Pollution 2022, 8(1): 269-279
DOI: 10.22059/POLL.2021.330761.1187
RESEARCH PAPER

Temporal Analysis and Forecast of Surface Air Temperature: case study in Colombia

Jhoana P. Romero–Leiton[1]*, Diego Torres[2] and Manuel Romero[3]

1. Facultad de Ingeniería, Universidad Cesmag, Pasto, Colombia
2,3. Fundación Universitaria los Libertadores, Bogotá, Colombia

Received: 19.09.2022, Revised: 25.10.2022, Accepted: 24.11.2021

ABSTRACT
In this work, we study the short-term dynamics of the Surface Air Temperature (SAT) using data obtained from a meteorological station in Bogotá from 2009 to 2019 and using time series. The data that we used correspond to the monthly mean of the historical registers of SAT and three pollutants. A descriptive analysis of the data follows. Then, some predictions are obtained from two different approaches: (i) a univariate analysis of SAT through a SARIMA model, which shows a good fit; and (ii) a multivariate analysis of SAT and pollutants using a SVAR model. Suitable transformations were first applied on the original dataset to work with stationary time series. Subsequently, A SARIMA model and a VAR(2) with its associated SVAR model are estimated. Furthermore, we obtain one-year forecasts for the logarithm of SAT in both models. Our forecasts simulate the natural fluctuation of SAT, presenting peaks and valleys in months when SAT is high and low, respectively. The SVAR model allows us to identify certain shocks that affect the instant relationships among variables. These relations were studied by the impulse-response function and the VAR model variance decomposition. Although the statistical methods used in this study are classical, they continue being widely used in the environmental field, presenting god fits, and the results obtained in this study are consistent with environmental theories.
Keywords: Time series, Pollutants, SARIMA, SVAR.

INTRODUCTION

SAT is an important variable that has been studied in a wide range of environmental applications including vector-borne diseases, bionomics, terrestrial hydrology, biosphere processes, climate change, among others (Benali et al., 2012). Even, the temporal changes of SAT have been used as a prominent indicator of global climate change (Aghelpour et al., 2019). The spatio-temporal SAT patterns can often be highly variable and complex due to the heterogeneity of the environmental factors that control the energy balance of the land-atmosphere system (Benali et al., 2012). These factors include the presence of certain atmospheric pollutants. The atmospheric pollutants that are normally measured in the urban atmosphere basically come from vehicles used for transport, from stationary sources of combustion (industries) and from waste disposal processes. Some of the most relevant atmospheric pollutants are suspended particles, sulfur compounds, nitrogen compounds, carbon oxides and photochemical oxidants. However, the presence of pollutants in the atmosphere not only affect the behaviour of SAT but they also affect public health. Most air

* Corresponding author Email: jpatirom3@gmail.com

pollutants have effects on human health, although their effects are different. Indeed, some studies have revealed that particulate matter (PM) can penetrate the respiratory system via inhalation, causing respiratory and cardiovascular diseases, reproductive and central nervous system dysfunctions, and cancer (Grivas et al., 2008). Therefore, many researchers have focused on studying SAT and the behaviour of atmospheric pollutants to obtain useful information to make forecasts and take control measures. In fact, information on SAT and atmospheric pollutant concentrations usually comes from automatic continuous monitoring stations. These stations capture records at different frequencies of the pollutant's concentrations and they also record the intensity of meteorological variables. Many mathematical and statistical studies have used these data; for instance, through multiple regression and interpolation methods, time series methods, geographical information systems (GIS) techniques, machine and deep learning methods (see e.g., (Alonso and Renard, 2019, Agbazo et al., 2019, Shen et al., 2020)), among others.

