MODEL-BASED ESTIMATION AND PREDICTION OF ATMOSPHERIC POLLUTANTS IN THE ENVIRONMENT

UROOSA TAGAR & RANJAN VEPA
School of Engineering and Material Science, Queen Mary University of London, UK

ABSTRACT

Air pollution in urban areas is mainly due to the intense use of motorized transport for travelling. To monitor the change in the concentration of pollutants in the atmosphere on a real time basis, we may use any control-theory engendered filter such as a Kalman filter, an extended Kalman filter, an unscented Kalman filter or a particle filter, which are statistically based methods that rely on some form of dynamic model of the changes in the concentration, and are therefore suitable for monitoring concentrations of carbon monoxide (CO), carbon dioxide (CO_2), nitric oxide (NO), unburnt hydrocarbons, nitrogen dioxide (NO_2) and other pollutants from cars and jet engines. In this study air quality monitoring data for the years 2013 to 2017 (5 years) was obtained from the website of Department for Environment, Food and Rural Affairs (DEFRA), UK. A formal statistical model that can be used to predict the dynamics of pollutants in the atmosphere or any other similar pollutant is not available and it is really necessary that such a model is developed based on the observed data. A simple second order model of this data was established using MATLAB software. After that the concentrations of NO, NO_2, ozone and particulate matter (PM_{10}) in the atmosphere at a particular location within London city, were estimated using certain techniques namely, Least-squares estimation technique, Theil–Sen regression technique and an auto-regressive predictive filter technique also known as the linear predictive coding (LPC) estimator. Based on the second order model, a Kalman filter technique was implemented to predict the concentration of the pollutants. Results obtained from these techniques were compared with each other and generally indicate a steady decline of the pollutant concentrations while the Kalman filter and LPC estimator predict a periodic reduction in the concentration of the same pollutants over the 5 year period.

Keywords: air pollution, concentration estimation, prediction, nitric oxide, nitrogen dioxide, particulate matter, ozone.

1 INTRODUCTION

Various species of pollutants and toxic chemicals are characterised by several emission pathways into the atmosphere. They could be emitted by point sources in relation to the size of large cities due to relatively rare events in time such as, accidental release at nuclear power plants or volcanic eruptions or from regionally distributed sources such as emission of photochemical smog generators on roads and highways as well as forest fires. Moreover, time interval of such releases could quite varied. These air pollutants generally are known to be carried hundreds or thousands of kilometres from wherever they are released to distant points on the surface of the globe depending on the nature and composition of pollutants. Consequently they could affect the health of several groups of humans (see for example, Murray and Lipfert [1]) and result in a long-term damage to both the population and the environment. Such incidents can also have a huge economic impact due to a variety of reasons. For example the volcanic eruption of Eyjafjallajökull in Iceland over a period of six days in April, 2010 caused enormous disruption to air travel to and from London Heathrow, following the cancellation of several hundreds of flights because of the damage volcanic ash that was ejected to the atmosphere could cause to aircraft. Consequently the high episodic concentrations of air pollutants (e.g., nitrogen oxides, particulates, and carbon monoxide) in major population centres across the world, is a matter of deep concern due to their harmful effects on human health and the environment.

WIT Transactions on Ecology and the Environment, Vol 252, © 2021 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR210021

Statistical analysis (see for example Wilcox [2]) of air pollution data has been extensively advanced in the recent decade (see for, e.g., Gyarmati-Szabo et al. [3] and references therein). For short term predictions of pollution concentration, statistical methods are traditionally more suitable than physically based deterministic models [4]. The long-term goals of the research in this area, were succinctly stated by Gyarmati-Szabo et al. [3], as:

i. prediction of critical levels of pollutants to give out health warnings to local populations;
ii. identifying and predicting changing trends in high concentration levels;
iii. assessing temporal changes (secular or periodic) in air pollution levels due to the impact of human activities on the environment, either directly due to changing emission patterns or indirectly due to climate change.

