On Trend Association Analysis of Time Series of Atmospheric Pollutants and Meteorological Variables in Mexico City Metropolitan Area

Victor Almanza and Ildar Batyrshin

Mexican Institute of Petroleum
Eje Central Lazaro Cardenas 152, Mexico City, Mexico
{vhalmanz,batyr}@imp.mx

Abstract. The paper studies trend associations between atmospheric pollutants and meteorological variables time series of Mexico City Metropolitan Area (MCMA) by applying the Moving Approximation Transform (MAP). This recently introduced technique measures and visualizes associations of the dynamics between different time series in the form of an association network. The paper studies associations between 5 atmospheric pollutants (SO_2, O_3, NO_2, NOx and $PM_{2.5}$) and 7 meteorological variables (mean wind velocity, minimum, average and maximum values of both temperature and relative humidity) measured daily during one year in three meteorological stations located in different zones of MCMA. These associations were studied for 4 seasons characterized by different meteorological conditions. For considered stations atmospheric pollutants and meteorological variables for different seasons positive and negative associations have been found and explained.

Keywords: Time series data mining, trend associations, MAP transform, atmospheric pollutants.

1 Introduction

Air pollution in urban areas is an issue of mayor concern due to its undesirable effects such as the public health impact, environmental damage and climate changes among others. For instance, there are epidemiological studies based on time series analysis to infer respect to the contribution of particle matter less than 10 and 2.5 micrometers to respiratory diseases in hospital admissions [1-3]. Methods of Soft Computing and Nonlinear Dynamical Systems have also been applied. For example, in [4] both fuzzy and neural networks (NN) were applied to develop a system which forecast concentrations of ozone daily maximums employing registers of four monitoring sites in Seoul, where patterns in the time series were fundamental for preparing the datasets for this system. In a similar way, Kukkonen et.al, [5] and Chaloulakou et.al [6] found that NN yielded better estimates for predicting pollutants concentrations. Also, Liu [7] developed a computational model in order to infer information the dynamics of chemical reactions that produce air pollution, and Cheng [8] improved the pollutant standard index by means of applying an entropy function. Nevertheless, Dillner [9]

J.-F. Martínez-Trinidad et al. (Eds.): MCPR 2011, LNCS 6718, pp. 95–102, 2011.

and Hyvönen, et. al [10] deepens in the understanding of the pollution phenomena since they focus on the source of the particles and aerosol formation respectively. They used cluster analysis of aerosol time series.

Since Mexico City is considered a megacity, two main pollutants are of particular concern in the MCMA, ozone and particulate matter (PM), because of frequent exeedance to air quality standards several days a year. Long-range transport could influence air quality, and hence effects could be felt in regions far from their sources [11]. This is why it is important to infer about the physical and chemical behavior of air pollution.

The objective of the paper is to apply novel time series data mining techniques based on local trend associations [12] to pollutant time series in order to obtain information about possible associations between meteorological conditions and air pollution for different year seasons in three meteorological stations located in MCMA. Meteorological conditions near these stations differ one from another depending on wind strength and wind direction. The information obtained with this method, can confirm expected relations between meteorological parameters and air pollutants as well as new insights about such relations in different seasons in MCMA.

The paper has the following structure. Section 2 describes the data used in analysis. The Section 3 gives short description of the method of local trend associations of time series. Section 4 presents obtained results and their discussion. Section 5 contains conclusions.

2 Data Used for Analysis

Air pollution and meteorological time series of three stations of the Atmospheric Automatic Monitoring Network (RAMA by its Spanish initials) were considered for the analysis (Fig.1). These are Pedregal (PED), Tlalnepantla (TLA) and Merced (MER) stations. They were chosen because they represent different pollution zones in the MCMA. MER is located in the center of the Mexico City near heavily, paved and curbed surface streets with light-duty vehicles and modern heavy-duty diesel buses. PED is in a suburban neighborhood near clean, paved residential roads, lightly traveled and presents no nearby industries. TLA is both an industrial and residential area with nearby electronics manufacturing, corn milling, and metal fabricating facilities [13].

The pollutants considered were ozone (O3), sulfur dioxide (SO2), nitrogen dioxide (NO2), nitrogen oxides (NOx), and particulate matter less than two micrometers (PM2.5); and the meteorological variables were wind direction (WD), wind velocity (WV) temperature (T) and relative humidity (RH). The raw time series consisted of hourly data for year 2004, but in this analysis only the daily maximum were considered since for pollutants the maximum concentration achieved in the day is more representative than the average since the latter could diminish the level of exposure of the population. In the case of meteorological variables, minimum, average and maximum values time series were constructed. Missing data were handled by simple interpolation.

