Aerosol and Air Quality Research, 19: 1614–1624, 2019
Copyright © Taiwan Association for Aerosol Research
ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2018.08.0315

Topological Characterization of Haze Episodes Using Persistent Homology

Nur Fariha Syaqina Zulkepli*, Mohd Salmi Md Noorani, Fatimah Abdul Razak, Munira Ismail, Mohd Almie Alias

School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

ABSTRACT

Haze is one of the major environmental issues that have continuously vexed countries worldwide, including Malaysia, for the last three decades. Therefore, this study aims to investigate the differences between the topological features of months with and those without haze episodes observed at air quality monitoring stations located in the areas of Jerantut, Klang, Petaling Jaya and Shah Alam. We employ persistent homology, which is a method of topological data analysis (TDA) that focuses on connected components and holes in the data, to characterize the local particulate matter (PM_{10}). The summary statistics reveal drastic changes in the lifetimes of the topological data from every station during haze episodes, highlighting the possibility of developing an early detection system for haze based on our approach.

Keywords: Haze; Particulate matter; Persistent homology; Time delay embedding; Topological data analysis.

INTRODUCTION

Haze phenomenon is one of the major environmental issues that occur continuously in countries across the world, Malaysia included (Wen *et al.*, 2016). The country is also afflicted, with the Klang Valley region being one of the areas that are particularly affected by such phenomenon (Latif *et al.*, 2018). The severity of this issue can be attributed to the large-scale forest and plantation fires that occur in Indonesia's Sumatra Island. Nevertheless, it is worsened further by local emissions from domestic industries, vehicle usage, and open burning activities (Afroz *et al.*, 2003; Wen *et al.*, 2016). In Malaysia, the chronological history of haze episodes can be underlined with the severe incidents recorded in the year 2005, 2013, 2014 and 2015 accordingly (DOE, 2018a). This issue has continually emerged year after year in Malaysia and the neighboring countries of Brunei and Singapore, whereby the prolonged duration of the problem has spurred the study into investigating haze episodes.

Particulate matter (PM_{10}) are small particles floating around in the air in the form of smoke, dirt, and dust that originate from factories, vehicles and farming activities (Schwartz *et al.*, 1996; Payus *et al.*, 2013). During haze episodes, PM_{10} is typically highlighted as a major air pollutant in the Southeast Asian regions, particularly the Klang Valley, Malaysia (Afroz *et al.*, 2003; Azmi *et al.*, 2010). Acknowledged as a harmful pollutant, inhalation of the material leads to diminishing lung function and causes various respiratory diseases, especially acute exacerbation of asthma (Schwartz *et al.*, 1996). According to the Malaysian Ambient Air Quality Guidelines (MAAQG), an average of PM_{10} concentration over 24-hours that exceeds 150 µg m^{-3} is considered to be unhealthy for human health. Therefore, a standard technique used to identify haze phenomenon is by observing the concentration of PM_{10} whereby haze emergency is declared when it reaches the emergency level (> 500 µg m^{-3}) (DOE, 2018b).

Several studies have explored into a comparison of PM_{10} concentration with MAAQG standards, undertaken by Azmi *et al.* (2010), Abdullah *et al.* (2012), Ling *et al.* (2010) and Rahman *et al.* (2015) respectively. They have consequently concluded that the concentration in the air quality monitoring stations located in Klang Valley exceeded the acceptable level recommended by the MAAQG during haze episodes. Various differences have also been revealed between the concentration of PM_{10} according to the locations of air quality monitoring stations, encompassing rural, urban and industrial areas respectively. Moreover, Azmi *et al.* (2010) and Abdullah *et al.* (2012) have also revealed that urban areas logged higher PM_{10} concentrations compared to rural areas. A study by Yusof *et al.* (2010) has provided further analysis using statistical models, lognormal and Weibull distributions to investigate the relationship between PM_{10} concentration and monsoon season. Similarly, other methods like chemometric analysis (Azid *et al.*, 2015),

* Corresponding author.
Tel.: +6012-543-2006
E-mail address: farihasyaqina@yahoo.com

fuzzy comprehensive evaluation method (Zhao *et al.*, 2010) and chaotic approach (Hamid and Noorani, 2014) have also been utilized in assessing data on air pollutants. Furthermore, haze detection can also be found to be fast gaining scholarly traction. Previously, works focusing on the topic were geared towards analyzing satellite imagery data using remote sensing methods (Makarau, 2014). Meanwhile, recent research is specifically related to haze periods and focused on analyzing particulate matter concentration to study its morphology during haze episodes (Zeb *et al.*, 2018). Yu *et al.* (2018) have also contributed by investigating human health risk in an indoor and outdoor environment that is exposed to the phenomenon, whereas another study has looked into mitigating severe urban haze episodes (Sharma and Balasubramanian, 2018). It should be noted that this literature has analyzed PM_{10} in a quantitative manner, whereas to the best of our knowledge, no effort has been expended on its qualitative aspects. Therefore, this paper is addressing the research gap by providing an analysis of the qualitative aspects and structures of PM_{10}, particularly on their topological features via persistent homology.

Persistent homology is a relatively new method that is robust under perturbations of input data, independent of coordinates and dimensions alike, and offers a solid representation of qualitative features of the input data (Otter *et al.*, 2017). These particular features of the data are captured accordingly as specific parameters change. As the approach is fundamentally based on the mathematical field of topology, the qualitative features in question encompass topological features like connected components, holes, voids and more. The qualitative approach of persistent homology is also particularly useful in handling complexity of data due to noise, high dimensionality or incomplete structure. This poses great challenges to researchers dealing with real-world data since data cleaning is required in order to remove the noise, which might result in information lost during the process (Jorquera *et al.*, 2000; Elangasinghe *et al.*, 2014). By contrast, persistent homology retains all information from data. The noise that may occur in multiple scale levels is filtered out by persistent homology and significant features are captured (Ghrist, 2008). Furthermore, its robustness has led other researchers to explore the method further in various fields, such as Gidea and Katz (2018)'s effort on the capability of the holes (1-dimensional features) to detect financial crisis in stock market data. Meanwhile, Emrani *et al.* (2014) have investigated wheeze signal detection via the persistency of topological features between wheeze and non-wheeze signals. Moreover, the summary statistics of topological features undertaken by Mittal and Gupta (2017) has shown that persistent homology is usable in early detection of bifurcations and chaos in complex systems. The explorations on persistent homology have also been tackled in various fields, encompassing the classification of breast cancer (Dewoskin *et al.*, 2010), viral evolution (Chan *et al.*, 2013) and protein structure (Gameiro *et al.*, 2015). A good and comprehensive review regarding the current applications of persistent homology may be sourced from Otter *et al.* (2017). To the best of our knowledge, there is a negligible amount of research output on environmental issues by persistent homology to date. Therefore, this study is conducted to fill the research gap by exploring the effectiveness of persistent homology in characterizing and detecting one of the pressing environmental issues, namely haze phenomenon.

