Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Data Mining (CIDM 2007)

Noise Reduction Approach for Decision Tree Construction: A Case Study of Knowledge Discovery on Climate and Air Pollution

Kyoko Fukuda

*Environmental Science Program, Dept. of Mathematics
and Statistics, Computer Science and Software
Engineering
University of Canterbury
Private Bag 4800, Christchurch, New Zealand
kfu11@math.canterbury.ac.nz*

Abstract – **Data mining is more effective on noisy time series with appropriate data pre-processing. Singular Spectrum Analysis (SSA) is explored as the noise reduction approach for a decision tree classifier for noisy data. SSA provides groups of additive components, from low to high frequency, by decomposing the noisy time series. In this study, the noisy climate data is decomposed by SSA and is used to construct decision trees to predict the carbon monoxide (CO) air pollution levels. Analysis shows that separating out seasons from the annual data helps the algorithm; the classification accuracy improvements vary by season, with the maximum improvement (from 60.7% to 77.3%) found in summer by removing 6.42% of the high frequency signals, while autumn showed no improvement. Examining decision tree structures provides threshold climate values that impact on different CO levels, e.g., a light wind speed of ≤ 2.5 m/s and any level of temperature inversion formation is found to associate with the high CO level (> 0.70 mg/m^3). Overall, data pre-processing using SSA is encouraging to improve the results of any time series data mining approach. Examining decision trees of the climate and air pollution helps increase knowledge about the data, and the studied approaches can be adaptable for various future environmental studies.**

Index Terms – **Air pollution; climate; decision trees; Singular Spectrum Analysis.**

I. Introduction

The causal effect of air pollution on human and environment health is a worldwide problem. Air pollutant levels even below standard concentrations are known to affect human health, with increases in respiratory symptoms, chronic cough, bronchitis and chest illness, and deterioration in pulmonary function [1].

The study area, Christchurch, in New Zealand (with a population of about 334,000 and an area of 452 km^2) suffers from a serious winter air pollution problem due to domestic heating, e.g., burning wood and coal, and poor air dispersion due to a combination of winter weather and its topographic factors, primarily a medium sized hill located adjacent to the city, which traps air pollutants in a temperature inversion layer; see details in [2]. The main winter air pollutants are CO from domestic heating and motor vehicles, PM from domestic heating, SO_2 from industry and NO_2 (a product of the oxidation reaction of NO) from motor vehicles [3]. Recent investigation of particulate matter of diameter below 10 μg/m^3 (PM_{10}) and the acute respiratory morbidity rate in the study area reports that low PM_{10} levels (less than 10 μg/m^3) can even impact on different age ranges, in particular, very young (under five years) and older ages (55 years and over), and its association varies between female and male and by season [4]. Short and long-term air pollution levels are affected by changes in local climate and global climate [5].

In recent years, data mining, a process of knowledge discovery in databases (KDD), is also found to be a useful tool among environmental scientists [6, 7], due to its flexibility to handle problems in environmental systems, which are often ill-structured and non-linear domains [7], and involve multidisciplinary factors, e.g., global and local ecological, social and economical factors. To investigate the air pollution and climate data set that is generally noisy and skewed, a primary step is to reduce the noise, although determining the noise component of such a noisy and skewed structure can be difficult. Attribute selection can be used to remove the outliers as a data pre-processing step, but it may lose the time sequence, as the air pollution and climate time series are associated, day-to-day. Hence, smoothing methods are ideal. Generally, Generalized Additive Models (GAMs) [8], a statistical method for smoothing non-linear time series, is commonly applied to identify response-predictor relationships. Recently, reference [9] used the wavelet transform as a data pre-processing step to extract the trends of air pollution levels in order to apply further neural network models, since data mining algorithms with pre-processed data sets generally work efficiently to provide improved results [9, 10].

In this paper, two investigations are carried out. Firstly, Singular Spectrum Analysis (SSA) is introduced as the noise reduction approach for the data pre-processing method to apply a data mining technique, a decision classifier (J4.8 from WEKA [11]). The noisy climate time series is decomposed and separated out from noise by SSA to form several additive components, which are used to construct decision trees to predict the different air pollution levels of carbon monoxide

(CO). Decision tree classification accuracy is then examined to see how SSA helped the algorithm. Secondly, the decision trees obtained are examined in detail to provide threshold climate values that impact on different CO levels. The investigation helps support knowledge on the cause and effect relationship of climate and air pollution profile.

A. Singular Spectrum Analysis in data mining

Singular Spectrum Analysis (SSA) is an innovative model-free nonparametric method of time series analysis, a mixture of mathematical and statistical analyses: namely classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing [12]. For example, SSA has been applied to digital signal processing [13], oceanographic research [14]. Recently it has been applied to an air pollution study [15], and SSA decomposed structures have been applied to data mining techniques for image segmentation [6, 16]. However, in this study, SSA is used, for the first time, for data pre-processing to help the data mining algorithm by removing noisy structures from the data set.