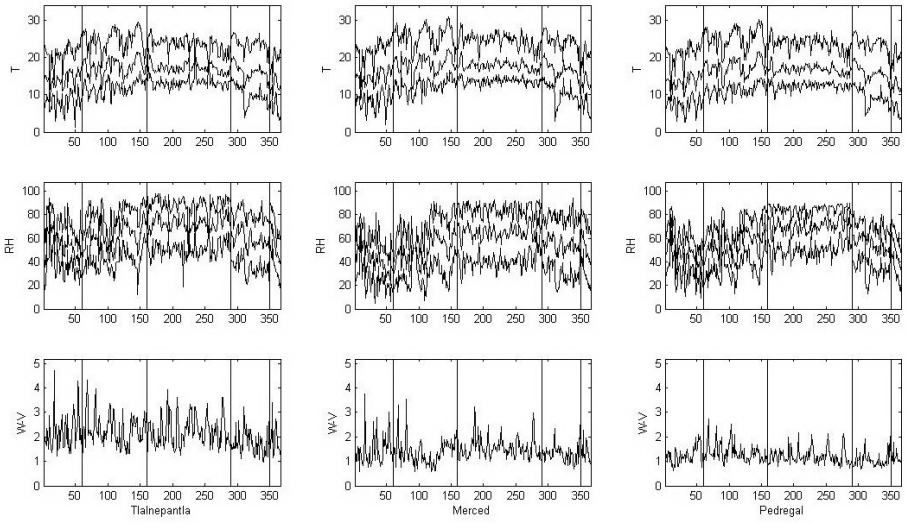


Fig. 1. Meteorological time series of three stations separated on 4 seasons (vertical lines denote borders between seasons)

3 Basic Notions of MAP Transform and Trend Associations

A time series (y,t) is a sequence $\{(y_i,t_i)\}$, $i \in I = (1,\ldots, n)$, such that $t_i < t_{i+1}$ for all $i = 1,.., n\text{-}1$, where y_i and t_i are real numbers called time series values and time points, correspondingly,. A time series (y,t) will be denoted also as y. A window W_i of a length $k > 1$ is a sequence of indexes $W_i = (i, i+1,\ldots, i+k\text{-}1)$, $i \in \{1,\ldots, n\text{-}k+1\}$. The sequence $y_{W_i} = (y_i, y_{i+1},\ldots, y_{i+k-1})$ of the corresponding values of time series y is called a partial time series induced by window W_i. A sequence $J = (W_1, W_2,\ldots, W_{n-k+1})$ of all windows of size k, $(1 < k \le n)$, is called a moving (or sliding) window. Such mowing window is used, for example, in statistics in moving average procedure for smoothing time series when the value in the middle of the window replaced by the mean of values from this window.

Suppose J is a moving window of size k and $y_{W_i} = (y_i, y_{i+1},\ldots, y_{i+k-1})$, $i \in (1,2,\ldots, n\text{-}k+1)$, are corresponding partial time series in time points $(t_i, t_{i+1},\ldots, t_{i+k-1})$. A linear function $f_i = a_i t + b_i$ with parameters $\{a_i, b_i\}$ minimizing the criterion

$$Q(f_i, y_{W_i}) = \sum_{j=i}^{i+k-1} \left(f_i(t_j) - y_j\right)^2 = \sum_{j=i}^{i+k-1} \left(a_i t_j + b_i - y_j\right)^2 . \quad (1)$$

is called a moving (least squares) approximation of y_{Wi}. The solution of (1) is well known and optimal values of parameters a_i, b_i can be calculated as follows:

$$a_i = \sum_{j=i}^{i+k-1} \left(t_j - \overline{t_i}\right)\left(y_j - \overline{y_i}\right) \Big/ \sum_{j=i}^{i+k-1} \left(t_j - \overline{t_i}\right)^2, \quad b_i = \overline{y_i} - a_i \overline{t_i} . \quad (2)$$

Where $\overline{t_i}$ and $\overline{y_i}$ are mean values of t and y.