This paper intends to investigate haze episodes using persistent homology by analyzing the topological features of PM_{10} data in Malaysia. The proposed technique has been applied to daily average PM_{10} data from four air quality monitoring stations of Jerantut, Klang, Petaling Jaya and Shah Alam from the year 2000 until 2015. The objective of this study is to investigate topological features for months with and without haze episodes, thus rendering the data on PM_{10} being partitioned according to the respective month. The topological features are then extracted for connected components (0-dimensional features) and holes (1-dimensional features), with the resulting information represented in persistence diagrams. Next, the summary statistics of the topological features, like an average of all lifetimes of the connected components and maximum lifetimes of all holes, are calculated for each persistence diagram. In this paper, the content has been organized accordingly, whereby the subsequent section consists of a discussion on data involved in this study. Then, the data transformation processes and the method of persistent homology will be explained in next two sections. Extensive summary statistics of topological features will be introduced in the section before its resulting outcomes are discussed.

DATA

The daily average data for PM_{10} encompassing the year 2000 to 2015 for four air quality monitoring stations (i.e., Jerantut, Klang, Petaling Jaya, and Shah Alam) has been obtained from the Department of Environment (DOE), Malaysia. Klang and Shah Alam are categorized as urban areas, while Petaling Jaya and Jerantut are categorized as respectively of the industrial and rural areas, accordingly (DOE, 2018c). Due to its rural location, Jerantut, in particular, has been chosen as the background station for comparative purpose (Azmi *et al.*, 2010; Banan *et al.*, 2013). Missing data has been treated with the mean substitution method (Pigott, 2001). The MAAQG standards are necessary to represent the safety level that causes no adverse health effects to human. Therefore, this study has adhered to the safe level as recommended by MAAQG to act as the benchmark for air quality level of each station.

Fig. 1 shows the time series of the daily average for PM_{10} from 1 January 2000 until 31 December 2015, and the MAAQG for 24-hour average PM_{10} at 150 µg m^{-3}. The presence of several peaks (rectangles) is indicative of the values of PM_{10} concentration exceeding the MAAQG during haze episodes that have occurred in August 2005, June 2013, March 2014, September 2015, and October 2015 accordingly (DOE, 2018a). Thus, these months have been highlighted as the main focus of this research. It is worth noting that based on the chronology of haze episodes (DOE, 2018a), the selected months have been reported of

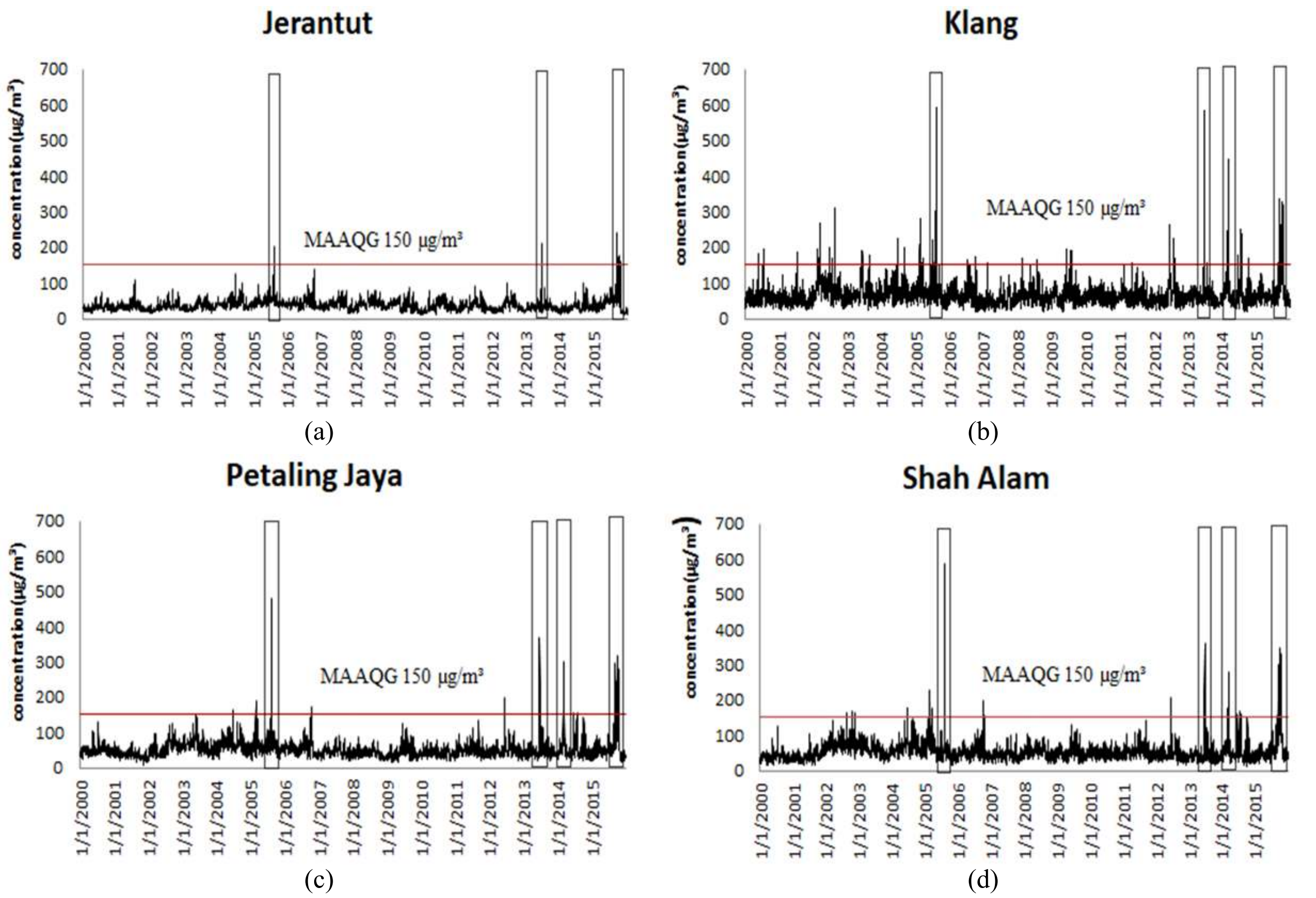


Fig. 1. Time series of daily average of PM_{10} for (a) Jerantut, (b) Klang, (c) Petaling Jaya and (d) Shah Alam air quality monitoring stations from 1 January 2000 until 31 December 2015.

experiencing severe haze. The descriptive statistics of the months are shown in Table 1 accordingly, with Klang, Petaling Jaya and Shah Alam stations showing higher concentrations of PM_{10} compared to Jerantut station during the haze episodes. This is attributable to the different locations of the air quality monitoring stations, with Klang and Shah Alam being located in an urban area, Petaling Jaya in an industrial area and Jerantut (background station) in a rural area (DOE, 2018c).

