are highlighted in the Table 4. For the sake of brevity the columns Gender-Lengthitude from Table 4 are hidden and only new columns are presented in Table 5. The output presented in Table 5 will be passed into the AssociateDatasets function and the closest AQM station will be associated with the patient accordingly.

4.3. AssociateDatasets()

The last part of the algorithm generates a new dataset by combining the two datasets. The AssociateDatasets function reads all AQM stations information, air pollution dataset and updated hospitalizations dataset. It iterates over the stations and gets all the dates in the datasets for that station. The next step iterates through these dates and combines the air pollution reading of that date with the number of hospitalizations on the same date for the station. In the end, it saves the new record, and the new AQM station-wise dataset is generated with the combination of health and air pollution dataset.

Table 6 shows the sample data from the generated dataset. The generated dataset is in a simple row-format. The sample dataset contains the daily average air pollutant readings for the specific date and the count of hospitalizations on the same date. All eight monitoring stations from the case study are represented in the sample dataset. Also, the AQM station Kuala Selangor only records \( PM_{10} \) reading at the time of data collection. The generated dataset can be used for various associations and predictive models. Furthermore, we have generated the dataset for the hospitalization count for the case study, but the researchers can alter this to include any other health record characteristic.

4.4. The radius parameter

Our case study is based in Klang Valley, Malaysia, a highly populated urban area with a focus on industrialization. Table 7 shows the number of patients/subjects which are included in the generated air pollution health dataset. Cheras and Petaling Jaya AQM stations are based in Kuala Lumpur and are only 5.5 km apart. As Kuala Lumpur is a populated urban area, the number of patients/subjects is much higher than the other monitoring stations. After these two AQM stations in Kuala Lumpur, Petaling Jaya has the highest number of patients. The lowest number of patients is associated with Bukit Changgang AQM station, which is a small town in Klang Valley.

The results also show an interesting role played by the radius variable in FindDistance function. The radius parameter plays an integral part in the inclusion criteria. It has a significant impact on the number of hospitalizations included in the generated dataset. Table 7 demonstrates the included hospitalization records with a varying radius of stations. It is clearly visible that the radius parameter of the proposed algorithm has a substantial impact on the new dataset, with excluded records dropping from 341,234 to 150,171 with a change of radius from 5,000 meters to 10,000 meters. The trend of reduction in excluded records continues when the radius value is increased. With the proposed algorithm calculating individual distances for every patient, the researcher can increase the radius without worrying about the overlap between the AQM stations.

To summarize, the propose algorithm took the hospitalization dataset in the form of Table 3 and the air pollution dataset presented in Table 2, and converted into useable tabular format presented in Table 6. The output of the algorithm can now be passed into any Artificial Intelligence model for the association or prediction of health impact of air
pollution. The results show that the researcher will save time and effort by automating the process of generating the air pollution health dataset and the automated association of the patients with the closest AQM station will help decrease the exposure misclassification by removing the patient’s who live far away from the AQM stations.

To the best of our knowledge, no studies are found that provide an efficient and automated way to generate air pollution health dataset. Some studies do the patient’s association and dataset generation manually (Dons et al., 2014), and some studies used ArcGIS for geocoding process, which is entire platform of GIS tools for mapping, spatial analysis, and visualization (Ribeiro et al., 2019; Usmani et al., 2020b; Zhang et al., 2015), but most of the studies mention the association parameter, as described in the Table 1, but do not provide the method used to merge the datasets. Hence, it is not possible for us to do a comparison with other methods specific to this domain.

Although the case study is based in Klang Valley, Malaysia, the algorithm is not specific to the case study. The proposed algorithm can be used to associate air quality and health datasets in any location around the world. The proposed algorithm is not limited to the data from AQM stations and can be used for data integration with the internet of things (IoT) based air pollution monitoring system. The MapQuest API used to find coordinates of address is also freely available with data for all countries. The future directions include the addition of more parameters in both datasets. Additionally, using only the residential address as the spatial parameter can be considered as the limitation of this algorithm (Huang and Baterman, 2000), but using the radius parameter in the algorithm decreases the exposure misclassification by removing patients who live far from the AQM stations. This limitation can be improved upon by adding more spatial parameters in future work.

5. Conclusion

Air pollution is among the major environmental issues facing the world today and has severe health consequences. Scientists have increasingly used air quality and health datasets to research the health effects of air pollution. The common association parameter used to merge these datasets is the residential address. Due to concentrated AQM stations in urban areas, a patient can have multiple monitoring stations around them. In most studies, the distance of the station from the patients’ coordinates is not considered, which can lead to exposure misclassification. The proposed algorithm provides the facility to associate the patient with closest monitoring station and generate the air pollution health dataset accordingly. The proposed algorithm also includes a radius facility to exclude patients that live far away from AQM stations, decrease exposure misclassification and improve the accuracy of research findings.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgment

This research is funded by Taylor’s University under the research grant application ID (TUFR/2017/004/04) entitled Modeling and Visualization of Air-Pollution and its Impacts on Health. We are also thankful to Department of Environment, Malaysia for providing the AQM station dataset and Ministry of Health, Malaysia for providing the hospitalization dataset.

References