Using SSA provides two benefits. Firstly, the decomposed structures help improve the results of the tree construction algorithm. Secondly, SSA helps identify noise in the structures, as it decomposes the noisy time series into several *additive* components – separating out several high and low frequency signals from the original time series – that can be grouped and are reconstructed to form the new time series. Note that the signals obtained by SSA decomposition differ from those obtained by filtering out frequency bands with the Fourier transform, as they are generated from eigenvectors and as such are not purely related to frequency. This allows exploring by adding or removing such *additive components* (low to high frequencies) to construct the decision tree. During this process, it identifies *which* components can potentially be noise, and the improvement can be examined by the classification accuracy. For example, adding insignificant components (generally high frequency) to the main structures (low frequency) can lower or have no influence on the classification accuracy. On the other hand, removing significant components (including some high frequencies) may lower the classification accuracy, which suggests that these components are unlikely to be noise.

B. Knowledge discovery for climate and air pollution

Extracted decision trees with high classification accuracies are investigated to understand the cause and effect relationship between climate and air pollution levels. This is carried out by examining *how* the decision pathway of climate attributes contributes for changes in air pollution levels, such as *which* climate variables influence air pollution levels, to *what* degree. Note that this study aims to provide knowledge from examining the decision trees via a data mining tool rather than providing prediction rules, to be used to predict air pollution levels in an unknown data set. This is because the studied data set is not large enough to demonstrate accurate prediction rules, but it can be at least used as a knowledge discovery tool.

To enhance the relationship between climate and air pollution level, decision trees are generated from the training data sets of SSA components that are each made up of a single season (dividing the annual data set into four seasons) and the annual data set (all seasons), to compare how the decision pathways of climate influence air pollution levels differently as well as differences in the classification accuracy among different seasons.

II. Methods

A. Singular Spectrum Analysis

SSA decomposes six climate time series to provide input data sets (several additive components) for further data mining application. The SSA procedure has four steps [12, 15]. The first step is embedding, which transforms the original one dimensional time series,

$$F = (f_t) = (f_1, \dots, f_{N_t}) \,, \tag{1}$$

into an *L*-dimensional series,

$$X_i = \left(f_{i-1}, \dots, f_{i+L-2}\right)^T, \tag{2}$$

where $1 \le i \le K = N - L + 1$ and L is the window length (≤ *N*/2). The embedding process turns the one-dimensional time series *F* into the *L*-trajectory matrix,

$$X = [X_1 : \dots : X_K] \,. \tag{3}$$

which can be rewritten as,

$$X = (x_{ij})_{i,j+1}^{L,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{pmatrix}. \tag{4}$$

Note that the matrix *X* is a Hankel matrix, which has equal elements on the diagonals (*i*+*j* = *const.*).

The second step of SSA is to decompose the obtained trajectory matrix by the singular value decomposition (SVD). Let U_1, …, U_L represent the corresponding orthogonal eigenvectors of the matrix $S=XX^T$. Then denote V_i as the eigenvector of *S*, which corresponds to the eigenvalue λ_i for *i* = 1, ..., *d*, where *d* is the number of nonzero eigenvalues (*d* < *L* and 1≤ *i* ≤ *d*),

$$V_i = \frac{1}{\sqrt{\lambda_i}} X^T U_i \,, \tag{5}$$

then the result of the *SVD* of the trajectory matrix, *X*, becomes

$$X = X_1 + \cdots + X_d \,, \tag{6}$$

where

$$X_i = \sqrt{\lambda_i} U_i V_i^T \,. \tag{7}$$

The i^{th} eigentriple (ET) is constructed from the three variables in (7) that make up X_i: the singular value (the square root of the i^{th} eigenvalue) and two orthogonal vectors, the i^{th} right (V_i) and left (U_i) singular vector of the trajectory matrix.

Note that each ET has a different variance, and the sum of the variances for all ETs is 1. These are the additive components.

In the third step, similar ETs are grouped together. It is important to combine appropriate ETs, or in other words, keep the components as similar as possible, rather than mixing dissimilar components, e.g., mixing low and high frequencies, as it decreases the quality of the results in reconstructing the new time series (step 4). In this study, the ET grouping procedure is performed computationally using FastGrouping, a separately developed program that uses Fourier expansion to determine ET similarity [15]. The Fourier expansion [12, 15] provides a correlation coefficient, $\rho_{1,2}$, which is calculated from the cross power of the two series, $F = F_{(1)}+F_{(2)}$, obtained from different ETs in (8);

$$\rho_{1,2} = \sum_{k=0}^{N} \sqrt{\Pi_{f_1}^{N}(k/N)}\sqrt{\Pi_{f_2}^{N}(k/N)} \leq \left\|F^{(1)}\right\|\left\|F^{(2)}\right\|. \tag{8}$$