Definition 1. A transformation $MAP_k(y,t)= a,$ where $a = (a_1, \ldots, a_{n-k+1})$ is a sequence of slope values obtained as a result of moving approximations of time series (y,t) in moving window of size k is called a moving approximation (MAP) transform of time series y. The slope values $a_1, \ldots, a_{n-k+1}$ are called local trends.

Elements a_i, $(i = 1,\ldots, n-k+1)$ from $MAP_k(y,t)$ will be denoted as $MAP_{ki}(y,t)$.

In many applications time points $t_1, \ldots, t_n$ are increasing with a constant step h such that $t_{i+1} - t_i = h$ for all $i = 1,\ldots, n-1$. In such cases in MAP transform the set of time points $t= (t_1, \ldots, t_n)$ can be replaced by the set of indexes $I = (1,\ldots, n)$ as follows: $MAP_k (y,t) = (1/h)MAP_k (y,I)$ and the formula (2) for local trends can be simplified as follows [12].

As a measure of similarity between time series one can use measures of similarity between their MAP transforms. Some of these measures satisfy very nice properties of invariance to linear transformations of time and time series values.

Definition 2. Suppose $y = (y_1,\ldots,y_n)$, $x = (x_1,\ldots, x_n)$ are two time series and $MAP_k(y) = (a_{y1}, \ldots, a_{ym})$, $MAP_k (x)= (a_{x1}, \ldots, a_{xm})$, $(k\in\{2,\ldots,n-1\}, m= n- k+1)$, are their MAP transforms. The following function is called a measure of local trend associations:

$$coss_k(y,x) = \sum_{i=1}^{m}\left(a_{y_i} \cdot a_{xi}\right) \Big/ \sqrt{\sum_{i=1}^{m} a_{y_i}^2 \cdot \sum_{j=1}^{m} a_{x_j}^2}$$

Suppose p,q,r,s, $(p,r\neq 0)$ are real values and (y,t) is a time series. Denote $py+q = (py_1+q, \ldots, py_n+q)$ and $rt+s = (rt_1+s, \ldots, rt_n+s)$. A transformation $L(y,t) = (py+q, rt+s)$ is called a linear transformation of time series (y,t).

Theorem. Suppose L_1 and L_2 are two linear transformations of time series (y,t) and (x,t) given by the sets of parameters (p_1,q_1,r_1,s_1) and (p_2,q_2,r_2,s_2), respectively, where $p_1, p_2, r_1, r_2 \neq 0$, then

$$coss_k(L_1(y,t), L_2(x,t)) =$$
$$sign(p_1)\cdot sign(r_1)\cdot sign(p_2)\cdot sign(r_2)\cdot coss_k((y,t), (x,t)).$$

From this Theorem it follows a very nice invariance property of local trend association measure under various types of normalization of time series.

Analysis of associations between time series is based on the analysis of associations between them for different window size. The sequence of association values $AV(y,x)= (coss_2(y,x),\ldots, coss_n(y,x))$ for all sizes of window is called an association function [12]. A specific measure of association between time series is defined by the subset of window sizes $J \subset \{2,\ldots,n\}$ as a maximum or average of all associations $coss_k(y,x)$, $k\in J$. Examples of application of association measure to the classification of time series are considered in [12].

4 Results and Discussion

Trend association analysis was applied to pollutant and meteorological time series of three monitoring stations in MCMA, by means of the evaluation of local and global trends associations, in order to infer about possible dynamic relationships. Based on these associations the corresponding association networks were constructed. In these

networks only association values in the interval [0.5, 0.9] are discussed for each monitoring station and seasonal period.

In TLA for association values greater or equal to 0.8 most of the associations were for meteorological variables in all seasonal periods, mainly between the Relative Humidity and Temperature. For pollutants, in fall a positive association was the most relevant for the class {NOx, PM2.5}; and in winter only the class {NO2, NOx} with positive association exist. Since NOx is defined as the sum of NO plus NO2, it is expected that these chemical species correlate, but the reason to include it is because in the rest of the monitoring stations this class is present only in PED for the summer period. For association values greater or equal to 0.6 there is a cluster of positive association in the class {NO2, NOx, PM2.5} for the summer period, a positive association in the class {O3, PM2.5} for summer and winter, and an inverse association for the class {WV, O3} is present in fall and in winter. Example of association network is presented in Fig. 2, where solid lines denote positive associations and dashed lines denote negative associations. This latter class is also present in the same seasonal periods in MER station. So, it is possible that in TLA and MER, turbulent mixing promoted pollutant dispersion and transport which can reduce the concentration values of these species, since the maximum values in both stations were lower than in PED station. Moreover the complex wind patterns in MCMA [13], can influence the local wind patterns in PED possibly by slope flow. Moreover, in winter season the presence of cold fronts increases wind velocities.