TIME DELAY EMBEDDING

Prior to the implementation of persistent homology, the time series data sets must be transformed into point cloud data, which is achieved via the Takens method. The higher dimensional data sets will allow us to look into higher dimensional topological features, such as holes and voids. Basically topological features are features that are invariant after deformations such as stretching, splitting and cutting (Hatcher, 2002; Edelsbrunner and Harer, 2010; Ghrist, 2014). The idea of using a combination of Takens method and persistent homology is discussed in the work of Perea and Harer (2015) and references therein. The Takens method (Takens, 1981) stated that a time series x_0, x_1, …, x_{n-1} can be reconstructed in a phase space of dimension m, where each point in the phase space is given by the vector $x_n(m,\tau) = x_n, x_{n+\tau}, \ldots, x_{n+(m-1)\tau}$, τ is the time delay and m is the embedding dimension. The two parameters τ and m require careful selection to ensure clear extraction of the desired topological features. In this study, a comparison made between the months has indicated that the different settings of time delay and embedding dimension will affect the results. The two parameters have been fixed as $\tau = 1$ and $m = 3$. The time delay is chosen as trivial time delay, whereas the embedding dimension chosen is 3 as the 1-dimensional topological features are best visualized in three-dimensional space throughout the study. This decision is supported by Umeda (2017), whose work has fixed both parameters as $\tau = 1$ and $m = 3$, while others like Khasawneh and Munch (2014) and Khasawneh *et al.* (2018) chose $m = 3$ to visualize the 1-dimensional features. Similarly, previous studies by Pereira and Mello (2015) and Maletić *et al.* (2016) have also fixed m, whereas in other areas of research (not related to persistent homology), Sivakumar (2003) and Sivakumar (2002) each have chosen $\tau = 1$ to achieve better results for their respective research.

PERSISTENT HOMOLOGY

Computations of simplicial complexes are compulsory in extracting topological features from point cloud data. The 0-simplices represent vertices or points, 1-simplices

Table 1. Descriptive statistics for daily average of PM_{10} for the chosen months, August 2005, June 2013, March 2014, September and October 2015.

Month	Statistic	Station			
		Jerantut	Klang	Petaling Jaya	Shah Alam
Aug ’05	Min	35	36	43	26
	Max	205	590	482	587
	Mean	76.12903	139.5161	119.2581	114.8064516
	Std. Deviation	47.34888	138.0956	111.0225	135.0062762
Jun ’13	Min	19	36	20	21
	Max	211	581	370	362
	Mean	56.56667	122	84.26667	83.23333333
	Std. Deviation	43.06826	125.8639	82.18731	81.23940072
Mar ’14	Min	17	47	33	36
	Max	49	448	303	279
	Mean	28.67742	137.9355	94.64516	94.67741935
	Std. Deviation	7.943077	98.63364	65.17134	62.35457059
Sep ’15	Min	35	59	49	49
	Max	242	337	295	301
	Mean	101.8333	141.4333	123.2	135.4
	Std. Deviation	52.76629	69.77584	64.72403	66.17458623
Oct ’15	Min	13	52	24	42
	Max	176	326	320	346
	Mean	75.64516	158.6774	125.5484	147.4516129
	Std. Deviation	49.42034	76.16972	72.68509	78.30829616

represent edges or lines, 2-simplices represent triangles and 3-simplices represent tetrahedra, and so on. A simplicial complex is built by a combination of these simplices (see Fig. 2) accordingly, whereby its complex formation is from data points and thus indicating the dependency of topological features on a scaling parameter (filtration value), ε. Fig. 3(a) in particular shows an example of a simplicial complex formation. The formation commenced at $\varepsilon = 0$, where the four 0-simplices formed represents four points in a point cloud. As the value ε increases to 0.5, four circles are formed with each 0-simplex acting as the center and ε as the radius of the circles. The circles keep growing as the value ε increases until any pair of the circles intersects each other. From this intersection, a 1-simplex is formed by connecting two 0-simplices by a line (edge). As seen in Fig. 3(a), four 1-simplices have been formed at $\varepsilon = 1.4$, and the stage is marked with the formation of a simplicial complex via the combination of the 0-simplices and 1-simplices. The ε value then further increased to 2, with more circles intersecting each other and resulting in the appearance of two 2-simplices (triangles). At this stage, a new simplicial complex is formed (see Fig. 3(a)). It should be noted that the simplicial complex at $\varepsilon = 1.4$ is contained in the simplicial complex at $\varepsilon = 2$, which is also known as filtered simplicial complexes. In this work, the constructions of simplicial complexes are done using Vietoris-Rips simplicial complex, or otherwise also known as Rips complex. Rips complex is a set of k-simplices, such that the distance of any two points in k-simplices is less than or equal to 2ε (Hatcher, 2002; Edelsbrunner and Harer, 2010; Ghrist, 2014).

The birth and death points of topological features are captured depending on the formation of simplicial complexes and subsequently recorded in a diagram known as barcode (Fig. 3(b)). Each feature is represented by a horizontal line in the barcode, with the left endpoint of the line being the birth point and the right endpoint of the line considered as the death point. One can recognize any noise in the barcode by looking at the short lines, while the significant features are signified by the long lines (Ghrist, 2008). Furthermore, the black lines in the barcode represent the connected components, whereas the red line represents the hole (see Fig. 3(b)). Moreover, persistence diagram (Fig. 3(c)) is yet another representation of barcode that serves to summarize the birth and death of topological features in R^2. The x-axis and y-axis of the persistence diagram is the representation of the birth and death points respectively for each of the topological features. Besides, a point (b, d) with multiplicity q in a persistence diagram represents q features that have same birth, b, and death, d, points. Another feature that persists longer through the filtration stage is located way above the diagonal line in the persistence diagram. Additionally, the persistency of a feature is measured by the difference between the death and birth points, $d - b$, and subsequently known as the lifetime of the feature (Otter *et al.*, 2017).

Fig. 3 shows an example of connected components (0-dimensional) and hole (1-dimensional) extracted based on the formation of simplicial complexes. At the starting point, four connected components have appeared with a birth value at $\varepsilon = 0$, and as the filtration value increases the features have remained until 1-simplices appeared at $\varepsilon = 1.4$. It should be noted that the four connected components have collapsed into one at $\varepsilon = 1.4$. Interestingly, a new feature has appeared at this stage, represented by a hole appearance. As the values of ε have continued to increase