The normalized form of equation (8) is

$$\rho R_{1,2} = \frac{\sum_{k=0}^{N} \sqrt{\Pi_{f_1}^{N}(k/N)}\sqrt{\Pi_{f_2}^{N}(k/N)}}{\|f_1\|\|f_2\|}, \tag{9}$$

and the magnitude of $\rho R_{1,2}$ indicates the similarity of the spectra of the relevant two signals [15, 16]. Each eigentriple is successively paired with every other eigentriple, and for each pair of eigentriples, the value of $\rho R_{1,2}$ is computed in (9) using the pair of eigenfunctions (eigenvalues and eigenvectors) as $F_{(1)}$ and $F_{(2)}$. It is then computed again using the pair of principal components. Averaging the resulting two $\rho R_{1,2}$ values provides a single metric, which improves the fsensitivity. This provides more reliable results than when the eigenfunctions and the principal components are considered separately. Next, the $\rho R_{1,2}$ value is compared with a threshold (between 0.50 and 0.90) and the two eigentriples are placed in the same group if the metric is greater than the threshold. Lowering the threshold provides fewer ET groups, grouping ETs less accurately, and raising the threshold gives the opposite. Generally, a threshold between 0.70 and 0.85 is recommended.

The fourth and final step is called diagonal averaging. It is a linear operation for reconstructing time series from the additive components and ET groups that are chosen in step 3,

$$F = F_1 + \dots F_m, \qquad 1 \leq i \leq \mathrm{m}. \tag{10}$$

Each of the six climate time series is decomposed by SSA into a number of additive components (each constructed from a single ET or a group of ETs, and of the same length as the original time series, *F*), which are used to generate decision trees as follows.

B. *Noise reduction using SSA for data mining*

Investigating the effectiveness of using SSA for data preprocessing for data mining, a decision tree classifier is applied on climate attributes to predict three CO levels (high, H; medium, M; and low, L), and the classification accuracy is used to assess the improvement. Results are compared for the original time series (without the SSA data processing) and the SSA additive component time series. From each time series, the full length of the time series is divided into four seasons to compare the annual data set (full data set) and seasonal data sets. Hence, the following procedure, generating a decision tree, is repeated for a total of five data sets (one covering the whole year, and one for each of spring, summer, autumn and winter), for the original and each of the additive component time series.

Each data set is divided into three parts, and three training and three test data sets are created. For example, the first training data set will consist of the first two thirds of the data set, and the first test set will consist of the remaining third. Thus, three distinct training and test data sets are created. A decision tree classifier, J4.8 from WEKA [11], based on the C.4.5 algorithm [17], is used to generate a decision tree from each training data set, and is tested on the test data set to provide a classification accuracy. The average and standard deviation (SD) of the three classification accuracies obtained from three test data sets are used for the results.

The specific procedure (repeated for each training set) for generating decision trees for experimenting with the noise reduction method via additive components is as follows. Firstly, a decision tree is generated from a single data set, which covers a full year or a single season. Secondly, decision trees are generated from a data set for each additive component, first removing the structures for ET151-180 from the rest (ET1-150), and increasing the range of eigentriples removed until reaching ET3-180, leaving only ET1 and ET2 (ET1 is kept to provide a base for the components). Hence, six experiments are repeated to generate six single decision trees for each of the five data sets (the full and seasonally divided data set). Note that the experiment starts from removing the 8th additive component (ET151-180 in Fig. 3-H), and the 1st additive component (ET1 in Fig. 3-A) is not removed.

C. *Knowledge discovery from decision trees*

To introduce the outcome of applying the data mining technique on climate and air pollution, the decision trees are examined in detail to increase knowledge about the cause and effect relationship of climate and CO levels by investigating the decision pathways, such as which climate attribute is most responsible for the high CO level. Note that investigations in this paper are carried out by examining the decision tree with the best classification accuracy out of each group of three training data sets. Also to simplify results, the decision pathway is focused and summarised on only the *high* CO level, while the decision trees classify CO into three levels (H, M and L); results on M and L are not described here. To contrast seasonal climate impacts on the high CO level, examination of decision trees is focused on seasonally divided data sets, thus the full data set is not interpreted.

III. DATA SETS

Four years (October 1998 – September 2002) of air pollution and climate daily measurements were provided by an Environment Canterbury (ECan) air pollution monitoring station, located in a residential area, Coles Place, in Christchurch. Six climate measurements are used as input attributes to predict the CO levels: relative humidity (RH in %), temperature measured at 1m above the ground (TG in C°) and at 10m above the ground (TT), the temperature difference (TD = TG-TT), wind speed (WS in m/s), and wind direction (measured in degrees: 0° and 360° for north, 90° for east, 180° for south, 270° for west). Negative values of TD (Fig. 2) indicate the formation of a temperature inversion, which traps air pollutants under a layer of warmer air. The CO levels are categorised into three levels based on the lower and upper quartile (LQ and UQ), since its distribution is rightward skewed; low (L) ≤ 0.14 mg/m^3 at LQ, medium (M) ≤ 0.70 mg/m^3, and high (H) > 0.70 mg/m^3. Fig.1 and Fig. 2 show the original CO and climate time series respectively. Fig. 2 shows six climate attributes from left to right, RH to Wdir along the *x*-axis, where each climate attribute time series covers, from left to right, October 1998 to September 2002. Generally, all time series are noisy, except CO, TG, TT and TD, which show reasonably strong seasonal structures with some high frequencies (Fig. 1 and 2). Note that all data were scaled by dividing each value by the maximum in order to improve the ease of comparison between the SSA results of the climate and air pollution data.

