Modelling multiple seasonalities with ARIMA: Forecasting Madrid NO2 hourly pollution levels

Matias Luis Avila (matiasluis.avila@alumnos.uc3m.es)
Carlos III University of Madrid

Andres M. Alonso
Carlos III University of Madrid

Daniel Peña
Carlos III University of Madrid

Research Article

Keywords: Time series, multiple seasonalities, ARIMA, NO2 pollution forecasting, Madrid Spain, Prophet

Posted Date: April 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2860239/v1

Additional Declarations: No competing interests reported.

Modelling multiple seasonalities with ARIMA: Forecasting Madrid NO_2 hourly pollution levels

M.L. Avila[*], A.M. Alonso[†], D. Peña[‡]

April 2023

Abstract

Multiple seasonalities often appear in high-frequency data. In this context multiple seasonal components are usually modelled in a deterministic way by trigonometric functions or dummy variables. This assumption may be too strict. Instead, a more flexible model is to allow the seasonality to slowly change as a seasonal Autoregressive Integrated Moving Average model, where the seasonality is modelled as a stochastic processes. In this study, we propose to model them iteratively, combining different seasonal Autoregressive Integrated Moving Average models. To this end, we test the proposed methodology with Madrid's NO_2 hourly measurements of pollutants with daily, weekly and annual seasonalities, due to human activity and weather conditions. Here, we demonstrate the usefulness of our approach by comparing it with other methodological approaches proposed for this type of data. In an extensive exercise involving 15-year hourly forecasts, we show that the proposed procedure performs very well in predicting hourly pollution over a 24-h horizon and improves on alternative procedures. Additionally, the impact on the predictions of covariates such as wind speed, temperature and festivities were evaluated.

Keywords: Time series, multiple seasonalities, ARIMA, NO_2 pollution forecasting, Madrid Spain, Prophet.

1 Introduction

Nitrogen dioxide (NO_2) is one of the most common pollutants in urban areas, a toxic gas potentially responsible for short and long-term injuries to the respiratory system (World Health Organization, 2006). Road traffic is the main source of this contaminant since most of it is generated under high temperatures, such as the ones that occur inside a combustion engine (Querol, 2018). We can distinguish a direct origin of this pollutant in emissions from motor vehicles and industrial processes, and an indirect source through the chemical oxidation reaction of nitrogen oxide (NO) into NO_2 due to ozone (O_3) in the atmosphere (Kurtenbach et al., 2012).

In many cities and regions continuous monitoring of NO_2 levels are carried out and there are action protocols in case high concentration levels are observed or expected (see, for instance, González-Enrique et al. 2021 and Medrano et al. 2021). The main objective of this study is to predict the time series of hourly levels of NO_2, which would serve as a complementary tool to the action protocols. When modelling this univariate time series we will also include some explanatory variables as covariates, such as the current average wind speed and a dummy variable that differentiate working days from holidays.

One of the most common methods for forecasting is to apply Arima Autoregressive Integrated Moving Average (ARIMA) models (Box and Jenkins, 1976) often with a single seasonality. Other methods have been developed to consider more seasonalities such as TBATS (Livera et al., 2011) and Prophet (Taylor and Letham, 2017), where the seasonal components are modelled by trigonometric functions. Assuming that a deterministic seasonality may be too strict, a more flexible model is to allow the seasonality to slowly change as in a seasonal ARIMA model, where the seasonality is modelled as a stochastic processes. Since the TBATS implementation does not allow for the inclusion of covariates, we will compare our methodology with Prophet. Moreover, Prophet has been recently used to model pollutants time series (Shen et al., 2020).

The rest of the paper is organised as follows: Section 2 reviews the literature regarding modelling time series with multiple seasonalities. We also comment on methods dealing with the forecasting of NO_2 pollution. Section 3 describes the NO_2 time series as well as the covariates that will be used in this work. Section 4 presents our proposed methodology. Section 5 describes the alternative method used as a benchmark model (Prophet). Section 6 compares our methodology with Prophet in our data. Section 7 gives the main concluding remarks.

[*]Department of Statistics, University Carlos III of Madrid, Madrid, Spain. matiasluis.avila@alumnos.uc3m.es 0009-0003-2665-2242

[†]Department of Statistics and Institute Flores de Lemus, University Carlos III of Madrid, Madrid, Spain. amalonso@est-econ.uc3m.es 0000-0002-6112-2867

[‡]Department of Statistics, University Carlos III of Madrid, Madrid, Spain. daniel.pena@uc3m.es 0000-0002-9137-1557

2 Background and related work

Air pollutants such as NO_2 represent a great opportunity to model multiple seasonalities time series as shown in Section 3.1.2. When forecasting pollution levels we find two main approaches. On one hand, we have mathematical formulations of the various physical and chemical processes that mediate the changes in pollutant concentrations, mainly carried out by meteorologists (Nieuwstadt and van Dop 1982 and Baklanov and Zhang 2020). On the other hand, we have statistical techniques establishing relations and patterns among historical data. Here we find different methods, such as using linear regression (Agirre-Basurko et al., 2006), ARIMA (Garg and Jindal, 2021), Prophet (Shen et al., 2020) and (Garg and Jindal, 2021), Artificial Nerural Network (ANN) (Agirre-Basurko et al. 2006 and González-Enrique et al. 2021), Convolutional Neural Network (CNN) (Garg and Jindal, 2021), Long Short Term Memory (LSTM) (Garg and Jindal 2021 and González-Enrique et al. 2021) or an ensemble of multiple models predictions (Medrano et al., 2021). None of these methodologies addresses the multiple seasonalities modelling, although they obtain good forecasting performance. Unlike statistical-ML approaches, physical-chemical models explicitly consider the relationship between precursor emissions, the resultant pollution and the chemical interactions between them and other factors. Therefore, they can predict pollution levels during unusual meteorological conditions and circumstances (i.e. road traffic reduction during COVID-19 lockdown). The main drawback of these models is that they require the establishment of the linkages between precursor emissions and final pollution, accurate meteorology data and incorporation of physical and chemical processes, such as diffusion. Compared with physical-chemical models, statistical methods can be more convenient and practical.