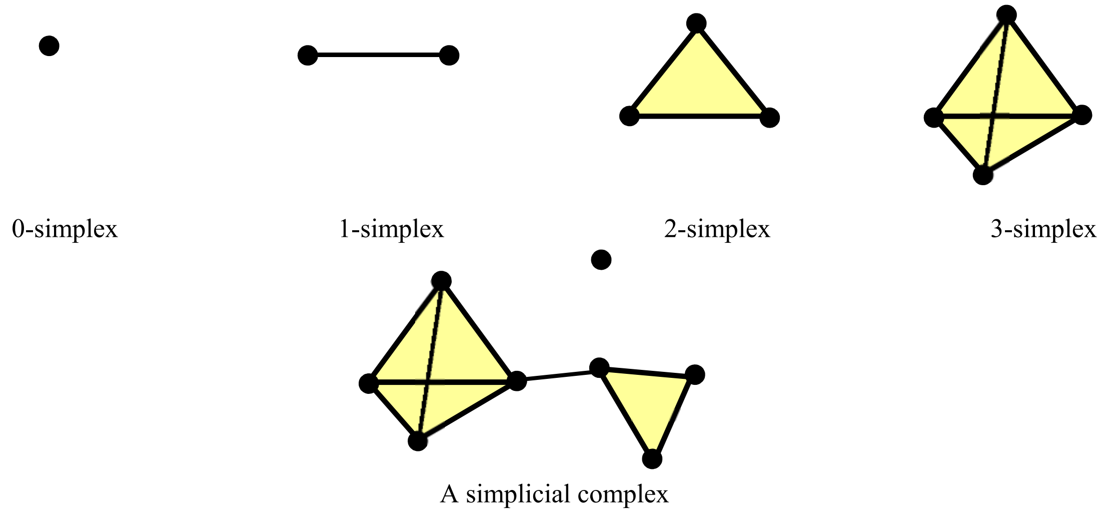
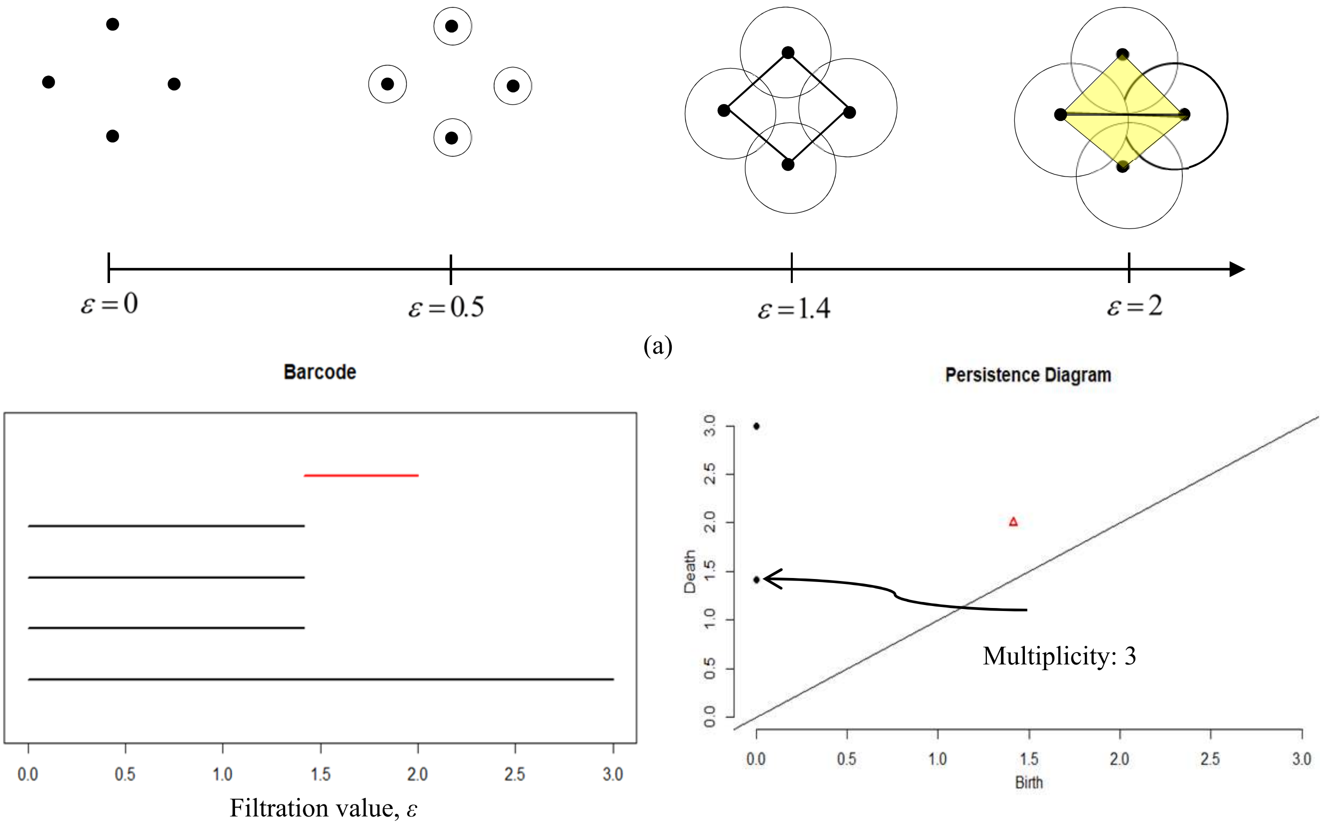


Fig. 2. k-simplices for $0 \le k \le 3$.



(a)

(b)

(c)

Fig. 3. (a) Formation of simplicial complexes with respect to the filtration values, ε. (b) Barcode and (c) persistence diagram for the formation of simplicial complex illustrated in (a). In (b) and (c), the black lines and black dots represent connected components in (a), while red line and red triangle correspond to hole in (a).

to 2, two 2-simplices are subsequently formed and close the hole. This causes the death of the hole and renders only one connected component to persist by the end of the filtration. Persistence diagram in Fig. 3(c) is another representation of these topological features, simplifying the barcode in Fig. 3(b).

SUMMARY STATISTICS OF TOPOLOGICAL FEATURES

A persistence diagram consists of k-dimensional features with 0-dimensional features representing the connected components, 1-dimensional features representing holes, and

2-dimensional features representing voids etc. Meanwhile, a persistence diagram ω_m consists of n features, $\omega_{m_i} = (b_i, d_i)$, with b_i and d_i ($i = 1, 2, \ldots, n$) indicating their birth and death points respectively. All of the features are summarized using summary statistics and described below: The first summary statistic is the sum of all lifetimes (Eq. (1); Pereira and Mello, 2015) of k-dimensional features. If the value of the sum is close to 0, the persistence diagram has practically short-lived features for each particular dimension, k.

$$\mathrm{sum}_k = \sum_{i=1}^{n}\left(d_i - b_i\right) \tag{1}$$

The second summary statistic is the average of all lifetimes (Pereira and Mello, 2015; Mittal and Gupta, 2017) of k-dimensional features, as described in Eq. (2). A small value of average indicates that for a particular dimension, k the data set has mostly short-lived features and vice versa.

$$\mathrm{avg}_k = \frac{\sum_{i=1}^{n}\left(d_i - b_i\right)}{n} \tag{2}$$

Next, the third summary statistic is the maximum of all lifetimes (Pereira and Mello, 2015; Mittal and Gupta, 2017) of k-dimensional features. For each dimension k, the value of the maximum lifetimes of all features is defined as:

$$\max_k = \max_{\omega_{m_i}}\left(d_i - b_i\right) \tag{3}$$

which will determine the most significant k-dimensional feature.

In this study, the summary statistics of 0-dimensional and 1-dimensional features have been calculated using Eqs. (1)–(3). The changes of values in each summary statistics have been observed to track the evolution of the topological features for months with and without haze accordingly.