In a recent study, Gyarmati-Szabo et al. [3], a novel approach to extreme value modelling is proposed and applied to the problem predicting pollutant concentrations in the atmosphere over the city of Leeds (West Yorkshire, UK), on both long (e.g., daily or yearly) and short (e.g., quarter of an hour) time scales. Although estimation methods based on linear Kalman filtering were introduced several decades ago (see for example Desalu et al. [5]), recent applications (Samaranayake et al. [6]) have focussed on nonlinear Kalman filtering techniques such as extended Kalman filtering and large scale ensemble Kalman filtering based on the use of a variety of nonlinear models. Advanced regression based methods have also been pursued as evidenced by the work of Ortiz and Friedrich [7]. A detailed review of the theory and application of those forecasting models have been presented by Bai et al. [8]. A host of different modelling approaches have also been developed, ranging from single species continuous time models characterised by ordinary differentials for the evolution of the state (see for example Liu et al. [9]) to dispersive and distributed models characterised by partial differential equations for the evolution and spread of a species in time and space, as in Leelőssy et al. [10].

In this paper, an approach similar to that proposed by Gyarmati-Szabo et al. [3] is adopted, to study pollutant concentrations in the atmosphere over the city of London at one location (London, Bloomsbury), on long time scales such as yearly and 5-yearly periods, based of available hourly measurements of pollutant concentrations from 2013 to 2017. The concentrations of NO, Nitrogen dioxide (NO_2), Ozone and particulate matter (PM_{10}) in the atmosphere at a particular location within London city, were estimated using different techniques namely, Least-squares estimation technique, Theil–Sen regression technique and an auto-regressive predictive filter known as the linear predictive coding (LPC) estimator technique. Based on the second order model, a Kalman filter was implemented to predict the concentration of the pollutants. Results obtained from these filters were compared with results of least-squares estimation method and Theil–Sen regression are similar and generally indicate a steady decline of the pollutant concentrations while the Kalman filter and LPC estimator predict a periodic reduction in the concentration of the same pollutants over the 5 year period. It is expected that a physical model would be developed and used for prediction over a longer time scale and for estimation and classification of the pollution components in the atmosphere.

2 METHODOLOGY

Air quality measurements give only quantitative information about ambient concentrations without identifying causes of air quality problems. There are over 1,500 sites across the UK that monitor air quality. They are organised into networks that gather a particular kind of

information, using a particular method. There are two major types – automatic and non-automatic networks that gather air pollution information.

For monitoring and reporting air pollution the UK has been divided into two zones: non-agglomeration zones and agglomeration zones. There are fifteen non-agglomeration zones which match:

i. The boundaries of England's Government Offices for the Regions; and
ii. The boundaries agreed by the Scottish Government, Welsh Government and Department of the Environment in Northern Ireland.

Based on the above zones, there are 16 regions and 16 urban areas which are used for reporting latest real-time monitoring data of air pollutants. These are a subset of the UK agglomeration and non-agglomeration zones.

This section provides details of the study area including geometric characteristics and air quality measurements data of a five year time frame for above mentioned pollutants.

2.1 Case study

This case study concerns the estimation and prediction of the concentrations of nitric oxide (NO), nitrogen dioxide (NO_2), ozone and particulate matter, (PM_{10}) at a particular site in city of London.

2.2 Air quality monitoring site

The study area is Bloomsbury in the city of London. Bloomsbury has no official boundaries, but can be roughly defined as the square of territory bounded by Tottenham Court Road to the west, Euston Road to the north, Gray's Inn Road to the east, and either High Holborn or the thoroughfare formed by New Oxford Street, Bloomsbury Way and Theobalds Road to the south. Bloomsbury merges gradually with Holborn in the south, with St Pancras and King's Cross in the north-east and with Clerkenwell in the south-east.

2.3 Data collection and analysis

In this research study, the air quality measurement data was collected from the website of UK Air Information Resource, Department for Environment, Food and Rural Affairs (DEFRA) UK with their permission. The data that has been used is the hourly concentration of NO nitrogen dioxide (NO_2), ozone and particulate matter (PM_{10}) in a central corridor of London, called Bloomsbury, from 2013 to 2017 years.

2.4 Concentration of the pollutants

Concentration data of the pollutants from year 2013 to year 2017 is collected from the UK Air Information Resource [11], Department for Environment, Food and Rural Affairs (DEFRA).