Traditional statistical approaches employed to modelling seasonal patterns include the Holt-Winters exponential smoothing approach, i.e. applied in sales forecasting (Winters, 1960), and the ARIMA models, which can easily accommodate a seasonal dependency. Technically, an ARIMA model can be extended to obtain models with two or more periodic components to take into account multiple seasonal cycles (Box and Jenkins, 1976), for example, we can estimate a double seasonal ARIMA (Mohamed et al., 2011). Regarding the exponential smoothing approach, Taylor (2003) adapts the Holt-Winters method to accommodate and forecast short-term electricity demand with a double seasonal exponential smoothing (intraday and weekly cycles) at the expense of assuming the same daily cycle for each of the seven days of the week. Despite these considerable improvements, the incorporation of new seasonality extensions in these types of approaches may produce over-parameterisation and estimation problems. Similarly, another way to include the seasonal dependencies is to add seasonal lagged observations as covariates, for example, by adding the different lagged values of the target variable as predictors in an ANN (Allende et al., 2002).

An alternative strategy is to decompose the time series into different components and model each of them separately, instead of modelling the original time series as a whole. The main difference between this methodology and a model seasonal dependency, i.e. in a seasonal ARIMA, is that coefficients are estimated jointly, while time series decomposition is carried out by a parallel estimation. For example, we can stratify the time series based on a calendar frequency like Ramanathan et al. (1997), where they forecast electricity loads by modelling each hour of the day separately. Another illustrative example is the lifting scheme used by Lee and Ko (2011) for load forecasting. They decomposed the historical load series into different load subseries at different resolution levels, displaying the different frequency characteristics of a load, and modelled each subseries separately. In theory, when decomposing and modelling each component separately with any of the previously mentioned techniques, we benefit from reducing the model complexity compared to forecasting the original time series as a whole. However, by stratifying the forecasting task the interaction between the different seasonal components is neglected. In these approaches, the seasonalities are deterministic and can be modelled by trigonometric functions via Fourier series or dummy variables. This assumption may be too strict, but allow an easy way to address the multiple seasonality problem. Some approaches, i.e. TBATS and Prophet, use trigonometric functions to deal with seasonalities and are often used in modelling multiple seasonalities. Unlike Prophet, TBATS does not allow to include covariates, nor special calendar effects such as holidays.

Finally, another alternative is to apply seasonal differencing, which removes the gross features of seasonality from a series, as well as most of the trend, leaving a time series composed of the changes from one season to the next.

ANN have received a great deal of attention and turned into an alternative to classical statistical methods, but removing the seasonal (Nelson et al., 1999) and trend (Zhang and Qi, 2005) components were needed to achieve better forecasting results. Moving from simple ANN to deep ANN, forecasting practitioners started focusing on CNN and Recurrent Neural Network (RNN) and, more recently, towards the LSTM which is an evolution over traditional RNNs. Despite its vast and increasing literature, few articles explicitly address the multiple seasonalities modelling with Deep Learning (Bandara et al., 2021).

In this paper, we propose to model the time series by a periodic model with multiple seasonalities iteratively combining different seasonal ARIMA models (see Section 4.1). Instead of using trigonometric functions or dummy variables, a more flexible model is to allow the seasonality to slowly change as in a seasonal ARIMA model, where

the seasonality is modelled as a stochastic process. Similar to the time series decomposition approach, we propose to stratify the original series into 168 subseries (one per day hour and weekday) and model each with a properly fitted ARIMA model. However, unlike simple decomposition techniques, we go one step further, reconstructing the stratified series from the previous step and modelling the remaining patterns such as the regular dependency in the second step.

A similar approach was used by Caro et al. (2020) to forecast short-term electricity load. However, both approaches differ since: (1) Our approach is more flexible since we consider AR and MA components in both steps. We also allowed the AR and MA coefficients to vary based on the day hour and week day while they only consider variation based on the day hour. (2) We consider three seasonalities while they only considered two. (3) The regular dependence between one hour and the next in our approach comes from the multiplicative structure of ARIMA models, while their approach allows the regular coefficients to vary depending on the hour using a periodic autoregressive model.

3 Exploratory Analysis and Data

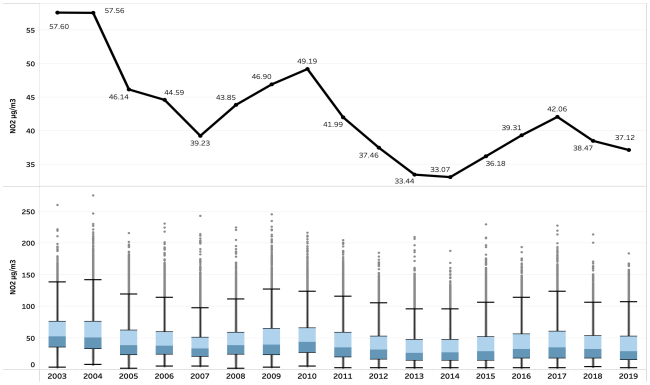
3.1 Pollution

The series we are interested in forecasting corresponds to the NO_2 hourly observations in the Barajas Village air quality measuring station, located in the district of Barajas (Longitude: 3º 34' 48.10" W. Latitude: 40º 28' 36.93" N. Altitude: 620 m) in the northeast of Madrid, Spain. These data cover 17 years of hourly values and is obtained from the Madrid City Council's website. The sample covers the period from 01-01-2003 at 00:00 until 31-12-2019 at 23:00 (149016 observations). The historical record from this station is one of the longest among the 24 stations located in Madrid and it contains very few missing measurements. The Madrid's Air Quality System, run by Madrid's City Hall, is in charge of the stations' maintenance as well as responsible for gathering, storing, validating and analysing the pollution measurements.
It is worth noting that these hourly observations are not the NO_2 emissions but the concentration of this compound (expressed in $\mu g/m^3$) released into the atmosphere where it will interact with different meteorological phenomena which will alter its concentration and spread.