Remark. We acknowledge that according to Cohen-Steiner *et al.* (2007) and Cohen-Steiner *et al.* (2010), the summary statistic, sum_k, is stable whereas the stability of the avg_k is still in doubt. However, there may be merit in taking into account avg_k as considered by Pereira and Mello (2015) and Mittal and Gupta (2017) but in general it should be done with care. Based on our observations using our data sets, small variation of the input data leads to small variation of avg_k values.

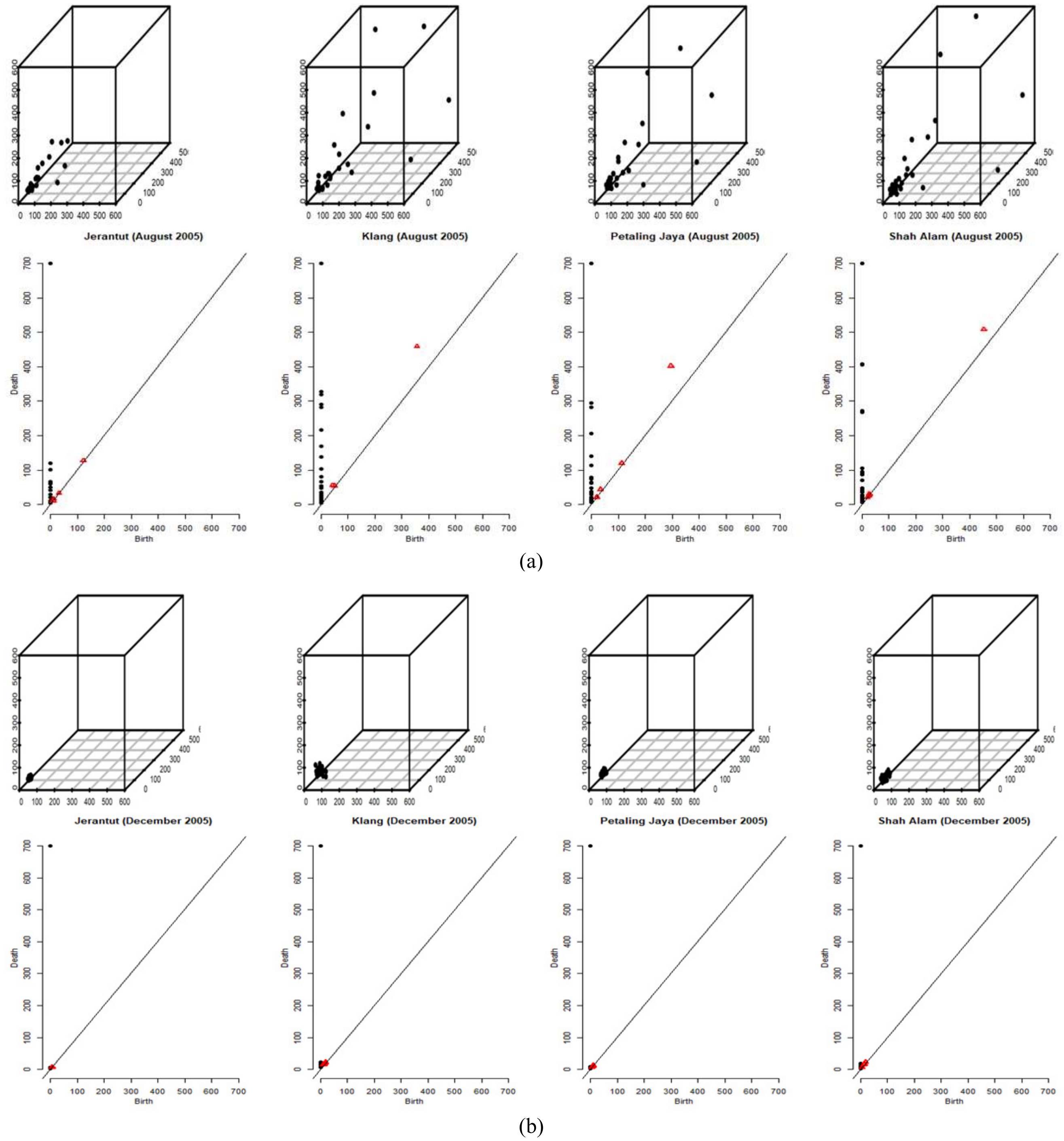
RESULTS

This study has analyzed the time series of daily average for PM_{10} that is partitioned according to month for four air quality monitoring stations located in Jerantut, Klang, Petaling Jaya, and Shah Alam for 16 years (2000–2015). The time delay embedding with time delay $\tau = 1$ and embedding dimension $m = 3$ have been applied for each month of data on PM_{10}, allowing each month to generate a point cloud data. Persistent homology is then applied to each point cloud with maximum filtration value, $\varepsilon_{\max} = 700$. The evolution of topological features based on Rips complexes is next observed in the range of filtration value, $\varepsilon = [0, 700]$, starting from the simplest form which is under-approximation until it displays over-estimation (i.e., one large Rips complex). The computation of persistent homology in this study has been completed using the R-package TDA (Fasy *et al.*, 2017).

All extracted topological features have been represented in persistence diagrams, following which the summary statistics for each persistence diagrams have been calculated. Fig. 4 has displayed a comparison between persistence diagrams generated in August 2005 where severe haze has been reported (DOE, 2018a), in contrast with December 2005 where no haze has occurred. The point clouds generated using time delay embedding with respect to the persistence diagrams are also shown in Fig. 4. The month of December has been selected due to the lesser likelihood for haze to occur during the north-east monsoon season (November–March), as Malaysia receives more rainfall during the period and increases the removal rate of PM_{10} (Yusof *et al.*, 2010).

From the persistence diagrams shown in Fig. 4, the topological features (i.e., black dots represent connected components, red triangles represent holes) for a month with haze episode recorded (i.e., August 2005) is spread away from the origin. In contrast, a month without haze (i.e., December 2005) has shown that the features accumulate close to the origin. For each station, a hole is present among the holes (red triangles) as in Fig. 4(a), which is located the furthest from the origin and has the highest value of birth and death points compared to other holes. It should be noted the corresponding feature in Jerantut is not too far from the origin compared to other stations, implying that the stations in Klang, Shah Alam, and Petaling Jaya have experienced haze of higher severity compared to Jerantut. Summary statistics are calculated to summarize the persistence diagrams, whereby the resulting outcomes are shown in Figs. 5 and 6.