The class {O3, PM2.5} is also relevant in MER and PED but only in winter season. This result suggests that the photochemical activity induces aerosol formation. However, as stated previously the higher concentration in PED is possibly related to stagnation conditions. The cluster {NO2, NOx, PM2.5} is also present in PED station for winter season. This can suggest that road traffic is an important source for the aerosol particles in these regions [14].

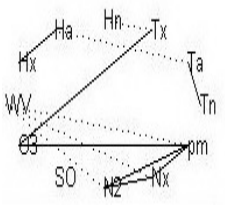
Association values greater or equal to 0.6 in TLA show a cluster for the class {O3, NO2, NOx} in spring season and a cluster for {Tmax, NOx, PM2.5} for fall season. These clusters are not present neither in PED nor in MER. Both clusters suggest higher traffic activity for TLA in this season. Other clusters of interest are the class {PM2.5, SO2, RHmax} and the class {PM2.5, SO2, O3}, which are relevant only for TLA in winter season. These clusters imply that in TLA, the sulfur content present in the atmospheric particles can be higher than in MER and PED. Composition analysis for particles in a study conducted in the MCMA showed that sulphate content was higher in TLA [15] because of large emissions of SO2 proper of the industrial zone where TLA is located in. So, the MAP transform seems to be capturing important dynamics in the time series.

In this association level the class {RHmax, PM2.5} is only present for MER in fall season. It is an inverse association that suggests the possibility that an increase in precipitation causes a decrease in particle concentration through scavenging [16]. Besides, the class of the inverse association {WV-PM2.5} is important only in TLA and MER in winter season, suggesting that transport is an important source for decreasing the concentration of particles in the north part of MCMA.

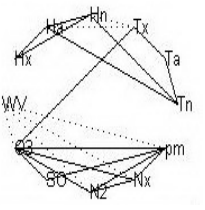
Finally for the association value greater or equal to 0.5 more classes are obtained, but the most relevant ones are the class discussed above for TLA {PM2.5, SO2, O3}

which now is present in MER for winter season and in PED for winter season too. The other class is the inverse association {RH, PM2.5} in MER for fall season.

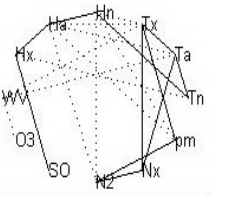
It is worth to mention that in this stage the analysis can support other studies regarding the source contribution to the evolution of atmospheric pollutants. Moreover it is possible to apply this approach to information of aerosols measurements, sulfate ions concentration and precipitation among others in order to complement the present study.



a) Trend associations for Tlalnepantla in Summer



b) Trend associations for Pedregal in Winter



c) Trend associations for Merced in Fall

Fig. 2. Examples of association networks found for different seasons and stations

5 Conclusions

In this work new methods of data mining were applied to analyze relationships between dynamics of pollutants and meteorological time series of three representative monitoring stations in the MCMA for year 2004. This approach calculates associations between elements of the systems under consideration as associations between the respective time series. The association networks obtained by Moving Approximation Transform were discussed for association levels spanning from 0.5 to 0.9. At high levels the positive association class {NOx, PM2.5} was the most relevant for TLA station suggesting a high contribution from mobile sources. At moderate levels, the positive association class {O3, PM2.5}, was present in summer in winter in all the stations but not simultaneously. Besides the cluster {PM2.5, SO2, O3} gives insight about the photochemical activity in the north part of MCMA. Moreover, the method captures the inverse association {WV, O3} which suggest removal of particles by turbulent mixing, especially in winter seasons and in less degree in summer season only in TLA and MER. In wet season the inverse association {RHmax, PM2.5} suggest scavenging of particles in TLA and MER only. So, PED seems to be influenced by the local winds to promote stagnation conditions, which can be a reason for high concentration values in this year.

However it would be useful to include Volatile Organic Compounds, precipitation, and solar radiation time series in order to find new associations, and at another stage to consider time series of respiratory illness in order to find possible associations by

zone and by pollutant, which could be an aid in applying statistical time series analysis such as GAM or ARMA. Finally it is possible that MAP approach also could serve for finding patterns for fuzzy rules construction.