IV. RESULTS AND DISCUSSIONS

A. Extraction of additive components

In this study, a window length, *L*, of 30 (~one month) was selected, as it was one of the dominant frequencies of the air pollution time series. FastGrouping with a threshold of 0.85 provided a number of ET groups. Eight heterogeneous ET groups (and variances, shown as percentages, in brackets) were extracted as input data sets for the data mining application, shown in Fig. 3-A to H respectively: ET1 (90.2%), ET2 (2.72%), ET3 (0.68%), ET4-40 (4.14%), ET41-80 (1.46%), ET81-126 (0.70%), ET127-150 (0.11%), and ET151-180 (< 0.01%). Fig. 3 shows six climate attribute ETs of RH to Wdir in the same manner as Fig. 2. Generally, the first three additive components (ET1 to ET3) hold important structures. The first eigentriple, ET1, which is made by the lowest frequency, has a large variance, describes the general trends, and provides the base structure, so it is always added to the other ETs. ET2 and ET3 generally describe the seasonal structure and change points, or structural changes respectively [5, 15]. The ETs after ET4 are made by high frequencies with reasonably small variances, thus they are grouped with similar components to form larger components. Note that these structures are generally not used, if the purpose of the study was to extract the smooth time series (see details in [5, 15]).

B. Comparison of classification accuracies

Table 1 shows summary results of decision tree classification accuracy (in %) using the original climate time series and different grouped additive components (ETs) based on the full and seasonally divided data sets. Successively larger numbers of ET groups are removed from Case 1 (removing < 0.01% of the entire structure) to Case 6 (removing 7.10% of the entire structure) in Table 1. Table 1 also shows the proportion (in %) of each CO level; H, M, and L, as a brief indicator. Generally, application of the decision tree classifier is better than simply guessing the CO levels, if its score is better than this number. Table 3 shows the confusion matrix for the best classification accuracy within the full data set and each seasonally divided set. Note that the number of instances is the sum of three test data sets.

Dividing the full data set (annual) into each season shows the different classification accuracies among seasons. The mean and standard deviation (SD) of the original and all cases in Table 1 show that the highest classification accuracy is found from winter (76.4±5.7%), which is higher than applying the full length of the data (66.7±2.7%). In fact, the average classification accuracy of the seasonally divided data (67.8±2.7%) is found to be slightly higher than simply applying the full data set. This suggests that even though the sample size has became one fourth of the full length of the data, separating out seasons in environmental data sets that have seasonality successfully helps the algorithm by highlighting relevant characteristics. On the other hand, the lowest classification accuracies compared with the full data set are found from spring (61.7±3.1%) and summer (63.5±8.0). This may due to the low proportion of high CO levels (9.3% and 0.6%) in the spring and summer data, although the reason for the higher classification accuracy in summer compared to

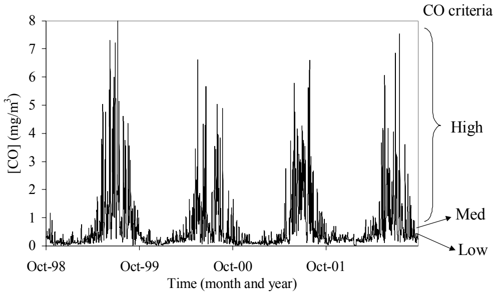


Fig. 1 Original time series of CO.

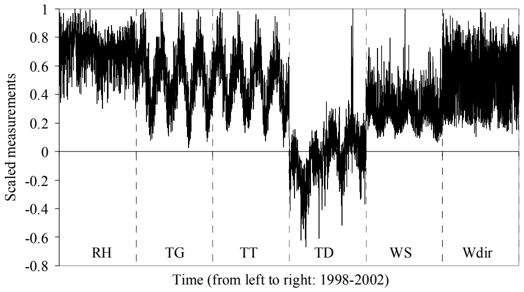


Fig. 2. Original time series of six climate attributes.

spring may be that the summer data set consists of almost half M (44.9%) and half L (54.6%), as it only predicts either M or L, which may confuse the algorithm less, compared with the spring data set.

For generating decision trees, removing some structures (up to 7.10%) from the original time series has shown some improvement over the original time series, although it varies by seasons and different additive ETs (Table 1). For each series, the classification accuracy peaks after a certain number of high frequencies have been removed. This point can be used to identify which structures are significant to capture the best decision trees.

The original times series and Table 1-Case 1 shows the same classification accuracy for all data sets. Removing such small and high frequency structures, ET151-180 (< 0.01% in Fig. 3-H), would not influence the classification accuracy, and may not help the algorithm. This suggests that ET151-180 structures can be noise or insignificant.