In this work, we will focus on time series techniques. Time series models have been widely used in a broad range of scientific applications to forecast variables, including climatology. Among the time series models frequently used on environmental studies, we have the Seasonal Autoregressive Integrated Moving Average (SARIMA) and the Vector Autoregressive (VAR) models (Aghelpour et al., 2019, Wang & Niu, 2009). Thus, the present study aims to analyse the individual SAT and the SAT atmospheric pollutants short–term dynamics using data reported for one monitoring station of the Bogotá Air Quality Monitoring Network (RMCAB, by its acronym in Spanish). This data is assessed by means of a descriptive analysis and then it is processed using an univariate and a multivariate time series models. For the univariate case, the SARIMA model (Section 2.3.1) is proposed according to the seasonal trend of SAT. This model shows a good fit in the errors behaviour and it is be able to forecast accurate monthly mean values for logarithm of SAT during all 2020. For the multivariate case (Section 2.3.2), a VAR model and its structural and unrestricted form SVAR are estimated. Thus, SAT is related with atmospheric pollutants by means of a restricted VAR(2) model and its associated SVAR model to determine potential instantaneous relations among the variables. The unrestricted form of a VAR model identifies the structural impacts and stores them in a matrix, which is obtained via data-driven identification techniques, such as Changes in Volatility (CV) (Rigobon, 2003). Diagnosis of the restricted VAR model has some fit problems, but instant relations between endogenous variables identified by the SVAR model are determined. Finally, a forecast from VAR(2) for all 2020 is obtained.

MATERIALS AND METHODS

Bogotá is the capital of Colombia and it is one of the largest cities of Latin America (see Figure 1). It is located in the center of the country at an altitude of approximately 2600 meters. The population in Bogotá for 2020 is now estimated at 10,978,360 and it has an annual growth rate of 2.08 (Lozano, 2004). Bogotá also has one of the highest environmental deterioration rates of the country. Air pollution has increased dramatically recently, due mainly to the uncontrolled increase in the number of vehicles in the city. Although air pollution has been monitored in Bogotá since 1967, it was not until 1990 that the monitoring stations were spread widely throughout the city. It has since been reported that the most important sources of pollution in Bogotá are automobiles followed by bricks and battery plants, among others. The areas with the highest levels of atmospheric pollution are Puente Aranda, Carvajal, and Kennedy, which are mostly affected by PM_{10} and $PM_{2.5}$.

Furthermore, in 1990 the Secretary of Health of the District with the collaboration of the Japanese International Cooperation Agency (JICA) identified for the first time the following components of air pollution in Bogotá: Sulfur Dioxide (SO_2), Nitrogen Oxides (NO_X), Total Suspended Particles (TSP), Carbon Monoxide (CO), Hydrocarbons (HC), and Ozone (O_3). Local authorities now face the challenge of supporting the growth and development of the city, while at the same time minimizing the adverse effects of the associated air pollution and its consequences on health. Permanent surveillance is an important tool in this process. The RMCAB currently operates 13 fixed and 1 mobile station throughout the city. The RMCAB measures P M using DASIBI 7001 (Met One Instruments, Inc., United States) and METONE–BAM 1020 (MetOne Inc., USA) β attenuation particle monitors for PM_{10} and $PM_{2.5}$, respectively.

Fig 1: A Google Maps view of Colombia (left–hand side) and close–up of Bogota (right– hand side) showing the location of the monitoring station Puente Aranda.

The data analysed in this work corresponds to Puente Aranda station (see Figure 1), which is one of the 15 stations that are part of the RMCAB that are spatially distributed throughout the city. Puente Aranda is one of the most representative stations because it is located in the heart of the industrial zone of the city and it is of mixed monitoring (i.e. it covers both air quality data and meteorological variables). Hourly records of several pollutants and meteorological parameters were obtained from Puente Aranda station from January 2009 to December 2019. However, due to the notable absence of records in many of their variables, only the measurement of SAT and three air pollutants: PM_{10}, O_3 and NO_2 with complete records greater than 60% were taken. A brief description of those pollutants follows: (a) : PM_{10}, is particulate matter 10 micrometers or less in diameter. Particles in this size range make up a large proportion of dust that can be drawn deep into the lungs. The specific effect of particles depends on their composition, concentration and the presence of other pollutants, such as acid forming gases. (b) O_3 is a highly reactive pale-blue gas that is formed in the Earth's lower atmosphere, near ground level. Ozone is formed when pollutants emitted by cars, power plants, industrial boilers, refineries, chemical plants, and other sources (usually NO_X and NOx) react chemically in the presence of daylight UV rays and is usually measured in *ppb*. (c) NO_2 is formed as a sub-product from burning fossil fuels and it is a precursor of ozone. Breathing air with a high concentration of NO_2 can irritate airways in the human respiratory system. These exposures over short periods can aggravate respiratory diseases,

particularly asthma, leading to delicate respiratory symptoms. Like ozone, NO_2 is usually measured in *ppb*. Thus, with this information and by using the **R** (version 1.1.463) statistical software, a new database was generated with the monthly average of each pollutant for each year. Subsequently, the monthly series of SAT and pollutants were generated.