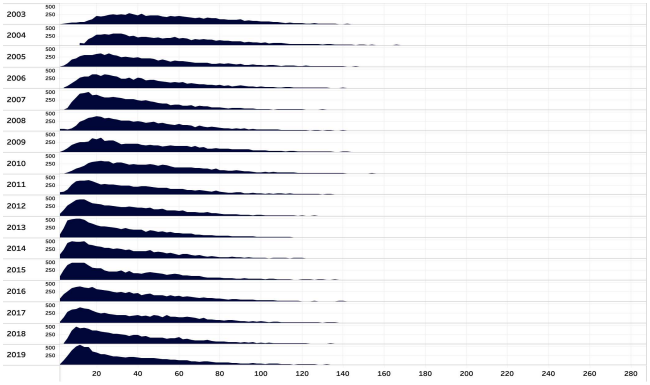
Fig. 1: Barajas Village air quality station (purple dot). Barajas Airport meteorological station (red triangle).

3.1.1 Trend

In Figure 2a we can see a global decreasing trend of NO_2 through the years combining periods of growth, fall and stagnation. For each year, the distribution of the hourly pollution measurements (Figure 2b) seems to be uni-modal with positive skewness. As the annual average pollution increases (decreases) so does its distribution shifting the distribution towards the right (left) and vice versa. Due to the nature of our data, pollution values are constrained on the first quadrant, being the minimum concentration level of NO_2 *0* $\mu g/m^3$ since negative values are preposterous.



(a) (Top) Annual average NO_2 $\mu g/m^3$ levels. (Bottom) One boxplot per year showing the hourly NO_2 $\mu g/m^3$ measurements.

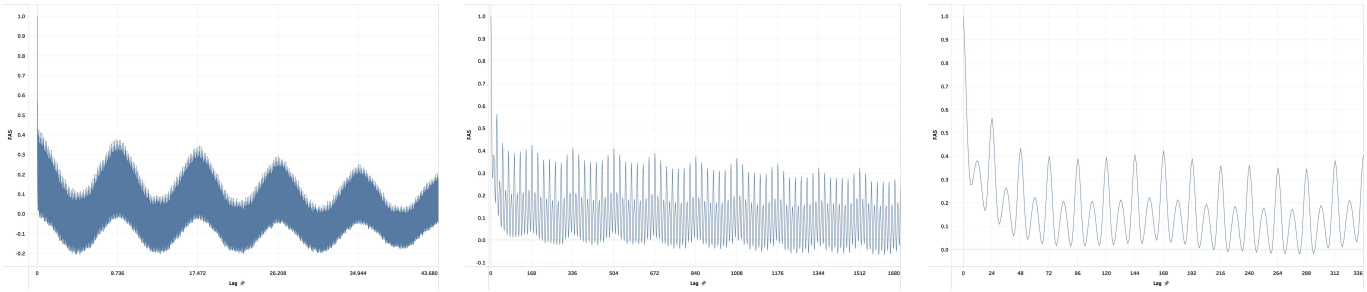


(b) One histogram per year showing the hourly NO_2 $\mu g/m^3$ measurements distribution.

Fig. 2: Descriptive analysis of hourly measurements of NO_2 $\mu g/m^3$ stratified by year, 2003–2019.

3.1.2 Seasonalities

We work out the Simple Autocorrelation Function (SAF) of the NO_2 time series and plot the autocorrelation values (see Figure 3) of the first 43680 lags (5 years), 1680 lags (10 weeks) and 336 lags (14 days). We can appreciate an annual seasonality in Figure 3a every 8736 lags (24 hours x 7 days x 52 weeks). If we zoom in, Figure 3b, we can see a weekly seasonality every 168 lags (24 hours x 7 days) and in Figure 3c a daily seasonality with two peaks spaced by 12 lags every 24 lags both of them.



(a) SAF first 43680 lags **(b)** SAF first 1680 lags **(c)** SAF first 336 lags

Fig. 3: Simple Autocorrelation Function (SAF) plots of the NO_2 pollution levels.

Figure 4 represents what will be called the Seasonal Decomposition Plot (SDP) of the average values of the series in the 17 years of the sample. These average values are computed for each possible seasonal effect, in this case for each given hour, each week of the year and each day of the week. This figure provides a visual understanding of the three seasonalities. (a) First, the columns show the seasonality of each day of the week within each year, where pollution levels are higher in winter than in summer. (b) Secondly, the rows show the strong weekly seasonality, where pollution drops during the weekend. (c) Third, each cell shows the daily seasonality that is complex and changes during the year and during the week. It has two peaks. The first one is relatively stable, but the second one swings significantly and especially along with the annual seasonality. In winter (summer) the first peak is around 8:00 (7:00) hours while the second one occurs around 19:00 (22:00) hours. It is worth noting that both peaks, within the same day, are closer during winter and farther away in summer. Similarly, the second peak is closer to the next day's first peak in summer and farther away in winter. Within the Saturday and Sunday (Monday to Friday) the first peak is around 7:00 (8:00) hours while the second peak occurs around 22:00 (21:00) hours.