Table 1-Case 5, the summer data, shows that the most significant classification accuracy improvement. Removing ET4-180, a total of 6.42% of the structure (in Fig. 3-C to H) shows the accuracy is 77.3%, up to a 16.6% improvement compared with the original time series classification accuracy (60.7% in Table 1). As previously mentioned, the average summer classification accuracy was the lowest. However, the summer data set is made up almost completely of two CO levels, M and L, and it contains fewer outliers and high pollution levels compared with winter. Hence, removing most of the high frequencies (including potential noise) or outliers that are obtained from 6.42% of the structures, ET4-180, may help the algorithm. Most of the classification errors are between M and L; 63 instances of L and 16 instances of M were misclassified as M and L respectively, but these errors are greatly reduced compared to the original data set, where 95 and 46 instances of L and M were misclassified respectively (see Table 2; original and Case 5, summer).

The spring data set (Table 1-Case 6) also shows similar findings, but the classification accuracy is lower (66.2%) than summer, and the improvement was 4.7% compared with the original time series (61.5%). However, the improvement for the spring data set is obtained from removing 7.10% of the structures, ET3-180. Note that the spring data set contained about 9.3% high CO levels (Table 1), although the algorithm works better with dominant low frequencies, ET1 and ET2 (Fig. 3-A and B), describing the seasonal oscillation by removing most of the high frequencies. However, interestingly, the use of only low frequencies improves the misclassification between M and H; 4 instances of H were misclassified as M, compared to 11 for the original time series (Table 2; original and Case 6, spring).

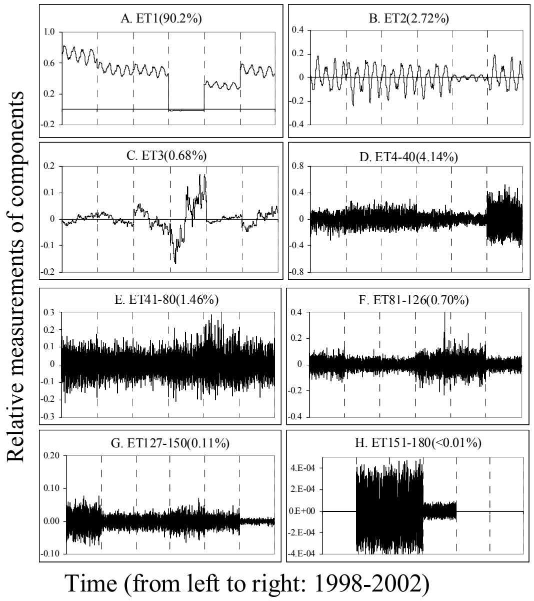


Fig. 3. Eight SSA climate additive components made by ET groups. Note that the dotted lines indicate a single climate attribute, RH, TG, TT, TD, WS and Wdir from left to right. The variance of each ET group is shown in brackets.

The winter (Table 1-Case 2) and full data set (Table 1-Case 5) classification accuracies (83.4% and 70.8% respectively) show about 3% improvement by removing 0.12% of high frequencies, ET127-180, and removing 6.42% of structures, ET4-180, compared with the original time series (80.1% for winter and 67.8% for the full data set). While the winter data set contains many high CO data points (60.9%), removing further high frequencies that are obtained after ET81 (Table 1-Case 3) decreases the classification accuracy, as it may remove truly high CO levels, which should not be considered as outliers. However, an interesting point is that removing high frequency eigentriples with very small variance, e.g. ET127-180, with 0.12% (Table 1-Case 2) shows an improvement, increasing the classification accuracy by about 3%. This may suggest that ET127-150 may be potential noise. From this improvement, the correct classification for M is increased from 106 to 120 instances (Table 2; original and Case 2 in winter). However, no correct classification for L is observed, which needs further investigation. The full data set shows higher classification accuracy as more high frequencies are removed up to ET4-180. Since the full data set lacks characteristics compared with seasonally divided data sets, removing all frequencies except the general trend (ET1), seasonal components (ET2) and change points (ET3) provides smoothed but detailed time series structures, that help to generate the decision tree with the best classification accuracy. The major classification improvement resulted from increasing the number of correctly classified M instances from 498 (original series) to 539 by decreasing misclassification between L and M (Table 2; original and Case 5).

An interesting observation is seen from the autumn data set. Removing any components did not change the classification accuracy, although removing 0.12% of ET127-180 (Fig. 3-G and H) kept the same classification accuracy as the original time series (71.2%). Therefore, these structures could be considered as potentially insignificant noise that can be eliminated even without changing the structures, and removal of these may or may not help the algorithm, as the variance of these structures are very small (0.12%). However, the differences between the original and removing ET127-180 (Table 2-Case 2) is that removing ET127-180 improves detection of H, increasing correctly classified instances from 66 to 72, but it decreases the correctly classified instances of M from 185 to 177. This point needs further investigation (Table 2-Case 2 for winter).