RESULTS AND DISCUSSION

A descriptive analysis of the new database generated from the records selected by the Puente Aranda station during the specified period of time (with monthly records) was carried out. Table 1 contains monthly information on the mean, median, coefficient of variation (CV), and bias for each variable. In general terms, both the mean and median SAT tend to be higher in the first months of the year, while CV allows us to identify the data distribution, which is more homogeneous when this value is small (i.e. when the variation is less). SAT shows a small variation (i.e. the values this variable takes are close each other). Variables with a CV greater than 0.3 usually have heterogeneity problems and the presence of atypical data, as shown by O_3 (see Figure 2 (C)) in the months of April and May. From Figure 2, we can infer about the periodicity of SAT and the average pollutants due to the seasons. A very low temperature variation is observed in the month of January compared to other months of the year. September is the month with the lowest temperature (13°C) and February has the highest temperature (16°C). With regard to pollutants, Figure 2 shows the presence of harmful concentrations of PM_{10} (higher than $50 g/m^3$ almost all months) outside the limits established by Resolution 2254 of 2017 of the Ministry of Environment and Sustainable Development (MADS, by its acronym in Spanish), where it is specified that the maximum permissible

Table 1: Descriptive information of the variables.

		En	Feb	Mar	Ap	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
T_{max}	**Mean**	14,24	14,55	14,6	14,57	14,58	14,19	13,77	13,84	14,18	14,25	14,28	14,18
	Median	14,08	14,39	14,41	14,49	14,53	14,12	13,91	13,91	14,1	14,28	14,25	14,23
	CV	0,04	0,05	0,05	0,04	0,03	0,02	0,03	0,04	0,03	0,03	0,04	0,03
	Beas	1,24	0,8	0,91	-0,14	0,04	-0,39	-0,98	0,1	0,15	-0,05	-0,3	-0,07
PM_{10}	**Mean**	56,1	62,02	59,53	52,76	47,76	39,9	36,61	40,99	46,45	54,34	60,44	57,95
	Median	54,57	61,63	58,42	54,34	46,72	37,97	33,56	39,27	41,81	55	61,11	53,08
	CV	0,19	0,06	0,13	0,11	0,14	0,24	0,28	0,27	0,24	0,16	0,16	0,2
	Beas	0,25	0,01	-0,07	0,24	0,71	-0,1	0,42	0,16	0,88	1,04	0,91	1,12
O_3	**Mean**	10,5	10,95	9,62	8,23	6,087	6,87	6,97	8,66	9,87	8,42	7,196	7,33
	Median	10,5	10,95	9,62	8,23	6,08	6,87	6,97	8,66	9,87	8,42	7,19	7,33
	CV	0,27	0,24	0,3	0,52	0,49	0,33	0,24	0,21	0,21	0,28	0,31	0,19
	Beas	-0,12	0,86	0,31	2,23	1,54	1,37	0,98	-0,09	0,02	0,21	0,15	0,49
No_2	**Mean**	20,23	20,84	22,44	20,14	17,15	13,66	13,75	16,49	19,49	23,41	22,77	20,14
	Median	20,09	19,73	21,73	21,64	17,3	13,37	12,99	14,88	17,83	24,07	22,11	21,3
	CV	0,23	0,2	0,17	0,22	0,27	0,31	0,35	0,28	0,3	0,26	0,18	0,4
	Beas	0,14	0,19	-0,2	-0,66	0,02	0,14	1,09	0,69	1,24	-0,58	1,06	-0,82

annual level of PM_{10} is $50_g/m^3$, which represents a public-health problem in the sector. Finally, from Figure 2, we can identify that the pattern of SAT variation has a behaviour similar to PM_{10} and NO_2 because they tend to increase slightly in the first months of the year and then decrease at the middle of the year and then register a new peak in October.

To generate the SAT time series, which was used later to do univariate forecasts, we used the **TSA** library of **R**. Usually, meteorological time series such as SAT are non-stationary.