An observation of the various peaks in Fig. 5(a) has revealed drastic increments of the sum of all lifetimes, sum_0, for connected components (0-dimensional features) during haze episodes (rectangles). This is indicative of the connected components persisting longer during haze as the filtration value varied compared to the normal months. This is further elucidated by the average of all lifetimes, avg_0, as shown in Fig. 5(b), which suggests that the persistence diagrams for months with haze consist of mostly long-lived features, hence producing the peaks. Figs. 5(a) and 5(b) also show the sum and average of the lifetimes of connected components extracted in the background station, Jerantut being the lowest compared to other stations. It should be noted that the maximum of all lifetimes are not calculated for connected components as the formation of simplicial complexes (Rips complexes) for PM_{10} persists until it becomes a large simplicial complex with birth, $b = 0$, and death, $d = 700$ (since maximum filtration value, $\varepsilon_{\max} = 700$), resulting in the same values of maximum for all lifetimes ($\max_0 = 700$) according to all selected months.



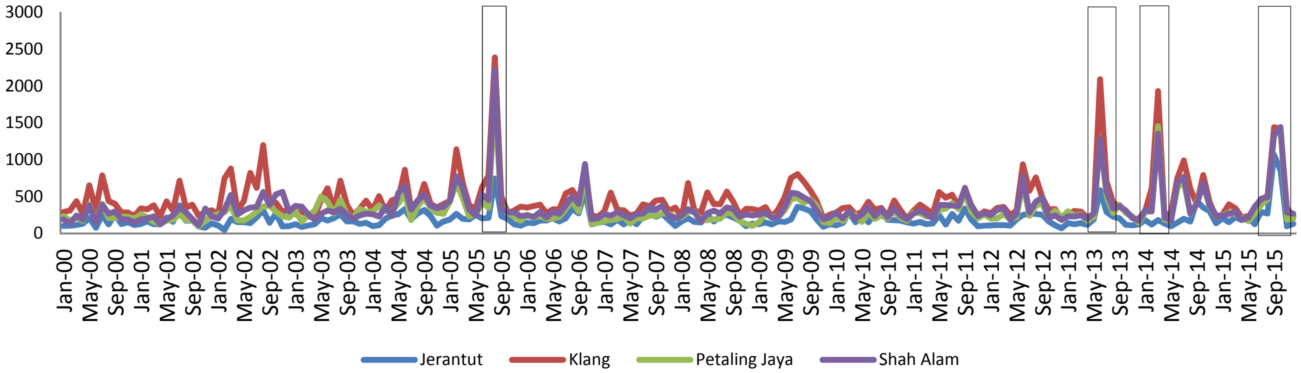
(a)

(b)

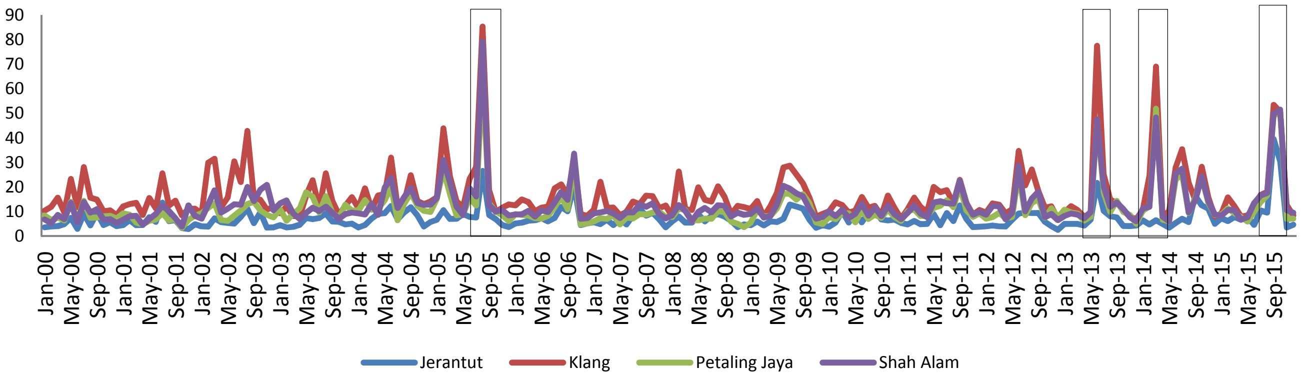
Fig. 4. Point clouds and persistence diagrams for the months with and without haze, (a) August 2005 and (b) December 2005, respectively, for Jerantut, Klang, Petaling Jaya and Shah Alam stations (left to right). The black dots and red triangles in persistence diagrams represent connected components and holes.

Furthermore, the summary statistics for holes (1-dimensional features) are shown in Fig. 6 and revealed drastic changes in the values of the sum, average and maximum of all lifetimes denoted as sum_1, avg_1 and max_1 respectively (see Figs. 6(a)–6c)) during haze episodes (rectangles). As seen in Fig. 6(c), the peaks (in rectangles) are the values of maximum of the lifetimes of all holes which imply that there is a hole with its lifetime being higher compared to other holes in the month that experienced severe haze. Hence, Figs. 5 and 6 have revealed the behavior of the topological features extracted by persistent homology to be consistent with real haze phenomena. When the haze occurred, persistent homology has identified this occurrence by showing the strong rising of topological features lifetimes.

The rectangles in Figs. 5 and 6 indicate the particular



(a)



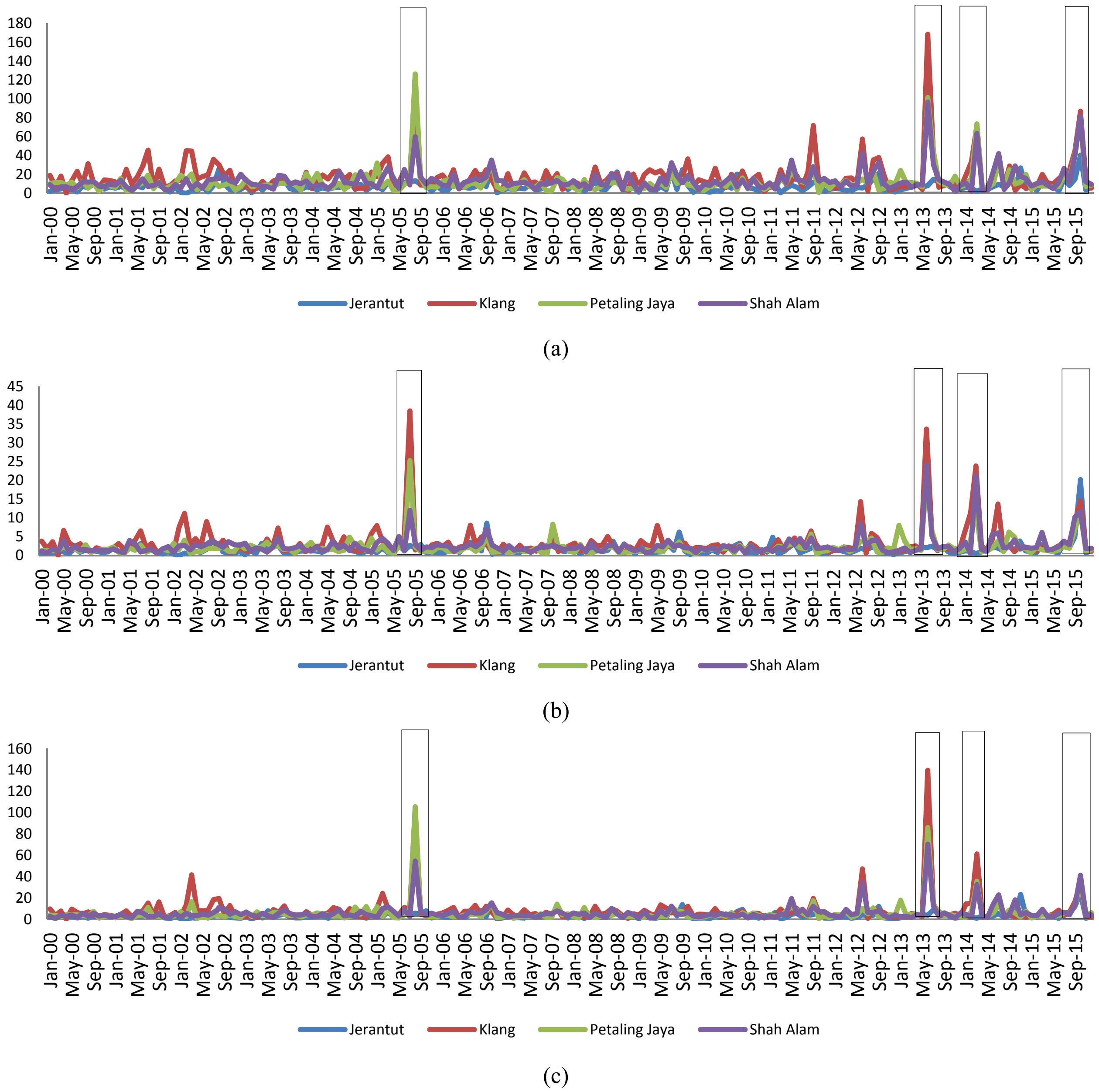
(b)