A preliminary inspection of the data generally not only indicated several season variations on a yearly basis but also variations over 5 year period. These variations dictated the choice of the time scales and the methods of analysis that could be used. The data analysis methods used in this study are the recursive least squares method, the Theil–Sen linear regression estimator, the linear predictive coding estimator and the Kalman filter.

2.5 Data analysis techniques

2.5.1 Least squares method

The least squares method (Simon [12]) is a form of mathematical regression analysis that finds the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points. Each point of data is representative of the relationship between a known independent variable and an unknown dependent variable. In our work the variables used for least square estimation are pollutant concentration measured in units ($\mu g/m^3$) and time in hours but represented as a whole year.

2.5.2 Theil–Sen estimator technique

Theil–Sen estimator is a method for robustly fitting a line to sample points in the plane (simple linear regression) by choosing the median of the slopes of all lines through pairs of points. As defined by Theil [13], Theil–Sen estimator of a set of two-dimensional points (x_i, y_i) is the median (m) of the slopes $\left(y_j - y_i\right)\Big/\left(x_j - x_i\right)$ determined by all pairs of sample points. Sen [14] extended this definition to handle the case in which two data points have the same x coordinate. In Sen's definition, one takes the median of the slopes defined only from pairs of points having distinct x coordinates. The Theil–Sen estimator has several important features such as unbiasedness (Wilcox [15], Wang and Yu [16]) and most suitable for prediction and forecasting applications.

2.5.3 Linear predictive coding estimator technique

Linear predictive coding (O'Shaugnessy [19]) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters.

In linear predictive coding, the signal of interest is modelled as,

$$y_{n+1} = \sum_{i=1}^{n} a_i y_i + e_i \,, \tag{1}$$

So as make the error e_i a minimum. An estimate of the prediction of the measurement y_i , $i = 1, 2...n$, is then given by,

$$\hat{y}_{n+1} = \sum_{i=1}^{n} a_i y_i \,. \tag{2}$$

2.5.4 Kalman filter estimation technique

The Kalman filter model (Brown and Hwang [20]) assumes that the state of a system, represented as a state vector, at a time t evolved from the prior state at time $t-1$ according to the below equation:

$$\mathbf{x}_{i+1} = \mathbf{A}_i \mathbf{x}_i + \mathbf{B}_i \mathbf{u}_i \,, \tag{3}$$

where $\mathbf{x}_i$ is the state vector containing the terms of interest for the system (concentration and rate of change of concentration) at time t, $\mathbf{u}_i$ is the vector containing any control or disturbance inputs, $\mathbf{A}_i$ is the state transition matrix which applies the effect of each system

state parameter at time $t-1$ on the system state at time t and $\mathbf{B}_i$ is a distribution matrix that distributes the inputs to the various individual state equations. The measurement model takes the form,

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x}_i + \mathbf{v}_i \,, \tag{4}$$

where $\mathbf{y}_i$ is the measurements vector that is linearly related to the state vector $\mathbf{x}_i$ and $\mathbf{v}_i$ is the vector containing disturbance inputs to the measurements.

The Kalman filter which is used to obtain an estimate of the state vector given by $\hat{\mathbf{x}}_i$, then takes the form,

$$\hat{\mathbf{x}}_{i+1} = \mathbf{A}_i \hat{\mathbf{x}}_i + \mathbf{B}_i \hat{\mathbf{u}}_i + \mathbf{K}_i \left(\mathbf{y}_i - \mathbf{H}_i \hat{\mathbf{x}}_i \right), \tag{5}$$

where $\mathbf{K}_i$ is the optimum Kalman gain multiplying an estimate of the measurement error.

3 RESULTS

3.1 Pollutants analysis using the least square technique and the Theil–Sen estimator technique

In the first instance, the Nitric Oxide (NO) concentration in the atmosphere was estimated using least square estimation method (LSEM).

Following the application of the LSEM, the pollutant concentrations were further analysed using Theil–Sen estimator technique. These results were also compared with LSEM results. Figs 1–5 show the comparative results of NO pollution data using LSEM and Theil–Sen estimation techniques of the years 2013 to 2017 (5 years) of London Bloomsbury.