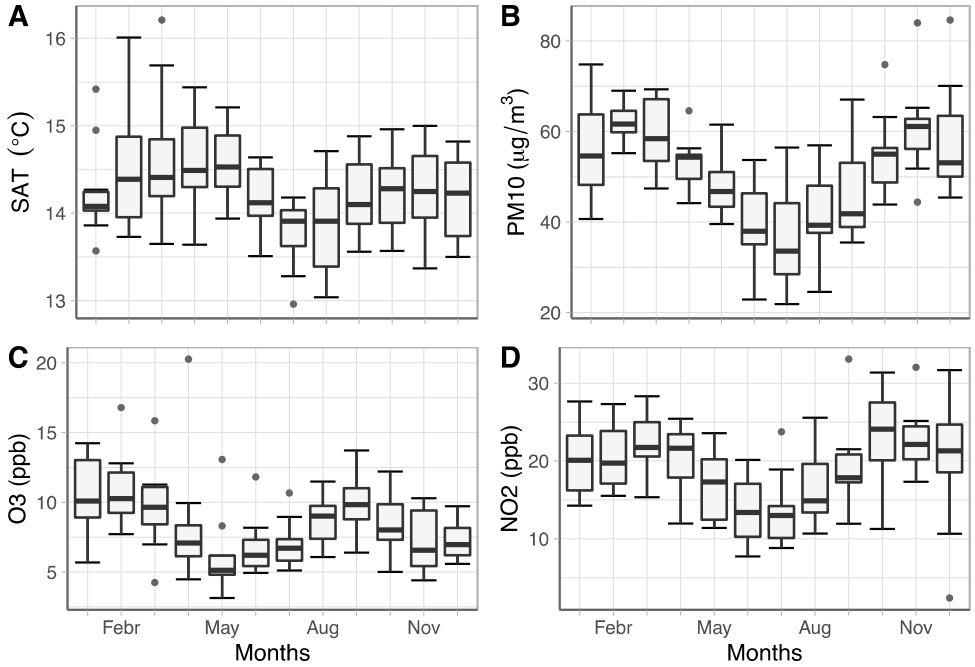

Fig 2: Box–plots for SAT and pollutants.

They do not oscillate around a constant mean (stationary in mean) and they do not make this oscillation among a constant interval (stationary in variance). Having said that, our original SAT time series was non-stationary, but it is necessary to be stationary to be able to generate predictions. Now, The Box–Cox test (Davidson & MacKinnon, 1985) was used to determine whether SAT time series was stationary in variance or not, and also to determine its possible transformation coefficient, from where it is suggested to use the logarithm of the data. To correct the non-stationary mean problem, we used the **autoarima** function to determine the number of differentiations needed. Finally, we determined that the logarithmic SAT time series had to be twice differentiated: once for ordinary differentiation and again for seasonality. The autoarima function also suggests the order of the SARIMA model which would fit the best. Then, the ACF and PACF of logarithms of SAT were obtained, while differentiations were applied fitting the model. From Figure 3 (A), we can see that ACF plot suggests an AR(5) model due to the exponential decay, being significant up to the fifth lag and new peaks every 12 lags; the same order was suggested by the autoarima function. From Figure 3 we can see two peaks: one at the first lag and the other near lag 24, indicating seasonality.

Subsequently, a SARIMA model with non-seasonal (5.2.0) and seasonal (2.0.0) parameters with 12 cycles per year is proposed (SARIMA(5, 2, 0)×(2, 0, 0)^{12}). Its diagnosis consisted in Ljung–Box test, which is usually applied to the residuals of an ARMA(p, q) fit to observe

how they are behaving, from where a p-value= 0.53 was obtained, and the references suggest a better approximation to the null–hypothesis distribution.