We can average the hourly pollution observations by different time units such as months (Figure A.3b), weekdays (Figure A.4b) and day hours (Figure A.5b) so that the annual, weekly and daily seasonalities emerge from the data (Figures A.* can be seen in the appendix). Another way, instead of averaging, is to compute a given quantile at each time unit and obtain a Timewise Quantile (TQ). For example, when plotting one boxplot per day hour sequentially (24 boxplots in total), we can spot the daily seasonality with both peaks thanks to the pattern drawn by the $1^{st}Quartile$, the $Median$ and the $3^{rd}Quartile$ of each boxplot which define the .25, .50 and .75 TQ, respectively (Figure A.5a). However, these TQ do not show the dynamics of the real daily pollution pattern since they are estimated at each day hour independently without considering previous nor following hours. Instead of plotting TQ we can use the Empirical Dynamic Quantiles (EDQ) proposed by Peña et al. (2019). The EDQ are defined as the time series in the set closer to the desired quantile. In one dimension, we can partition a finite set of points into q subsets of (nearly) equal sizes, each delimited by specific cut points known as quantiles. Likewise, in two dimensions we can divide a set of lines into q subsets of a nearly equal number of lines delimited by cut lines which are the EDQ. Figure 5 shows the .05 (Thursday 13-08-2015, light blue), .25 (Saturday 18-11-2006, blue), .5 (Monday 26-03-2007, red), .75 (Thursday 15-05-2008, green) and .95 (Wednesday 25-01-2017) EDQ.

To further study the seasonalities we introduce the following notation. We denote the NO_2 pollution time series in the time period $t = 1, ..., T$ by $\boldsymbol{Y}$ (Equation 1). Each observation, y_t in $\boldsymbol{Y}$ belongs to an hour $h \in \{0, 1, \ldots, 23\}$ in a weekday $d \in \{1, 2, \ldots, 7\}$ (Equation 2), therefore, there are 168 observations per week. We will denote the week number as w. For simplicity in the notation, we will assume that $t = 1$ corresponds to the hour $h = 0$, the day $d = 1$ (in our case, Wednesday) and the week $w = 1$. The relationship between y_t and $y_w^{(h,d)}$ is as follows: $h = ((t-1) \mod 24)$, $d = (((t-1) \operatorname{div} 24) \mod 7) + 1$ and $w = ((t-1) \operatorname{div} 168) + 1$, where div is the Euclidean division quotient and mod is the modulo operation.

$$\boldsymbol{Y} = y_1,\ y_2,\ y_3,\ \ldots,\ y_{24},\ y_{25},\ y_{26},\ \ldots,\ y_{168},\ y_{169},\ y_{170},\ \ldots \tag{1}$$

$$\boldsymbol{Y} = y_1^{(0,1)}, y_1^{(1,1)}, y_1^{(2,1)}, \ \ldots\ , y_1^{(23,1)}, y_1^{(0,2)}, y_1^{(1,2)}, \ \ldots\ , y_1^{(23,7)}, y_2^{(0,1)}, y_2^{(1,1)}, \ \ldots \tag{2}$$

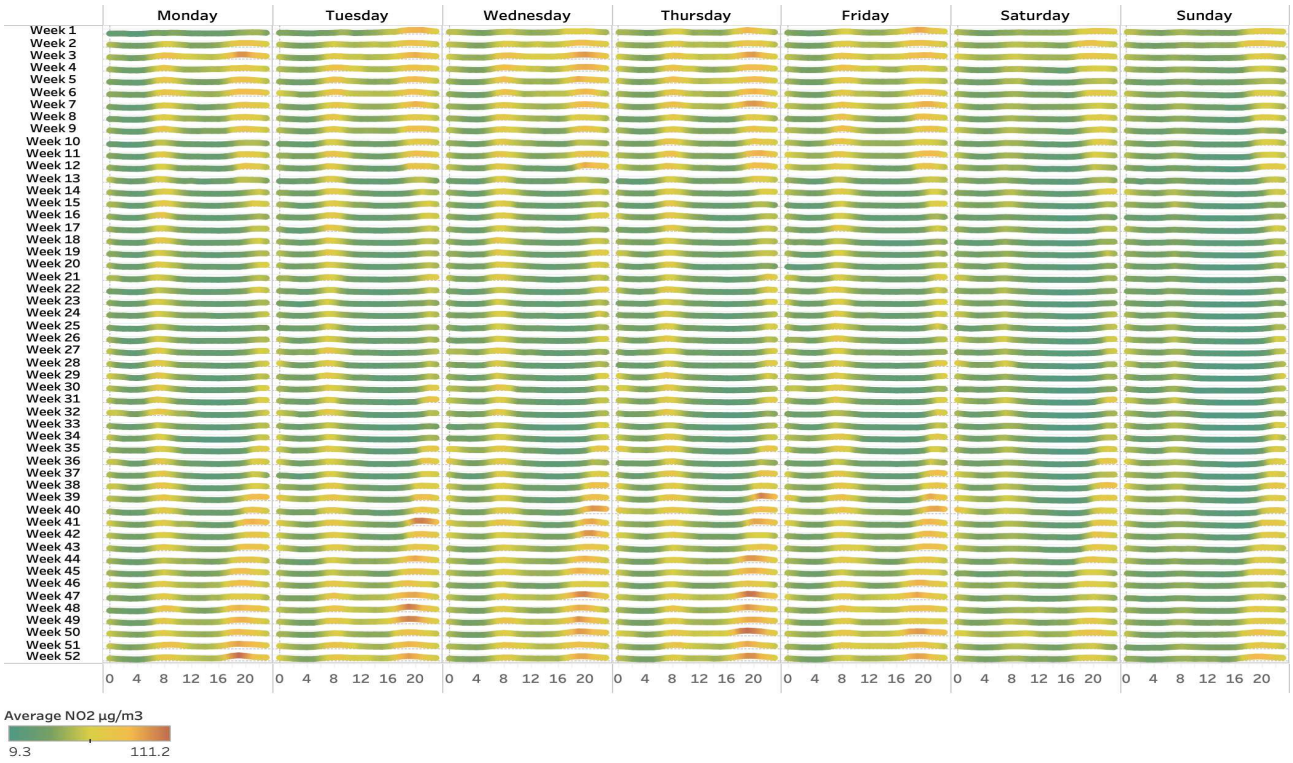


Fig. 4: Seasonal Decomposition Plot of the NO_2 $\mu g/m^3$ levels. One row per week of the year and one column per weekday and each cell includes a line with the average values over the 17 years at each of the 24 hours.