In Fig. 1, data analysed was of the period from January 2013 to January 2014, allowing an extra month time to show the trend of pollution. The pollutant concentration was expressed in µg/m^3 unit and time in terms of year. The scale bar is kept generalized for each analysed graph keeping the same number of days for each year but the data point used are

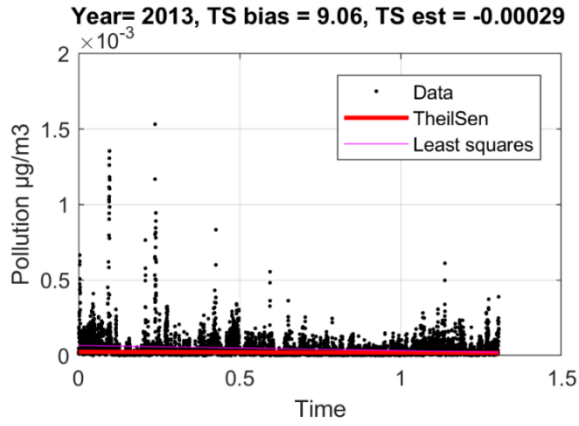

Figure 1: Comparative analysis of NO concentrations using LSEM and Theil–Sen techniques for 2013 (395 days).

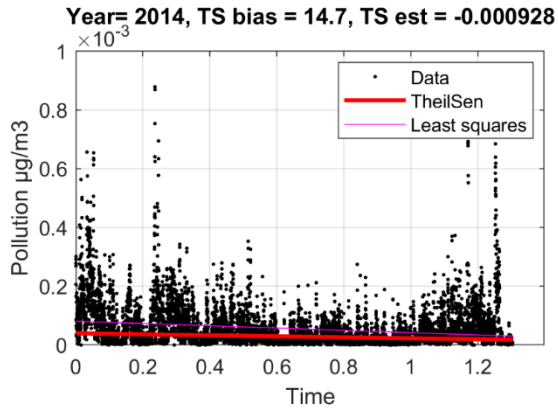

Figure 2: Comparative NO concentrations using LSEM and Theil–Sen, 2014 (395 days).

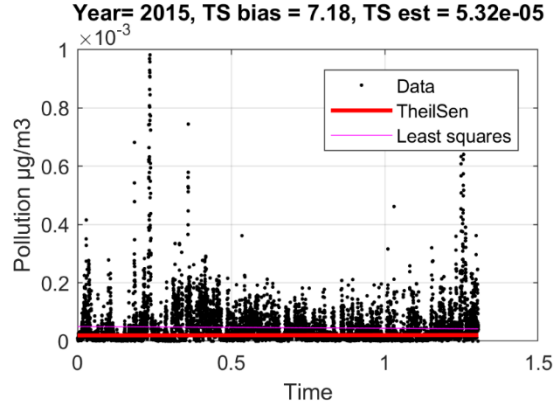

Figure 3: Comparative NO concentrations using LSEM and Theil–Sen, 2015 (395 days).

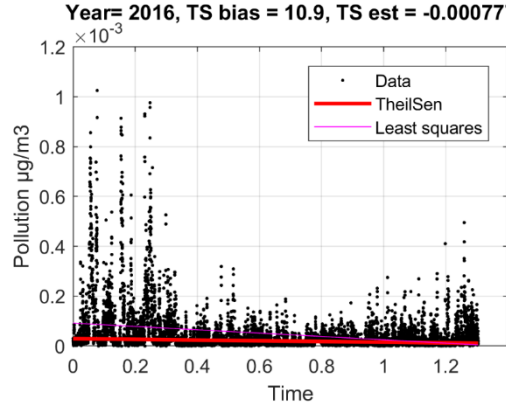

Figure 4: Comparative NO concentrations using LSEM and Theil–Sen, 2016 (395 days).

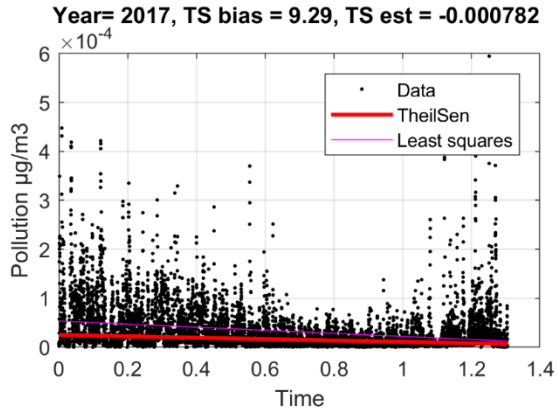

Figure 5: Comparative NO concentrations using LSEM and Theil–Sen, 2017 (395 days).