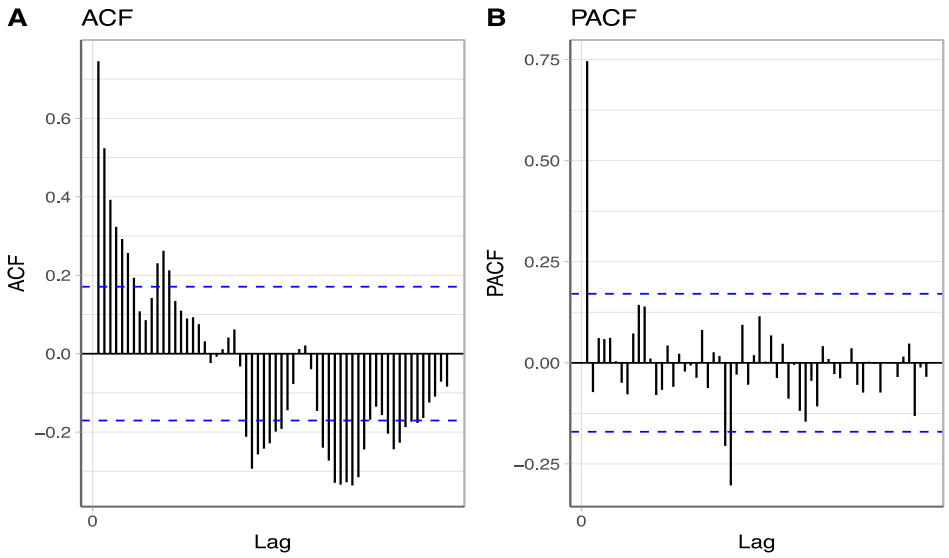

Fig 3: ACF and PACF form the logarithm of SAT.

To make forecasts, we used the **forecast** library of **R**, which allowed us to generate predictions from past records. We graphically show the forecasts of the obtained SARIMA model from the behaviour of its residuals, where their normality and non-correlation are measured in Figure 4. Additionally, Table 2 shows the logarithm of SAT forecasts for one year and Table 3 contains the range of summary measures of the forecast accuracy. To obtain raw forecast of SAT, the inverse function of the logarithm must be applied insomuch as the logarithm is a reversible function.

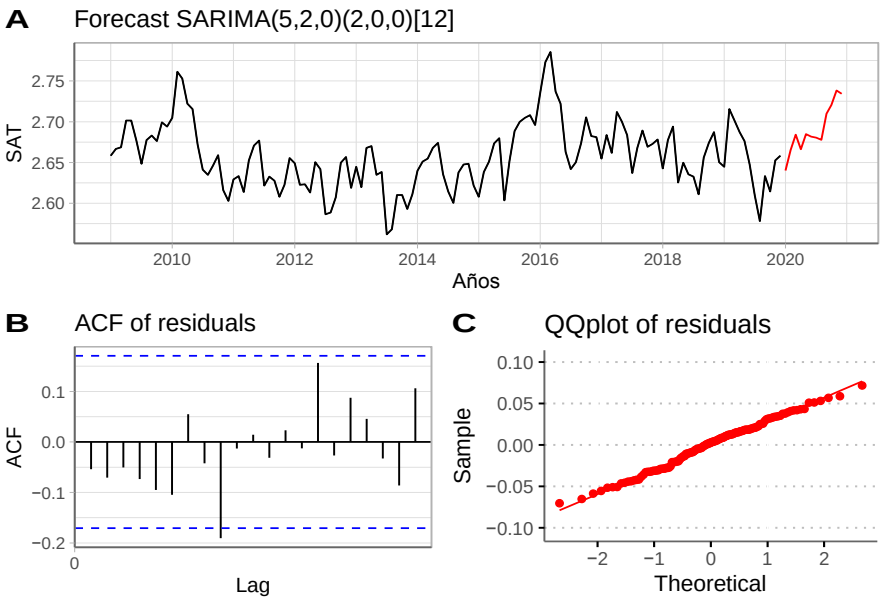

Fig 4 A. One-year forecast for the logarithm of SAT from the SARIMA model. **B**. Residuals ACF. **C**. Residuals Qqplot of the residuals. In **B** and **C** the normality of the residuals is observed.

Table 2: One–year predictions of the logarithm of SAT from SARIMA model.