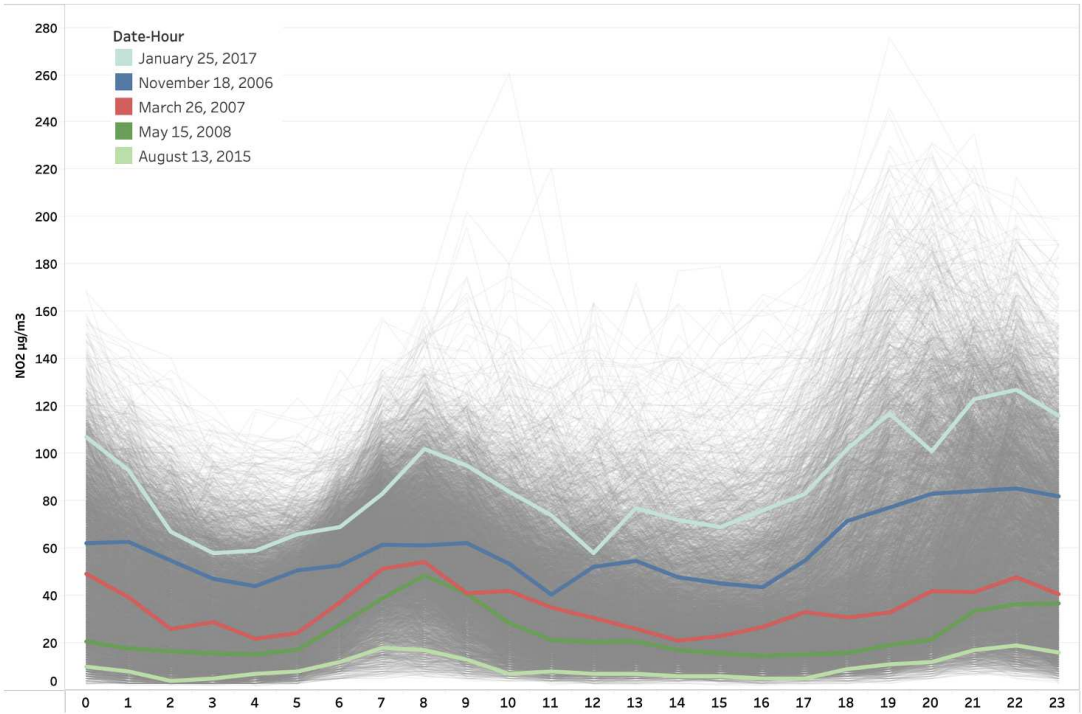


Fig. 5: Daily pollution time series. Coloured lines, ordered bottom-up, represent .05, .25, .5, .75 and .95 EDQ.

Let us consider the 168 weekly time series formed by each of the hours of the day, h, and the days of the week, d, and denote each of these series as $Y^{(h,d)}$, formed by the observations $y_w^{(h,d)}$ (Equation 3).

$$\begin{aligned} Y^{(0,1)} &= y_1,\ y_{169},\ y_{337},\ \dots &&= y_1^{(0,1)},\ y_2^{(0,1)},\ y_3^{(0,1)},\ \dots \\ Y^{(1,1)} &= y_2,\ y_{170},\ y_{338},\ \dots &&= y_1^{(1,1)},\ y_2^{(1,1)},\ y_3^{(1,1)},\ \dots \\ &\vdots \\ Y^{(23,7)} &= y_{168},\ y_{336},\ y_{504},\ \cdots &&= y_1^{(23,7)},\ y_2^{(23,7)},\ y_3^{(23,7)},\ \dots \end{aligned} \tag{3}$$

In Figure A.6 we plot the simple autocorrelogram function (ACF) for each of the 168 subseries. As we can see, the strength of the weekly and annual seasonalities vary depending on the weekday and day hour. Specifically, these subseries are more alike based on the day hour rather than based on the weekday. We can observe three main groups of subseries based on the day hour: series apparently white noise between 21:00 and 8:00; series with a small seasonality between 09:00 and 16:00 and series with a strong seasonality between 17:00 and 20:00.

3.1.3 Data imputation

A 0.4% of the observations in this time period are either missing or faulty. These observations (originated by a failure in the measuring station, lack of power supply, maintenance inspection, missing hourly observations due to switching between summer and winter time, etc) are flagged by the Surveillance System public service provided by Madrid's City Hall and need to be imputed.

These observations are spread out in time through the dataset and rarely group together. In Figure 6 we can observe one of these rare cases being one of the longest imputed sequences in the data. As shown in Section 3.1.2 there is a strong daily seasonality, therefore we will impute these values in the following way: First, we stratify the the original time series based on the day hour, obtaining 24 series $(Y^{(0)}, Y^{(1)}, \dots, Y^{(23)})$ defined as $Y^{(h)} for\ h \in \{0, 1, \dots, 23\}$. Then we will perform the imputation separately within each of the 24 series with a Weighted Moving Average of the previous and following four observations.

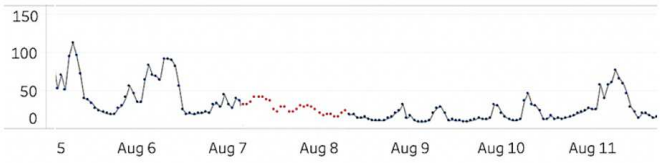


Fig. 6: Data imputation example from 07-08-2009 03:00AM to 08-08-2009 08:00AM. Black dotted line corresponds to real observations and red dots to imputed values.

3.1.4 Dispersion and level association

There is a positive relationship between the dispersion and the level of the NO_2 measurements. In Figure 7 each point represents the hourly observations grouped by year (17 points), week (884 points) and day (6188 points). We use the Median Absolute Deviation (MAD) as the dispersion metric, since it is more robust than the standard deviation and the median as the level metric. Based on the three subplots, we infer that as the pollution level increases so does the dispersion and vice versa. When modelling the pollution time series (see Section 4), denoted as $\boldsymbol{Y}$, we will first transform it with the logarithm so that we reduce the dispersion increase associated with the level. It is worth noting that *0* $\mu g/m^3$ is the minimum value therefore, before taking logarithms, we will add a small constant value c (i.e. c=*4*) to all values in the original time series. For simplicity in the notation, we will denote $\log(Y + c)$ as $\boldsymbol{Y}$ from now on.