hourly collection of pollutant concentration. The graphical representation of the concentration of nitric oxide (NO) shows that there are major three peaks over a year, and the concentration was much lower during the summer period of June and July (0.6–0.8 time scale bar). The data also shows that in year 2014, NO concentration was much less than the year 2013. A major rise in the NO peaks during the winter months of January and February, was also observed. Comparing with the other years data obtained, overall concentration was much higher in 2014. The highest concentration peak is above 500 µg/m^3. The concentration in the year 2015 was found to be much more compared to previous year but the major increase in concentration was in the winter period.

The slope of the least-squares regression line is the average change in the predicted values of the response variable when the explanatory variable increases by 1 unit. Fig. 6 refers to the slope and bias of the Theil–Sen estimator for NO pollution concentration over a 5 year time frame. Bias means that the expected value of the estimator is not equal to the pollution parameter. Intuitively in a regression analysis, this would mean that the estimate of one of the parameters is too high or too low. In 2017, the highest concentration peak was less than 300 µg/m^3. Figs 7–9 refer respectively to the Theil–Sen estimators for nitrogen dioxide (NO_2) concentration, ozone concentration and particulate matter (PM_{10}) in the atmosphere. The magnitudes of the slopes and biases of the Theil–Sen estimators in each are of similar order. This means these three techniques predicted the different pollutant's behaviour in a same manner over a fixed period of time. The continuous negative slope over 5 years for NO, NO_2 and ozone shows the negative correlation of pollutant concentration over time. It means the concentration was decreased with increasing time period and Theil–Sen estimator technique correlated same as the least square technique.

The slope has been changed for PM_{10} concentration and continuous positive trend in the slope of Theil–Sen over 5 years shown in Fig. 9 proves the positive correlation between concentration and time.

3.2 Pollutants estimation using the Kalman filter

The available data from the air quality measurements was also analysed using a lightly damped second order model Kalman filter, to capture the seasonal and annual variations of

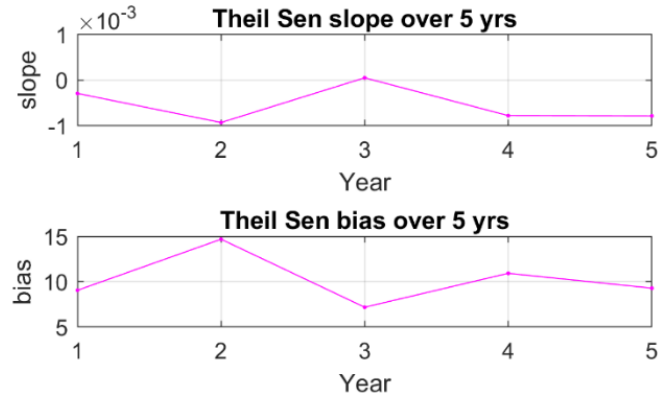

Figure 6: The Theil–Sen slope and bias over 5 years of NO concentration.

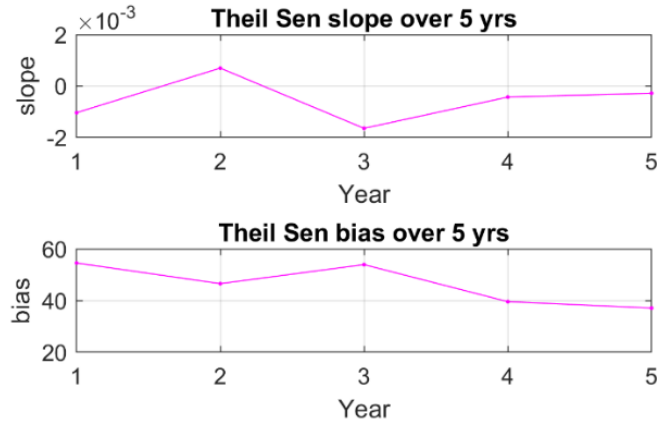

Figure 7: The Theil–Sen slope and bias over 5 years of NO_2 concentration.