	Ene	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Prediction	2.64	2.665	2.684	2.666	2.685	2.682	2.68	2.679	2.71	2.72	2.738	2.734

Table 3: SARIMA model error measurements with out-sample for SAT forecasts

ME	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**
6.28×10^{-5}	0.029	0.024	1.09×10^{-3}	0.895	0.647	-0.054

With respect to the time series used in this section, we could observe that after using logarithm transformations all time series are stationary. However, due to the stability of a VAR process, this transformation usually is not needed. Nevertheless, the logarithm was applied to obtain forecasts with the same units both in univariate and multivariate cases. Subsequently, we proceeded to determine de p order of our VAR(p) model. For this end, we used the **R** function **VARselect** from **vars** package, which internally compares among possible VAR(p) models with different p order, taking account the least AIC, BIC and FPE value for each model. At the end, the p order suggested by the information criteria was VAR(1). However, due to the low and insufficient number of lags taken account by VAR(1) model, VAR(2) was finally selected. Referring to diagnosis of VAR(2) model, heteroscedasticity and no-correlations of residuals were proved, but it is inconvenient because of its normality. This model also presents significant problems in its coefficients, possibly due to the fact that it is not capable of identifying contemporary structures hidden in the residuals by itself. Therefore, unrestricted VAR form (SVAR) is proposed.

When a shock of SAT occurs, all of the variables respond positively. Then, variables such as PM_{10} increase until the third month and start decaying exponentially. Other variables such as NO_2 decrease rapidly until obtain a negative response, they then increase until they finally attenuate. Whereas, when we see the response of SAT after a shock of itself and the different pollutants, it responds positively but it then decreases; in cases as PM_{10} and O_3, SAT response decreases enough to get a negative answer between third and seventh month for O_3 and fourth and tenth month for NO_2, respectively. Meanwhile, for a shock of NO_2, SAT responds negatively but it grows and decay quickly for the first tree months and then it has a negative response until it attenuates. It is interesting the reaction from O_3 when a shock of NO_2 occurs, because at the beginning the response is negative but it then grows rapidly until it reaches stability. This might be explained by the fact that NO_2 is a chemical precursor of O_3, referring that in presence of NO_2, O_3 tends to form in the environment. Figure 5 shows the instantaneous reciprocity observed between pollutants and SAT, where it is clearly distinguished that SAT influences the concentration of pollutants.

Figure 6 shows the FEVD, which consists of determining, graphically, how the systems of equations interact within the model (Lütkepohl, 2006). The variance decomposition is useful for forecasting errors and for visualizing the instantaneous impact on the relationship between the variables due to a shock to itself or to the shock of other variables, so that the FEVD acts as a complement to the impulse-response functions. If a variable largely explains its variance with its same innovations, then this variable will be of a more exogenous level than the others.

Thus, the variance decompositions of the four variables forecasts is presented. It can be seen that SAT explained around 50% of the forecast variance of itself, the rest appears to be explained by the pollutants. Additionally, 70% of PM_{10} was explained by O_3 and itself, and by SAT in less measure. The O_3 variance forecast mostly depended on NO_2 and PM_{10}, and

its relationship with its own lags and SAT were almost marginal. Finally, the contribution of NO_2 in the variation of the other variables was minimal, being the most exogenous variable. Although even if one of the pollutants seems to be exogenous, the three pollutants appear to be useful in explaining the variation in SAT forecasts.

Finally, due to the limited scope of SVAR models to generate forecasts, our predictions were obtained from the associated VAR model; the results for each variable are shown in Figure 7. Likewise, Table 4 shows the results for the logarithm forecasts of SAT via VAR(2) model and Table 5 contains its associated range of summary measures of the forecast accuracy.