3.2 Meteorological effects

The seasonal complexity of this pollutant is not negligible since different chemical reactions, meteorological phenomena and human activities interact altering the level of NO_2 diluted in the Boundary Layer (BL). The BL is the lowest part of the troposphere ranging from hundreds of meters to a few kilometres and directly influenced by the Earth's surface. The BL consists in three major dynamic parts A.1 that follow clear daily and annual pattern; a turbulent Mixed Layer, a less turbulent Residual Layer and a nocturnal Stable Boundary Layer. The Mixed Layer arises with the unstable conditions that begin to manifest at sunrise, when the warming of the Earth's surface is transmitted into the atmosphere. The different air components are well-mixed, giving this layer its name, due to the turbulence driven by vertical convective sources (heat transfer from a warm ground surface and radiative cooling from the top of the cloud). The height and air volume of this layer is defined by the Inversion Layer altitude that acts as a lid to the pollutants, establishing a close relationship between NO_2, along with other air pollutants, and the thickness of this air layer (Aron, 1983). The variation of this height is closely related with the heating of the ground by the sun, following a daily and

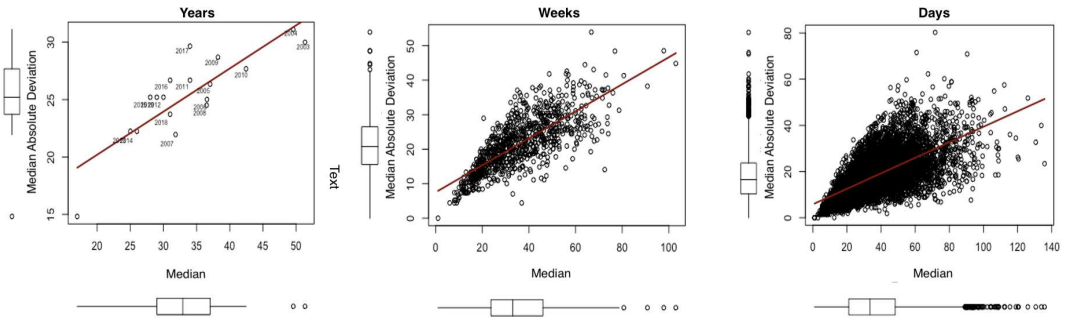


Fig. 7: Relationship between dispersion (vertical axis) and the level of NO_2 (horizontal axis) measured as the Median Absolute Deviation (MAD) and the median, respectively. A linear regression line is shown in red. Hourly observations are grouped per year (left subplot), per week (centre subplot) and per day (right subplot).

annual pattern A.2, reaching its maximum (minimum) height during the central hours of the day (night hours) and during the warm (cold) months of the year while diluting (concentrating) and consequently reducing (increasing) the NO_2 levels (Borge-Garcia and la Paz-Martí, 2017) as described in Section 3.1.2. As the sun goes down, the cessation of the sun's energy input causes the Mixing Layer to collapse with a rapid decrease in thickness. The resulting layer is the Residual Layer, whose initial conditions are the same as those of the preceding Mixing Layer. Below the Residual Layer, the Stable Boundary Layer is formed at ground level and it is characterised by static stability.

Stability does not favour vertical exchange between different layers and surface winds are characteristically weak. This stability, plus when the surface cools more than the air above it, due to radiational cooing of the surface, causes the formation of Nocturnal Inversions in the vicinity of the Earth's surface that typically erodes quickly after sunrise. The formation of these inversions is usually associated with anticyclonic conditions with low wind speeds. It is because of these inversions that there is often fog in the morning, which mixed with pollutants, creates smog in big cities. Turbulent activation either by wind or the warming of the land surface at dawn will lead to the development of the Mixing Layer which, when intensified, will eventually eliminate the Stable Boundary Layer created during the night. The pollutants emitted in the nighttime stable layer disperse relatively little vertically, but they do disperse rapidly in an hourly and horizontal manner. This behaviour is called fanning (Stull, 1988).

Wind is the horizontal movement of air between two points due to the atmospheric pressure differences between them. By transporting the pollutants dissolved in the air, wind produces its horizontal dispersion. When higher wind speeds occur, lower concentrations of NO_2 at ground level are generally measured, and vice versa, as it would result in more diffusion and mixing of contaminants. This relationship is not linear nor a constant effect. For instance, you can have low pollution levels, thanks to low pollution emissions, and no wind at all. Wind speed exhibits an annual and daily seasonality, both with one peak, as shown in the Seasonal Decomposition Plot of Figure 8, and the daily seasonality varies through the year. These patterns match the low pollution period between the first and second daily peaks of pollution through the year shown in Figure 4. Unlike man-made pollution, wind doesn't follow a weekly pattern. It is worth mentioning that unlike other pollutants, like PM10, rainfall has no washing effect over NO_2 as shown by Kwak et al. (2017).

Regarding chemical reactions and meteorological effects, we need to understand how NO_2 is formed. Within urban areas, we can distinguish a direct and indirect origin of this pollutant: NO_2 is primarily formed within engines in combustion processes and is emitted along with other pollutants such as NO into the Mixed Layer. The indirect origin of NO_2 proceeds from the NO emitted by road traffic, which is converted through an oxidation reaction by O_3 or peroxy RO_2 radicals into NO_2 ($NO + O_3;(RO_2) \longrightarrow NO_2 + O_2;(RO)$). Conversely, in the presence of solar radiation a photolysis reaction takes place ($NO_2 + h\nu \longrightarrow NO + O_3$) reducing the levels of NO_2 (Kurtenbach et al., 2012). The photolysis reaction, which follows a daily and annual pattern, takes place during daylight hours therefore its effect is more pronounced during summer. This pattern is aligned with the NO_2 daily and annual seasonality described in Section 3.1.2.