Acknowledgements

Victor Almanza thanks Dr. Gustavo Sosa for providing useful comments and suggestions for this work.

References

1. Roberts, S.: Biologically Plausible Particulate Air Pollution Mortality Concentration-Response Functions. Environ. Health Perspect. 112(3), 309–313 (2004)
2. Samoli, E., Analitis, A., Touloumi, G., Schwartz, J., Anderson, H.R., Sunyer, J., et al.: Estimating the Exposure–Response Relationships between Particulate Matter and Mortality within the APHEA Multicity Project. Environ. Health Perspect. 113, 88–95 (2005)
3. Galán, I., Tobías, A., Banegas, J.R., Aránguez, E.: Short-Term Effects of Air Pollution on daily Asthma Emergency Room Admissions. Eur. Respir. J. 22, 802–808 (2003)
4. Heo, J.K., Kim, D.S.: A New Method of Ozone Forecasting using Fuzzy Expert and Neural Network Systems. Sci. Total Environ. 325, 221–237 (2004)
5. Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G.: Extensive Evaluation of Neural Network Models for the Prediction of NO2 and PM10 Concentrations, compared with a Deterministic Modelling System and Measurements in Central Helsinki. Atm. Env. 37, 4539–4550 (2003)
6. Chaloulakou, A., Saisana, M., Spyrellis, N.: Comparative Assessment of Neural Networks and Regression Models for Forecasting Summertime Ozone in Athens. Sci.Tot. Environ. 313, 1–13 (2003)
7. Liu, Z., Lai, Y.C., Lopez, J.M.: Noise-induced Enhancement of Chemical Reactions in Nonlinear Flows. Chaos. 12(2), 417–425 (2002)
8. Cheng, W., Kuo, Y., Lin, P., Chang, K., Chen, Y., Lin, T., Huang, R.: Revised Air Quality Index Derived from an Entropy Function. Atmos. Environ. 38, 383–391 (2004)
9. Dillner, A.M.: A Quantitative Method for Clustering Size Distributions of Elements. Atm. Env. 39, 1525–1537 (2005)
10. Hyvönen, S., Junninen, H., Laakso, L., Dal Maso, M., Grönholm, T., Bonn, B., Keronen, P., Aalto, P., Hiltunen, V., Pohja, T., Launiainen, S., Hari, P., Mannila, H., Kulmala, M.: A Look at Aerosol Formation Using Data Mining Techniques. Atmos Chem. Phys. Discuss. 5, 7577–7611 (2005)
11. Molina, M., Molina, L.: Megacities and Atmospheric Pollution. J. Air Waste Manage. Assoc. 54, 644–680 (2004)
12. Batyrshin, I., Herrera-Avelar, R., Sheremetov, L., Panova, A.: Association Networks in Time Series Data Mining. In: NAFIPS 2005 Soft Computing for Real World Applications, Ann Arbor, Michigan, USA, pp. 754–759 (2005)

13. Edgerton, S.A., Bian, X., Doran, J.C., Fast, J.D., Hubbe, J.M., Malone, E.L., Shaw, W.J., Whiteman, C.D., Zhong, S., Arriaga, J.L., Ortiz, E., Ruiz, M., Sosa, G., Vega, E., Limon, T., Guzman, F., Archuleta, J., Bossert, J.E., Elliot, S.M., Lee, J.T., McNair, L.A., Chow, J.C., Watson, J.G., Coulter, R.L., Doskey, V.: Particulate Air Pollution in Mexico City: A Collaborative Research Project. J. Air Waste M. A. 49(10), 1221–1229 (1999)
14. Harrison, R., Deacon, A., Jones, M., Appleby, R.: Sources and Processes Affecting Concentrations of PM10 and PM2.5 Particulate Matter in Birmingham (U.K). Atm. Env. 31(24), 4103–4117 (1997)
15. Chow, J.C., Watson, J.G., Edgerton, S.A., Vega, E.: Chemical Composition of PM2.5 and PM10 in Mexico City During Winter 1997. Sci. Tot. Environ. 287, 177–201 (2002)
16. Tai, A., Mickley, L., Jacob, D.: Correlations Between Fine Particulate Matter (PM2.5) and Meteorological Variables in the United States: Implications for the sensitivity of PM2.5 to Climate Change. Atm. Env. 44, 3976–3984 (2010)