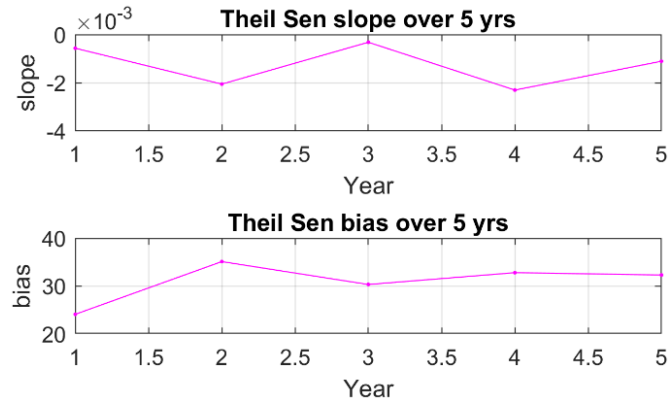

Figure 8: The Theil–Sen slope and bias over 5 years of ozone concentration.

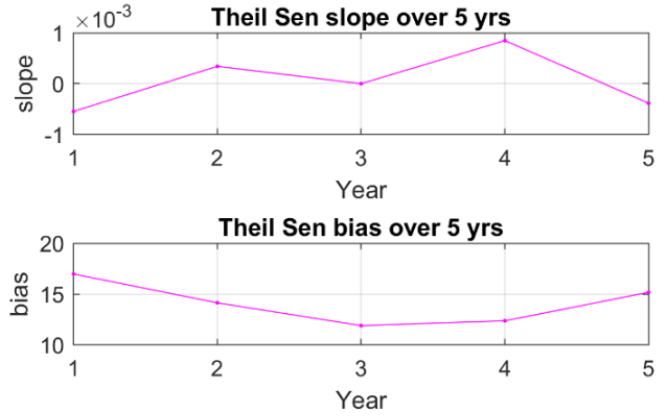

Figure 9: The Theil–Sen slope and bias over 5 years of PM_{10} concentration.

the pollutant concentration over the five year period. Typical estimates of NO concentrations are compared with the measurement and model predicted data ass shown in Fig. 10. Similar results were obtained for nitrogen dioxide (NO_2) concentration, ozone concentration and particulate matter (PM_{10}) in the atmosphere. We can see that filter estimate (Kalman filter) is predicting the pollutants in the same manner but with reduced error.

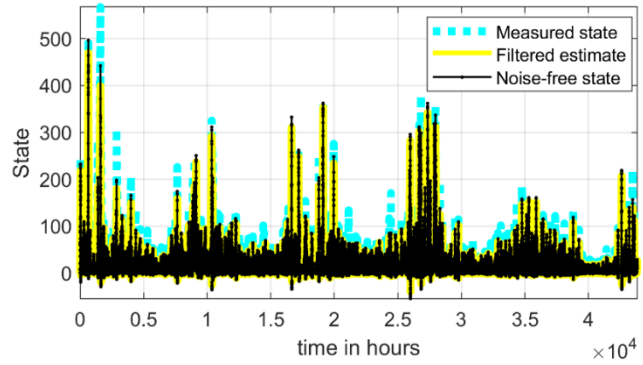

Figure 10: Comparison of measured, model predicted and Kalman filter estimated NO concentrations.

The noise free data represents the smooth data. The data points of pollutants are named as state, a general term to show al data. It can be seen that our model can easily be fit into all points and gives a smooth signal to easily predict the behaviour.

3.3 Pollutants analysis using the LPC estimator technique

The LPC estimation method was also applied to the pollutant concentrations. Typical estimates of NO concentrations are compared with the original measurement (labelled as

the "original signal" in the figure) in Fig. 11. The error in the LPC estimate is shown in Fig. 12. It can be seen that LPC estimator predicted and followed the original data points in a very same manner. It means LPC estimator can predict the pollution data points efficiently. The amplitude shows the frequency of signals produced by pollutant data points named sample number.