When modelling, we will use the hourly average temperature and wind speed at Madrid Adolfo Suarez Airport, located near the pollution measuring station in the district of Barajas. These data have been provided by the AEMET (*Agencia Estatal de Meteorología*) and ranges from 01-01-2003 at 00:00 until 31-12-2019 at 23:00.

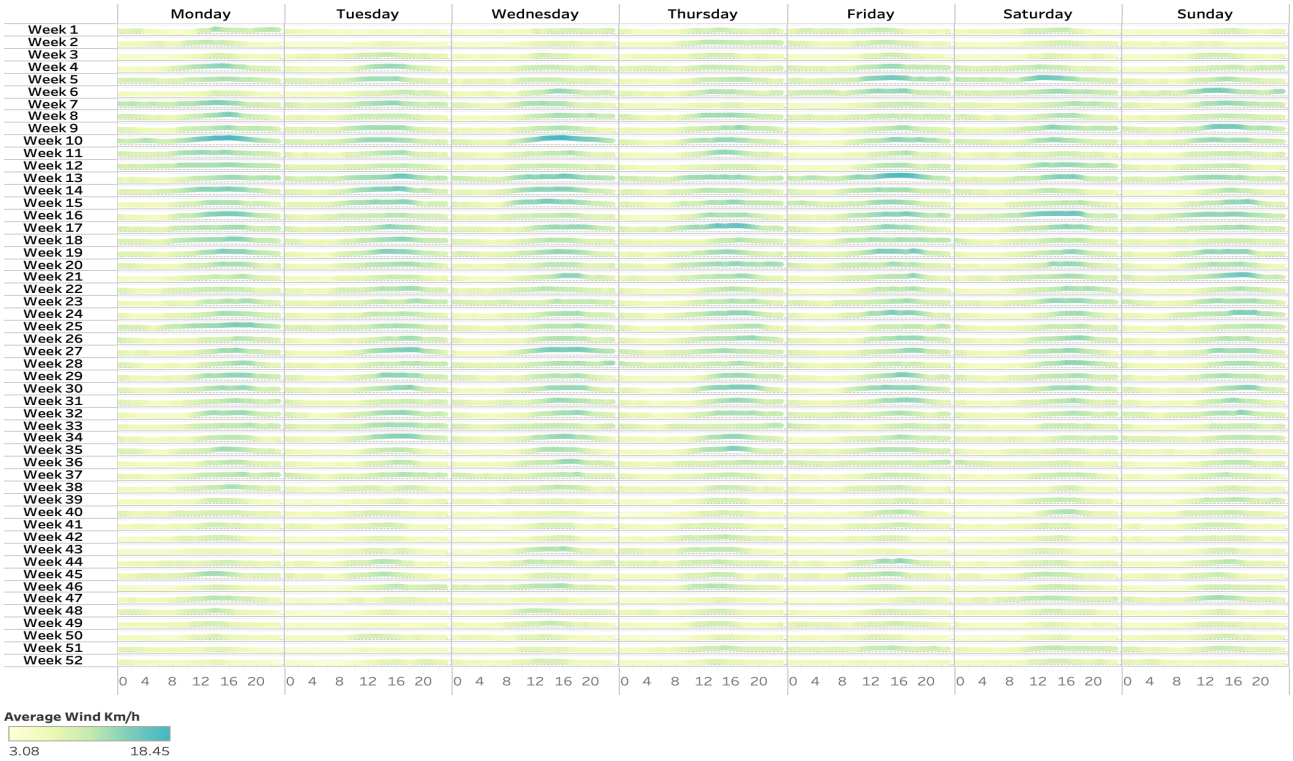


Fig. 8: Seasonal Decomposition Plot of the wind speed (Km/h) levels. One row per week of the year and one column per weekday and each cell includes a line with the average values over the 17 years at each of the 24 hours.