Overall, removing more high frequencies from the original time series improves the classification accuracy for spring, summer and the full data set. In particular, spring and summer contain fewer high levels of CO, removal of high frequencies such as potential noise, outliers or insignificant signatures helps the algorithm efficiently. For example, the maximum classification accuracy improvement was 16.6% for summer by removing 6.42% of the structures (most of the high

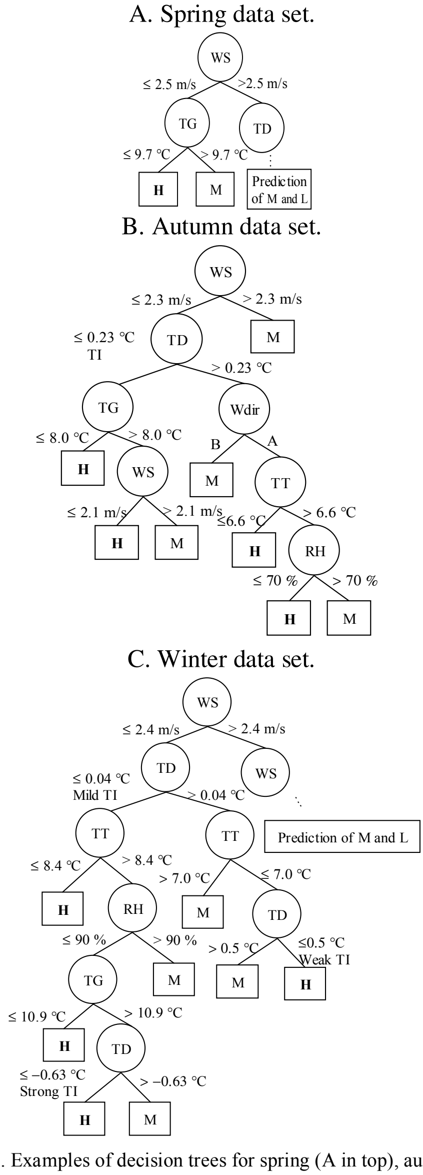
TABLE 1. Summary of decision tree classification accuracy using different SSA decomposed components.

(%)	Spring	Summer	Autumn	Winter	Average of single season	Full data set (all seasons)
Original proportion of CO levels						
H	9.3	0.6	28.5	60.9	24.8	25.0
M	59.3	44.9	58.7	36.4	49.8	49.8
L	31.3	54.6	12.8	2.7	25.3	25.2
Original time series						
C.A.	61.5	60.7	**71.2**	80.1	68.4	67.8
SD.	5.0	4.7	3.7	3.3	1.0	7.9
Case 1. Removing ET151-180 (<0.01%) in Fig. 3 - H from the rest (= adding G. ET127-150 on ET1-80)						
C.A.	61.5	60.7	71.2	80.1	68.4	67.8
SD.	5.0	4.7	3.7	3.3	1.0	7.9
Case 2. Removing ET127-180 (0.12%) in Fig. 3 - G and H from the rest (=adding F. ET81-126 on ET1-80)						
C.A.	62.4	59.5	**71.2**	**83.4**	69.1	66.9
SD.	4.1	4.5	1.6	1.3	2.5	9.3
Case 3. Removing ET81-180 (0.82%) in Fig. 3 - F to H from the rest (=adding E. ET41-80 on ET1-40)						
C.A.	60.7	55.7	69.0	79.3	66.2	65.9
SD.	1.2	6.5	7.2	3.0	2.3	8.9
Case 4. Removing ET41-180 (2.28%) in Fig. 3 - E to H from the rest (=adding D. ET4-40 on ET1-3)						
C.A.	56.0	58.5	67.1	68.2	62.5	62.0
SD.	2.8	1.7	2.2	4.2	2.7	5.3
Case 5. Removing ET4-180 (6.42%) in Fig. 3 - D to H from the rest (= adding C. ET3 on ET1-2)						
C.A.	63.5	**77.3**	68.7	72.8	70.6	**70.8**
SD.	1.6	0.5	0.6	2.2	0.5	5.1
Case 6. Removing ET3-180 (7.10%) in Fig. 3 - C to H from the rest (= adding B. ET2 on ET1)						
C.A.	**66.2**	72.0	68.5	70.9	**69.4**	65.8
SD.	2.5	1.0	1.1	2.0	1.6	2.2
Mean of the original and all cases within the same season	61.7	63.5	69.6	76.4	67.8	66.7
SD of the original and all cases within the same season	3.1	8.0	1.6	5.7	2.7	2.7

TABLE 2. Comparison of the confusion matrices between the original time series and the high frequency separated SSA additive components for all data sets. Note that the total number (sum of all three test results) of instances is shown. Numbers in bold indicate correctly classified instances.