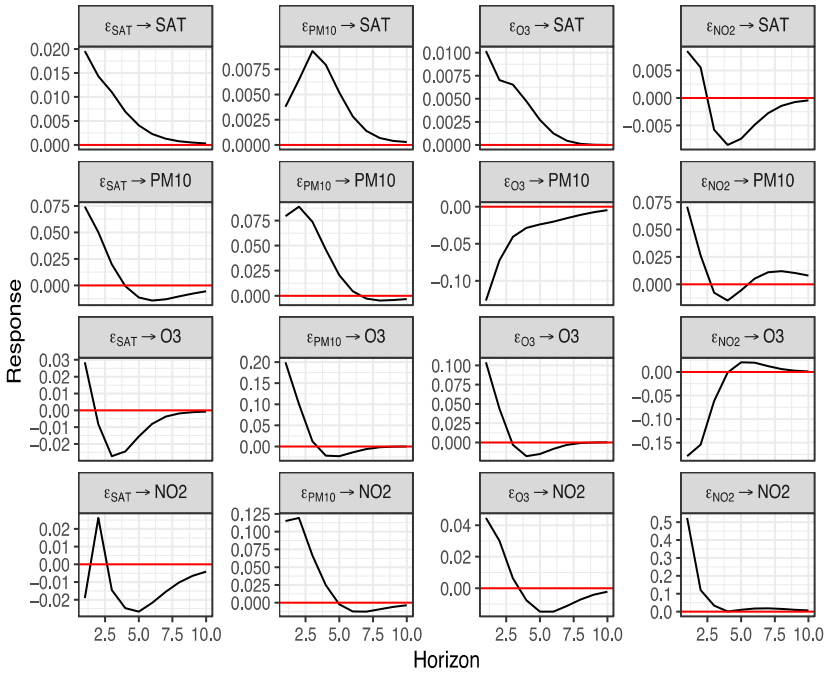

Fig 5: Individual impulse–response diagrams obtained from the SVAR model that show the periodicity of impulses due to the presence of seasons during the year, and the response variables.

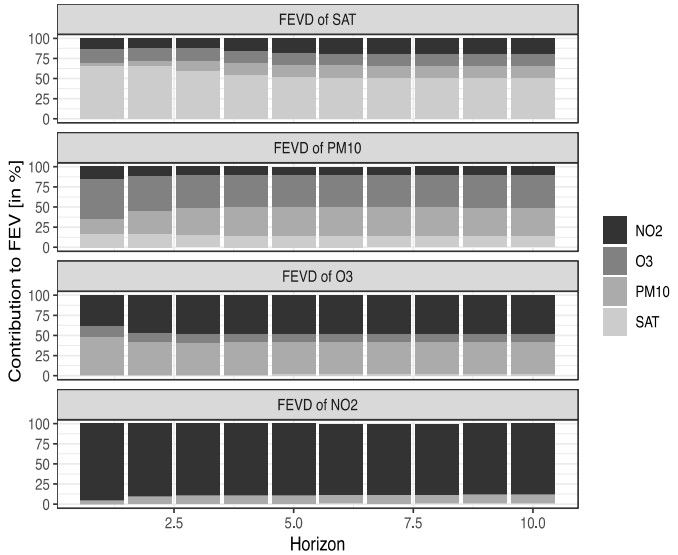

Fig 6: FEVD diagrams: the instantaneous influence between variables is observed.

Table 4: One-year predictions of the logarithm of SAT from associated VAR(2) model.

	Ene	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Prediction	2.67	2.69	2.694	2.693	2.694	2.667	2.636	2.64	2.663	2.667	2.669	2.662

Table 5: VAR(2) model error measurements with out-sample for SAT forecasts.

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2.5×10^{-17}	0.022	0.017	6.95×10^{-3}	0.666	0.481	-0.024

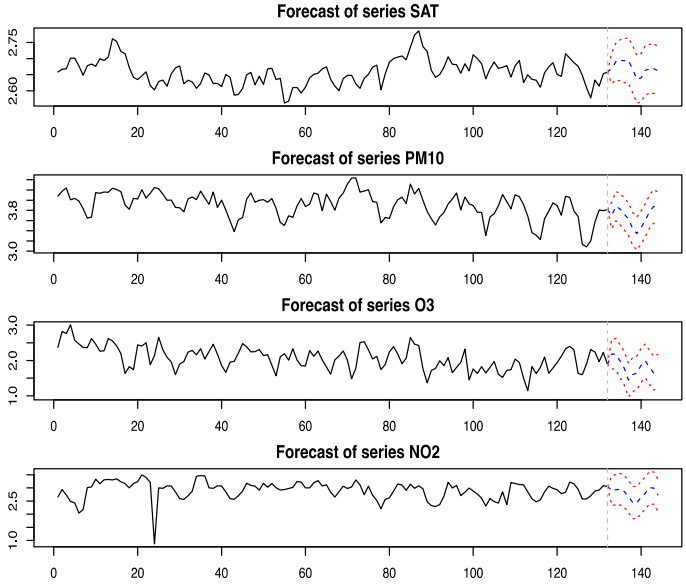

Fig 7. One-year forecast of the SAT logarithm and pollutants from the VAR(2) model associated with SVAR model.