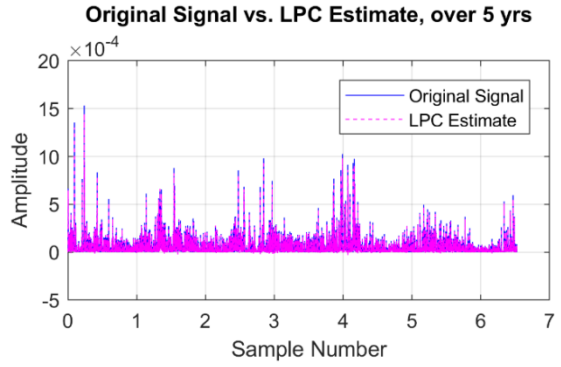

Figure 11: Comparison of the LPC estimate with the original measurement (labelled as the "original signal" in the figure).

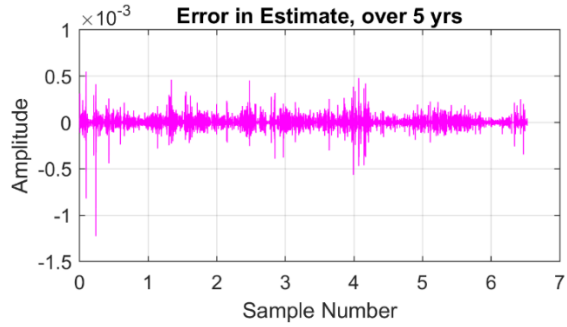

Figure 12: Error in the estimate of NO concentration over 5 years.

4 DISCUSSION AND CONCLUSIONS

In this study, application of the least square method, the Theil–Sen estimator, the Kalman filter and the LPC estimator have been successfully used to predict the above mentioned pollutant's concentrations and thereby forecasted the changes at this particular urban site called London Bloomsbury. The results generally indicate that over the 5 year time period, the overall concentration of the pollutants was predicted which was steadily declining with increase in time. However it is also possible to understand the limitations of the predictions. The linear fit of least square estimate and Theil–Sen estimate showed the same behaviour over 5 year data points. Negative slope of Theil–Sen estimate showed a declining trend of pollutants concentration over 5 year time slot. Furthermore, results of Kalman filter predicted the same behaviour as of measured state and noise free data (pollutant concentration). The frequency (amplitude) spectra of pollutant concentration by LPC

estimator follows the same signal of original data. A very low error estimate of LPC ranging between -0.5×10^{-3} to 0.5×10^{3} proves the efficiency of LPC and makes this technique suitable for predicting the behaviour of atmospheric pollutants for future. It can be concluded that these techniques can be efficiently used to predict the pollutant concentrations. Primarily, although the estimates are quite accurate, the prediction is only valid in the short term, which could be typically less than a year. It is desirable that predictions can be made in the longer term as well.

5 FUTURE WORK

It is planned to use more physically meaningful models to predict and forecast the changes in the pollution concentration over a longer prediction period. Air pollution dynamics can be modelled using mathematical representations to describe the novel relationship between emissions, meteorology, atmospheric concentrations, surface depositions and other factors [21]. A dynamic model of pollution generation, particularly gaseous pollutants may be developed based on system dynamic modelling considerations. The system dynamics models enable the quantity of each generated component such as nitric oxide to be estimated and predicted so effective and timely methods of pollution control may be developed. Furthermore, pollution generation in general is directly related to the growth of the population and hence can be related to the birth and death dynamics of the population. These basic models of the behaviour of simple populations utilise the standard tools of dynamical systems as outlined by Metz and Diekmann [22], McCauley and Murdoch [23] and Benenson [24]. Typical examples of population growth and death are the Logistic growth model, Predator–Prey models, and Lokta–Volterra competition. The Logistic growth model can and has been effectively used to model the generation of pollution. Finally, it is expected that the dynamic model developed will all be used for prediction, estimation and classification of the pollution generated and compared with the actual data.

ACKNOWLEDGEMENT

The first author wishes to acknowledge the financial support from the Pakistan Higher Education Council.

REFERENCES

[1] Murray, C.J. & Lipfert, F.W., Revisiting a population-dynamic model of air pollution and daily mortality of the elderly in Philadelphia. *Journal of the Air and Waste Management Association*, **60**(5), pp. 611–628. DOI: 10.3155/1047-3289.611, 2010.

[2] Wilcox, R.R., *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, 2nd ed., Springer: New York, 2010.