Spring				**Summer**				**Autumn**				**Winter**				**Full data set**			
Original time series																			
	H	M	L		H	M	L		H	M	L		H	M	L		H	M	L
H	**12**	11	1	H	**1**	0	0	H	**66**	15	0	H	**189**	26	0	H	**276**	67	2
M	22	**143**	44	M	1	**67**	46	M	39	**185**	36	M	35	**106**	10	M	86	**498**	150
L	0	62	**69**	L	0	95	**151**	L	0	16	**11**	L	0	2	**0**	L	3	163	**216**
Case 6 (Removing ET3-180)				**Case 5 (Removing ET4-180)**				**Case 2 (Removing ET127-180)**				**Case 2 (Removing ET127-180)**				**Case 5 (Removing ET4-180)**			
	H	M	L		H	M	L		H	M	L		H	M	L		H	M	L
H	**10**	4	2	H	**0**	1	0	H	**72**	24	0	H	**187**	13	0	H	**261**	102	9
M	21	**154**	35	M	1	**98**	16	M	33	**177**	34	M	37	**120**	10	M	99	**539**	124
L	3	58	**77**	L	1	63	**181**	L	0	15	**13**	L	0	1	**0**	L	5	87	**235**

frequencies), compared to non pre-processed data. On the other hand, removing only a very small amount of high frequency information, the ET127-180 structures (0.12%), improved the winter classification accuracy by 3%, compared to non pre-processed data. The autumn data set did not show any particular improvement, and it may require further investigation. The use of SSA additive components as inputs for generating decision trees may have future use for any noisy time series, as this provides better classification accuracy for some parts of the data, which can be helpful for the overall analysis. It also allows exploring the data set.


A. Spring data set.

B. Autumn data set.

C. Winter data set.

Fig. 4. Examples of decision trees for spring (A in top), autumn (B in middle) and winter (C in bottom). Note that all decision trees shown here focused on the high CO level, otherwise branches for predicting medium and low CO levels are not shown. The autumn decision tree (B) shows the wind direction as A (NE, E, SE direction) and B (S, SW, W, NW direction).

C. *Knowledge discovery from decision trees*

Fig. 4 shows the highest performing decision tree out of the three training data sets for each of spring, autumn and winter (with respective accuracies of 68.6%, 73.0% and 84.6%). Note that the investigation is focused on climate responses or impacts on the high CO levels. The summer decision tree did not indicate the pathway of the high CO level. Also the interpretation is based on major findings.

Since the input climate attributes are numerical, the decision trees have numerical threshold values. Note that the autumn decision tree (Fig. 4-B) shows wind direction (Wdir), which takes the value of A for easterly and southeasterly wind and B for southerly, southwesterly and westerly wind. The dominant wind direction is southeasterly, followed by southerly, southwesterly, easterly, and westerly. As previously mentioned in Section III, negative values of TD indicate the formation of temperature inversion (TI). The winter decision tree (Fig. 4-C) shows three TD nodes; the lowest (most negative) TD value suggests a strong TI (≤ -0.63 °C), whereas smaller (≤ 0.04 °C) and larger (≤ 0.5 °C) positive TD values suggest the mild and weaker TI formation.

The winter decision tree has the largest tree size (*TS*=29) of all the trees and the highest number of leaves (*NL*=15), suggesting that the decision process for the winter CO level is most complicated, whereas the spring decision tree has the simplest and smallest tree (*TS*=11 and *NL*=6), and the autumn decision tree (*TS*=15 and *NL*=8) lies between the spring and winter trees.

Common climate responses to the high CO level are found. The most important climate factor (found at the root of the tree) is WS with the value of ≤ 2.3-2.5 m/s (the threshold varies between seasons). The mean and standard deviation of wind speed in the study area are 2.60±0.97. Hence, when the wind speed is lower than the mean (light wind speed) the CO level is high. The second most important climate variable is TD. The autumn data set has milder TI formation (≤ 0.23 °C) than winter, as generally TI is often observed more in winter with lower temperatures. Three different levels of TI (strong, medium and weak) also associate with the high CO level. However, the spring decision tree shows the association of TD is more with M and L (Fig. 4-A). In spring, the ≤ 9.7 °C TG is responsible for the high CO level instead. Interestingly, only the autumn decision tree uses the wind direction attribute; southeasterly direction associates with high CO level via lower TT (≤ 6.6 °C), but when TT is above 6.6 °C with lower humidity (≤ 70%; dryer air), the association of the high level is detected (Fig. 4-B). A similar finding is found from winter (Fig. 4-C). The association of the high CO level is: during mild TI, via lower TT (≤ 8.4 °C); during dryer relative humidity (≤ 90%), via colder TD (≤ 10.9 °C) or via further strong formation of TI (≤ -0.63 °C). Also the weaker TI associates with the high CO level via lower TT (≤ 7.0 °C).

Overall, the responsible climate attributes for the CO level are light wind speed and temperature inversion formation. This is a reasonable finding, also seen from previous research in the study area [15]. As this study is the first attempt for applying the data mining technique, decision trees for knowledge discovery on the climate and air pollution, it is important to note that the exact threshold values and findings require further investigation, carried out by experts in this field.

V. Conclusions

The use of SSA as the noise reduction method for the data mining application, a decision tree classifier, successfully improves the classification accuracy, compared with the original time series. The improvements are more effective, when the data set (containing all four seasons) is divided into seasons, as the summer data set classification accuracy improved up to 16.7%, compared with the original time series, after removing 6.42% of the signal. However, the autumn data did not show any improvement, which may suggest that other attributes can describe the CO level better than the currently used climate attributes. The advantage of using SSA is to provide several additive components that can be added to or removed from the main structures, allowing exploration of the nature of the noisy time series data set.