Fig. 5. (a) sum_0 of all lifetimes and (b) avg_0 of all lifetimes for four air quality monitoring stations, Jerantut, Klang, Petaling Jaya and Shah Alam. The rectangles indicate the months with severe haze episodes, namely August 2005, June 2013, March 2014 and September and October 2015 (left to right).

month recorded with severe haze in the selected years, with a clear and drastic increase in the values of summary statistics in the month. These characteristics have also been shown by each station involved, substantiating the consistency of the topological characterization for months with and without haze accordingly. Overall, the values (i.e., summation, average, maximum) of all lifetimes for the topological features of the background station (Jerantut) are the lowest compared with the others. Thus, it is clear that the topological features extracted via persistent homology are capable of distinguishing the months with severe haze and without haze. While the existing methods (Azmi *et al.*, 2010; Ling *et al.*, 2010; Abdullah *et al.*, 2012; Rahman *et al.*, 2015) have served to directly quantify PM_{10} concentration without providing a qualitative understanding, this paper has successfully filled in research gap by revealing the changes of topological features seen between months with and without haze accordingly.

CONCLUSIONS

This study uses persistent homology, a method of topological data analysis (TDA), to evaluate the topological features embedded within PM_{10} concentration data. The data collected from all four stations involved in this study (Klang, Petaling Jaya, Shah Alam and Jerantut) display changes in topology for the months during which severe haze episodes were recorded. Although existing methods are capable of directly quantifying and analyzing PM_{10} concentration data in order to identify the months affected by haze episodes, this study proposes a new qualitative approach based on topology, which complements other methods by providing robust qualitative structures that filter noise without sacrificing data. The summary statistics of the topological features (viz., connected components and holes) indicate a significant increase in the sum, average and maximum values for the lifetimes of these features during haze episodes for all four stations. Furthermore, the lowest values are exhibited by the summary statistics for Jerantut, characterizing this location as a background site due to its low level of air pollution. Overall, our approach accurately identifies haze episodes by extracting topological features and analyzing their summary statistics, thereby demonstrating its applicability as the basis of an early warning system. We particularly believe that early signals can be detected by refining data analysis, and the progressive efforts spearheaded by this work are making waves in this area. We also recommend that future research address the as-yet undetermined stability of the summary statistic avg_k.



(a)

(b)

(c)

Fig. 6. (a) sum_l of all lifetimes, (b) avg_l of all lifetimes and (c) max_l of all lifetimes for four air quality monitoring stations, Jerantut, Klang, Petaling Jaya and Shah Alam.

ACKNOWLEDGMENTS

The authors would like to express their utmost gratitude to Universiti Kebangsaan Malaysia for the Research University Grant (DIP-2017-011, GGPM-2016-001), Ministry of Education Malaysia Research Grant FRGS/1/2017/STG06/UKM/01/1, and to Department of the Environment (DOE) for providing the air quality data. We also thank the reviewers for their helpful and critical comments.

REFERENCES

Abdullah, A.M., Samah, M.A.A. and Jun, T.Y. (2012). An overview of the air pollution trend in Klang Valley, Malaysia. *Open Environ. Sci*. 6: 13–19.

Afroz, R., Hassan, M.N. and Ibrahim, N.A. (2003). Review of air pollution and health impacts in Malaysia. *Environ. Res*. 92: 71–77.

Azid, A., Juahir, H., Ezani, E., Toriman, M.E., Endut, A., Rahman, M.N.A., Yunus, K., Nordin, M., Kamarudin, M.K.A., Hasnam, C.N.C., Saudi, A.S.M. and Umar, R. (2015). Identification source of variation on regional impact of air quality pattern using chemometric. *Aerosol Air Qual. Res*. 15: 1545–1558.

Azmi, S.Z., Latif, M.T., Ismail, A.S., Juneng, L. and Jemain, A.A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Qual. Atmos. Health* 3: 53–64.

Banan, N., Latif, M.T., Juneng, L. and Ahamad, F. (2013). Characteristics of surface ozone concentrations at stations with different backgrounds in the Malaysian Peninsula. *Aerosol Air Qual. Res*. 13: 1090–1106.

Chan, J.M., Carlsson, G. and Rabadan, R. (2013). Topology of viral evolution. *Proc. Natl. Acad. Sci. U.S.A.*. 110: 18566–18571.

Cohen-Steiner, D., Edelsbrunner, H. and Harer, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* 37: 103–120.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J. and Mileyko, Y. (2010). Lipschitz functions have L_p-stable persistence. *Found. Comput. Math.* 10: 127–139.

Dewoskin, D., Climent, J., Cruz-White, I., Vazquez, M., Park, C. and Arsuaga J. (2010). Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topol. Appl.* 157: 157–164.

DOE (2018a). Chronology of haze episodes in Malaysia. Department of Environment Malaysia. https://www.doe.gov.my/portalv1/en/info-umum/info-kualiti-udara/kroonologi-episod-jerebu-di-malaysia/319123. Last Access: 10 July 2018.

DOE (2018b). A guide to air pollutant index (API) in Malaysia. Department of Environment Malaysia. Enviro Knowledge Centre. https://enviro.doe.gov.my/. Last Access: 10 August 2018.

DOE (2018c). Environmental quality report 2015. Department Of Environment Malaysia. Enviro Knowledge Centre. https://enviro.doe.gov.my/. Last Access: 10 July 2018.