CONCLUSION

In general, no significant variations in SAT were observed over time, which is possibly due to the fact that the interval of time worked is small. With respect to the models obtained, the SARIMA univariate model was more stable, it presented a good diagnosis, and it adequately represents the internal seasonal variation. However, it clearly falls short when it comes to establishing relationships of dependency and causality with other factors that affect atmospheric dynamics. The multivariate SVAR model generated significant results that explain the co-integration between SAT and pollutants. However, the associated VAR(2) model presented certain violations of some assumptions of normality in the residual analysis, although it presented stability in its structural part. This may be due to the non-linear nature of the data. This would indicate that the estimation of a model that fits better could occur, for example, in the analysis of the fractal behaviour in meteorological variables; as proposed in (Agbazo et al., 2019).

Regarding the forecasts, both models generated predictions of the logarithm of SAT similar to each other. In both cases, the natural fluctuation recorded in SAT from month to month is correctly followed, coinciding with the peaks and valleys in the months with high and low temperatures, respectively. Nevertheless, because of the precision of the forecasts, the VAR(2) model was more accurate in its results. Although direct forecasts could not be obtained from SVAR models, it is evident that they are very useful for exploring the implicity of certain theoretical restrictions on the dynamic behaviour of our interest variables. Finally, in future works it would be interesting to generate this type of analysis at a regional scale,

taking into account the geographical distribution of the information collected, which is usually how atmospheric dynamics are studied. This could be achieved through the implementation of geostatistical methods (GIS systems) with more than one meteorological station data, also considering geomorphology and relief, or the implementation of other meteorological variables or other atmospheric pollutants to the model.

ACKNOWLEDGEMENT

The authors thank to RMCAB for the free data availability. Jhoana P. Romero thanks to Fundación Ceiba, Colombia. D. Torres and M. Romero thank to Fundación Universitaria los Libertadores, Colombia.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Agbazo, M., Koto N'gobi, G., Alamou, E., Kounouhewa, B., Afouda, A. and Kounkonnou, N. (2019). Multifractal behaviors of daily temperature time series observed over benin synoptic stations (west africa). Earth Sciences Research Journal, 23(4):365–370.

Aghelpour, P., Mohammadi, B. and Biazar, S. M. (2019). Long- term monthly average temperature forecasting in some climate types of iran, using the models sarima, SVR, and SVR-FA. Theoretical and Applied Climatology, 138(3-4):1471–1480.

Alonso, L. and Renard, F. (2019). Integrating satellite–derived data as spatial predictors in multiple regression models to enhance the knowledge of air temperature patterns. Urban Science, 3(4):101.

Benali, A., Carvalho, A., Nunes, J., Carvalhais, N. and Santos, A. (2012). Estimating air surface temperature in portugal using modis lst data. Remote Sensing of Environment, 124:108–121.

Box, G. E. and Jenkins, G. M. (1976). Time series analysis: forecasting and control revised. Holden–Day.

Grivas, G., Chaloulakou, A. and Kassomenos, P. (2008). An overview of the pm10 pollution problem, in the metropolitan area of athens, greece: assessment of controlling factors and potential impact of long range transport. Science of the Total Environment, 389(1):165–177.

Lozano, N. (2004). Air pollution in Bogotá, colombia: a concentration– response approach. Revista Desarrollo y Sociedad, (54):133–177.

Ninyerola, M., Pons, X. and Roure, J. M. (2000). A methodological approach of climatological modelling of air temperature and precipitation through gis techniques. International Journal of Climatology: A Journal of the Royal Meteorological Society, 20(14):1823–1841.

Rigobon, R. (2003). Identification through heteroskedasticity. Review of Economics and Statistics, 85(4):777–792.

Shen, H., Jiang, Y., Li, T., Cheng, Q., Zeng, C. and Zhang, L. (2020). Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data. Remote Sensing of Environment, 240:111692.

Singh, K. P., Gupta, S. andRai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environ- ment, 80:426–437.

Wang, W. and Niu, Z. (2009). Var model of pm2.5, weather and traffic in los angeles–long beach area. In 2009 International Conference on Environmental Science and Information Application Technology, volume 3, pages 66–69. IEEE.