Observing how the classification accuracy changes provides information on which components are essential to generate the decision tree or which help identifying the insignificant signatures in the noisy time series (potential noise). Generating the decision trees using climate attributes to predict the CO levels from different seasons provides knowledge of the responsible climate attributes or the pathway for the CO levels. In particular, the decision tree provides threshold values of each climate variable that are responsible for the change of CO levels. Detailed examination of the decision trees suggests that the most important climate condition is wind speed less than equal to 2.3 to 2.5 m/s, which associates with high CO levels. The second most important climate attribute is any level of temperature inversion formation. Note that the exact threshold value for each climate attribute requires further investigation from experts in the field, although results are used as indexes for future climate and air pollution study. In order to increase the sensitivity in generating the decision tree, the fuzzy decision tree technique may help reduce misclassification of the different CO levels (H, M, and L). However, the introduced noise reduction method via SSA is an encouraging data pre-processing method for any data mining techniques. The data mining approach in this study can be adapted and used as a knowledge discovery tool for various environment researches in future.

Acknowledgment

K. Fukuda thanks P. Pearson for giving me useful comments as well as helping with programming for the data preparation and editorial work. Thanks to Dr. B. Martin, and Chloe and Jack Duckling for their inspiration and encouragement towards this paper. Thanks to T. Aberkane (ECan) for providing climate and air pollution data sets. Thanks to the department of Mathematical and Statistics at the University of Canterbury for supporting the travel grant.

References

[1] J.Q. Koening, *Health effects of ambient air pollution, how safe is the air we breathe?*, Boston: Kluwer Academic, 2000.

[2] M. Kossmann and A.P. Sturman. "The surface wind field during winter smog nights in Christchurch and coastal Canterbury, New Zealand," *Int. J. Climatol.*, vol, 24, no. 1, pp. 93-108, Jun. 2004.

[3] A. Scott, and M. Gunatilake, "2002 Christchurch inventory of emissions to air (R04/03)," Environment Canterbury, Christchurch, 2004.

[4] K. Fukuda and T. Takaoka, "Analysis of air pollution (PM_{10}) and respiratory morbidity rate using *K*-Maximum Sub-array (2-D) algorithm," *ACM SAC'07 in CAHC*, Seoul, Korea, March 2007, in press.

[5] K. Fukuda, and I.L. Hudson, "Global and local climatic factors on sulfur dioxide levels: comparison of residential and industrial sites," in *Proc. of 20th IWSM*, Sydney, Australia, pp.187-194, Jul. 2005.

[6] K. Fukuda and P.A. Pearson, "Data mining and image segmentation approaches for classifying defoliation in aerial forest imagery," *3rd Biennial meeting of the IEMSs*, Burlington, VT, Jul. 2006.

[7] J.M. Spate et al., "Data mining as a tool for environmental scientists," *3rd Biennial meeting of the IEMSs*, Burlington, VT, Jul. 2006.

[8] M. Aldrin, and I.H. Haff, "Generalised additive modelling of air pollution, traffic volume and meteorology," *Atmos. Environ.*, vol 39, pp. 2145-2155, 2005.

[9] S-T. Li, and L-Y. Shue, "Data mining to aid policy making in air pollution management," *Expert. Syst. Appl.*, vol. 27, pp. 331-340, 2004.

[10] T. Li, Q. Li, S. Zhu, and M. Ogihara, "A survey on wavelet applications in data mining," *ACM SIGKDD Exploration*, vol. 4, no. 2, pp. 49-68, 2002.

[11] I. H. Witten and E. Frank, *Data Mining; Practical Machine Learning Tools and Techniques with Java Implementations*, 2d ed., San Francisco: Morgan Kaufmann, 2005.

[12] Golyandina, N., Nekrutkin, V., and Zhigljavsky, A., *Analysis of Time Series Structure: SSA and Related Techniques*, Boca Raton: Chapman & Hall/CRC, 2001.

[13] R. Kumaresan and D.W. Tufts, "Data-adaptive principal component signal processing," in *IEEE Proc. Conf. on Decision and Control*, Albuqueque, NM, pp. 949-954, 1980.

[14] J.M. Colebrook, "Continuous plankton records: zooplankton and environments, North-East Atlantic and North Sea," *Oceanol. Acta*, vol. 1, pp. 9-23, 1978.

[15] K. Fukuda, and I.L. Hudson, "Investigations of short-term (hourly) weather influences on CO, NO, NO_2, PM_{10} and SO_2 levels in Christchurch, New Zealand," in *Proc. of Intl. Conf. on Research Highlights and Vanguard Technology on Environmental Engineering in Agricultural Systems*, Ishikawa, Japan, pp. 45-52, Sep. 2005.

[16] K. Fukuda, and P.A. Pearson, "Investigation of Singular Spectrum Analysis and Machine Learning for Road Sign Location," In *Extended abstracts, 7th Intl. Assoc. for Pattern Recognition workshop on DAS 2006*, Nelson, NZ, pp. 29-32, Feb. 2006.

[17] J.R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo: Morgan Kaufmann, 1993.