Edelsbrunner, H. and Harer, J. (2010). *Computational topology: An introduction*. American Mathematical Society, Providence.

Elangasinghe, M.A., Singhal, N., Dirks, K.N. and Salmond, J.A. (2014). Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmos. Pollut. Res.* 5: 696–708.

Emrani, S., Gentimis, T. and Krim, H. (2014). Persistent homology of delay embeddings and its application to wheeze detection. *IEEE Signal Process Lett.* 21: 459–463.

Fasy, B.T., Kim, J., Lecci, F., Maria, C. and Rouvreau, V. (2017). Statistical tools for topological data analysis. arXiv: Mathematical Software: Available from https://cran.r project.org/web/packages/TDA/TDA.pdf.

Gameiro, M., Hiraoka, Y., Izumi, S., Kramar, M., Mischaikow, K. and Nanda, V. (2015). A topological measurement of protein compressibility. *Jpn. J. Ind. Appl. Math*. 32: 1–17.

Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bull. Am. Math. Soc.* 45: 61–75.

Ghrist, R.W. (2014). *Elementary applied topology*, Createspace, Seattle.

Gidea, M. and Katz, Y.A. (2018). Topological data analysis of financial time series: Landscapes of crashes. *Physica A* 491: 820–834.

Hamid, N.Z.A. and Noorani, M.S.M. (2014). A pilot study using chaotic approach to determine characteristics and forecasting of PM_{10} concentration time series. *Sains Malaysiana* 43: 475–481.

Hatcher, A. (2002). *Algebraic topology*. Cambridge University Press, Cambridge.

Jorquera, H., Palma, W. and Tapia, J. (2000). An intervention analysis of air quality data at Santiago, Chile. *Atmos. Environ*. 34: 4073–4084.

Khasawneh, F.A. and Munch, E. (2014). ASME 2014 International mechanical engineering congress and exposition 2014, American Society of Mechanical Engineers, Montreal, Canada.

Khasawneh, F.A., Munch, E. and Perea, J.A. (2018). Chatter classification in turning using machine learning and topological data analysis. *IFAC-PapersOnLine* 51: 195–200.

Latif, M.T., Othman, M., Idris, N., Juneng, L., Abdullah, A.M., Hamzah, W.P., Khan, M.F., Sulaiman, N.M.N., Jewaratnam, J., Aghamohammadi, N., Sahani, M., Chung, J.X., Ahamad, F., Amil, N., Darus, M., Varkkey, H., Tangang, F. and Jaafar, A.B. (2018). Impact of regional haze towards air quality in Malaysia: A review. *Atmos. Environ*. 177: 28–44.

Ling, O.H.L., Ting, K.H., Shaharuddin, A., Kadaruddin, A. and Yaakob, M.J. (2010). Urban growth and air quality in Kuala Lumpur city, Malaysia. *Environ. Asia* 3: 123–128.

Makarau, A., Richter, R., Muller, R. and Reinartz, P. (2014). Haze detection and removal in remotely sensed multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* 52: 5895–5905.

Maletić, S., Zhao, Y. and Rajković, M. (2016). Persistent topological features of dynamical systems. *Chaos*: 26: 053105.

Mittal, K. and Gupta, S. (2017). Topological characterization and early detection of bifurcations and chaos in complex systems using persistent homology. *Chaos* 27: 051102.

Otter, N., Porter, M.A., Tillmann, U., Grindrod, P. and Harrington, H.A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Sci*. 6: 17.

Payus, C., Abdullah, N. and Sulaiman, N. (2013). Airborne particulate matter and meteorological interactions during the haze period in Malaysia. *Int. J. Environ. Sci. Dev*. 4: 398–402.

Perea, J.A. and Harer, J. (2015). Sliding windows and persistence: An application of topological methods to signal analysis. *Found. Comput. Math.* 15: 799-838.

Pereira, C.M. and Mello, R.F.D. (2015). Persistent homology for time series and spatial data clustering. *Expert Syst. Appl.* 42: 6026–6038.

Pigott, T.D. (2001). A review of methods for missing data. *Educ. Res. Eval*. 7: 353–383.

Rahman, S.R., Ismail, S.N., Ramli, M.F., Latif, M.T., Abidin, E.Z. and Praveena, S.M. (2015). The assessment of ambient air pollution trend in Klang Valley, Malaysia. *World Environ*. 5: 1–11.

Schwartz, J., Dockery, D.W. and Neas, L.M. (1996). Is daily mortality associated specifically with fine particles?

J. Air Waste Manage. Assoc. 46: 927–939.

Sharma, R. and Balasubramanian, R. (2018). Size-fractionated particulate matter in indoor and outdoor environments during the 2015 haze in Singapore: Potential human health risk assessment. *Aerosol Air Qual. Res*. 18: 904–917.

Sivakumar, B. (2002). A phase-space reconstruction approach to prediction of suspended sediment concentration in rivers. *J. Hydrol*. 258: 149–162.

Sivakumar, B. (2003). Forecasting monthly streamflow dynamics in the western United States: A nonlinear dynamical approach. *Environ. Modell. Software* 18: 721–728.

Takens, F. (1981) Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, Rand D., Young LS. (Eds.), Lecture Notes in Mathematics, vol. 898. Springer, Berlin, Heidelberg, pp. 366–381.

Umeda, Y. (2017). Time series classification via topological data analysis. *Trans. Jpn. Soc. Artif. Intell.* 32: D-G72_1-12.

Wen, Y.S., Mohd Nor, A.F., Fazilan, N.N. and Sulaiman, Z. (2016). Transboundary air pollution in Malaysia: Impact and perspective on haze. *Nova J. Eng. Appl. Sci*. 5: 1–11.

Yu, S., Li, P., Wang, L., Wu, Y., Wang, S., Liu, K., Zhu, T., Zhang, Y., Hu, M., Zeng, L., Zhang, X., Cao, J., Alapaty, K., Wong, D.C., Pleim, J., Mathur, R., Rosenfeld, D. and Seinfeld J.H. (2018). Mitigation of severe urban haze pollution by a precision air pollution control approach. *Sci. Rep.* 8: 8151.

Yusof, N.F.F.M., Ramli, N.A., Yahaya, A.S., Sansuddin, N., Ghazali, N.A. and Madhoun, W.A. (2010). Monsoonal differences and probability distribution of PM_{10} concentration. *Environ. Monit. Assess.* 163: 655–667.

Zeb, B., Alam, K., Sorooshian, A., Blaschke, T., Ahmad, I. and Shahid, I. (2018). On the morphology and composition of particulate matter in an urban environment. *Aerosol Air Qual. Res*. 18: 1431–1447.

Zhao, X., Qi, Q. and Li, R. (2010). The establishment and application of fuzzy comprehensive model with weight based on entropy technology for air quality assessment. *Procedia Eng*. 7: 217–222.

Received for review, October 8, 2018
Revised, May 1, 2019
Accepted, May 28, 2019