[3] Gyarmati-Szabo, J., Bogachev, L.V. & Chen, H., Nonstationary POT modelling of air pollution concentrations: Statistical analysis of the traffic and meteorological impact. *Environmetrics*, **28**(5), 2017. DOI: 10.1002/env.2449.

[4] Catalano, M., Galatioto, F., Bell, M. & Namdeo, A., Improving the prediction of air pollution peak episodes generated by urban transport networks. *Journal of Environmental Science and Policy*, **60**, pp. 69–83.

[5] Desalu, A.A., Gould, L.A. & Schweppe, F.C., Dynamic estimation of air pollution. *IEEE Transactions on Automatic Control*, **AC19**(6), pp. 904–910, 1974.

[6] Samaranayake, S., Glaser, S., Holstius, D., Monteil, J., Tracton, K. & Bayen, A., Real-time estimation of pollution emissions and dispersion from highway traffic. *Computer-Aided Civil and Infrastructure Engineering*, **29**, pp. 546–558. DOI: 10.1111/mice.12078, 2014.

[7] Ortiz, S.T. & Friedrich, R., A modelling approach for estimating background pollutant concentrations in urban areas. *Atmospheric Pollution Research*, **4**, pp. 147–156. DOI: 10.5094/APR.2013.015, 2013.
[8] Bai, L., Wang, J., Ma, X. & Lu, H., Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*, **15**(780), pp. 1–44, 2018. DOI: 10.3390/ijerph15040780.
[9] Liu, B., Luan S. & Gao, Y., Modeling the dynamics of a single-species model with pollution treatment in a polluted environment. *Discrete Dynamics in Nature and Society*, pp. 1–8, 2013. DOI: 10.1155/2013/412409.
[10] Leelőssy, A., Molnár, Jr. F., Izsák, F., Havasi, A., Lagzi, I. & Mészáros, R., Dispersion modeling of air pollutants in the atmosphere: A review. *Central European Journal of Geosciences* **6**(3), pp. 257–278, 2014. DOI: 10.2478/s13533-012-0188-6.
[11] UK Air Information Resource, Department for Environment, Food and Rural Affairs. https://uk-air.defra.gov.uk/air-pollution/. Accessed on: 2 Apr. 2019.
[12] Simon, D., *Optimal State Estimations*, John Wiley: Hoboken, NJ, 2006.
[13] Theil, H., A rank-invariant method of linear and polynomial regression analysis. *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen*, **A53**, pp. 386–392, 1950.
[14] Sen, P.K., Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, **63**, pp. 1379–1389, 1968.
[15] Wilcox, R.R., Some results on extensions and modifications of the Theil–Sen regression estimator. *British Journal of Mathematical and Statistical Psychology*, **57**(2), pp. 265–280, 2004.
[16] Wang, X. & Yu, Q., Unbiasedness of the Theil–Sen estimator. *Journal of Nonparametric Statistics*, **17**(6), pp. 685–695, 2005.
[17] Fernandes, R. & Leblanc, S., Parametric (modified least squares) and non-parametric (Theil–Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment*, **95**, pp. 303–316, 2005.
[18] Peng, H., Wang, S. & Wang, X., Consistency and asymptotic distribution of the Theil–Sen estimator. *Journal of Statistical Planning and Inference*, **138**(6), pp. 1836–1850, 2008.
[19] O'Shaugnessy, D., *Speech Communication: Human and Machine*, Addison Wesley, 1978.
[20] Brown, R.G. & Hwang, P.Y.C., *Introduction to Random Signals and Applied Kalman Filtering*, John Wiley, 1992.
[21] Daly, A. & Zannetti, P., Air pollution modelling: An overview, *Ambient Air Pollution*, Arab School for Science and Technology and The Enviro-Comp Institute: USA, 2007.
[22] Metz, J.A.J. & Diekmann, O., The dynamics of physiologically structured populations, 68. *Lecture Notes in Biomathematics*, Springer–Verlag: Heidelberg, 1986.
[23] McCauley, E. & Murdoch, W.W., Predator–prey dynamics in rich and poor environments. *Nature*, **343**, pp. 455–457.
[24] Benenson, I., Modeling population dynamics in the city: From a regional to a multi-Agent approach. *Discrete Dynamics in Nature and Society*, **3**, pp. 149–170, 1999.