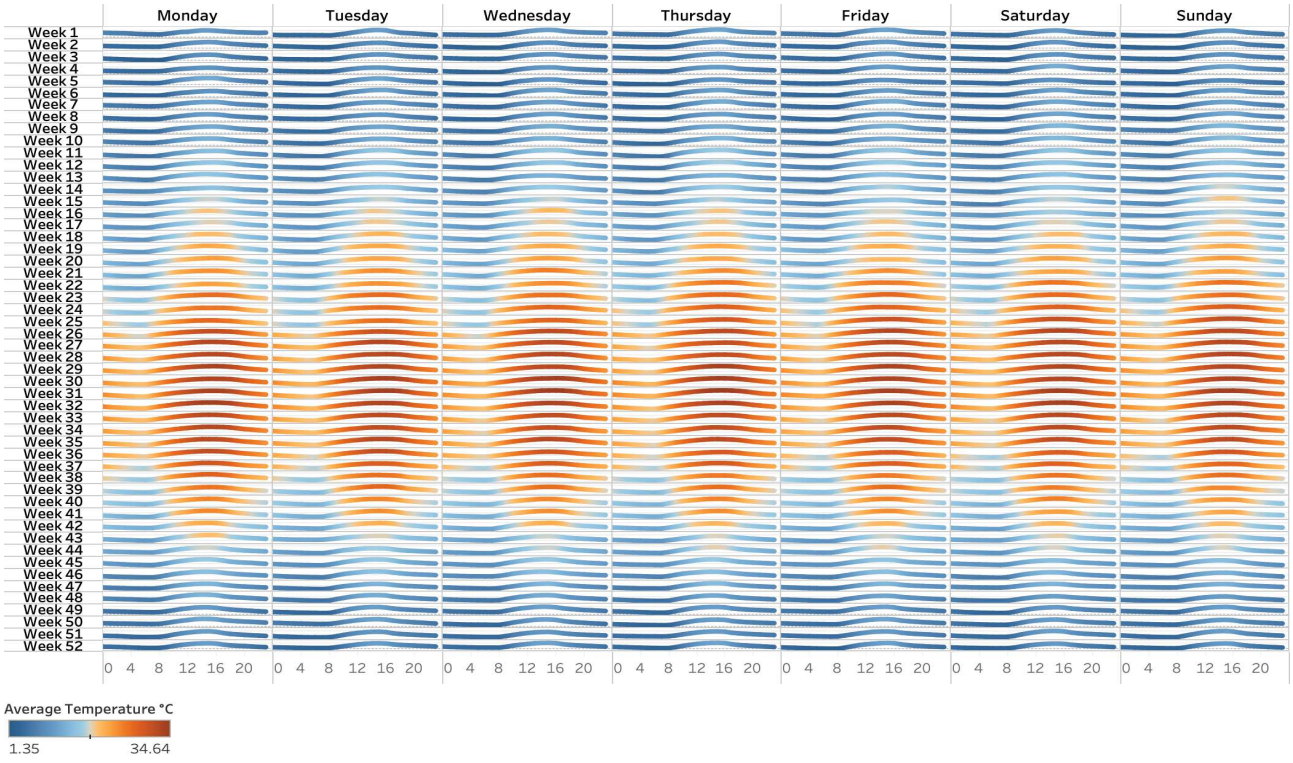


Fig. 9: Seasonal Decomposition Plot of the temperature. One row per week of the year and one column per weekday and each cell the average values of the 24 hours over the 17 years.

3.3 Holidays

We have a calendar with all the national, regional and local holidays that take place in Madrid ranging from 01-01-2003 until 31-12-2019 at 23:00. These data are obtained from Madrid City Council's website and Madrid Autonomous

Community's website. When modelling pollution levels, we will use a dummy variable to incorporate the effect of holidays. On average, NO_2 levels are lower on holidays during the weekdays (Figure 10). Holidays gazetted on Sundays are uncommon since most of them are substituted for the next working day in lieu of the usual date. This is not the case in Madrid for most Christmas and New Year's Day days that fall on Sundays and are not substituted. Therefore, Sunday holidays tend to have higher levels of pollution since they coincide with heavy traffic days.

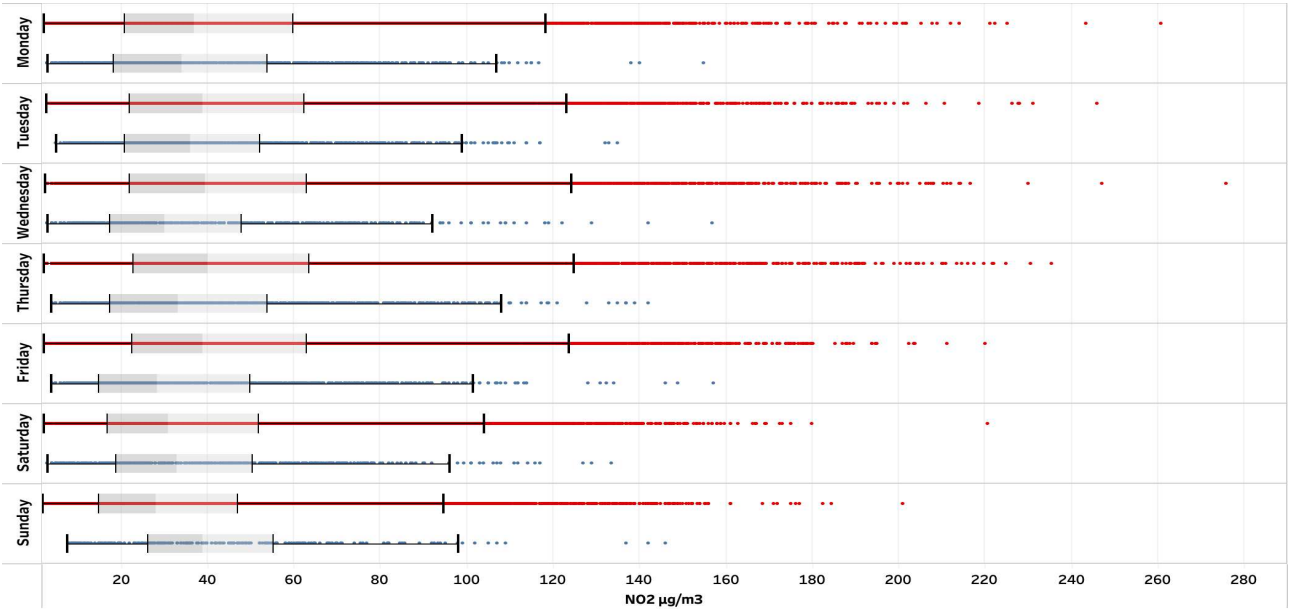


Fig. 10: Pollution levels Boxplots stratified by weekday and holiday (blue if holiday, else red).

4 Proposed Methodology

Single seasonality ARMA model can be extended to obtain models with two or more periodic components to accommodate multiple seasonal cycles as described by Box and Jenkins (1976). We propose a model that takes into account the regular dependency between hourly observations, the three seasonal components (daily, weekly and annual with seasonality 24, 168 and 8736 hours, respectively) combining multiple seasonal ARIMAs and covariables such as the current average wind speed and temperature. It is worth noting that this model is not linear since the seasonal component is composed by time-varying parameters depending on the day hour and weekday.

4.1 M-SARIMAX: Multiple Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors

Our approach consists on two sequential steps: First, we will model the annual and weekly seasonalities using the 168 weekly time series, $Y_w^{(h,d)}$. Then, in the second step, we will model the daily seasonality and the regular dependencies present, using a single time series conformed by the residuals from the first step plus the exogenous covariables effect.

4.1.1 Two steps modelling procedure

First Step. In the first step, we stratify the hourly NO_2 time series into 168 weekly time series $Y_w^{(h,d)}$, formed by each of the hours of the day (h) and the days of the week (d) as described in Section 3.1.2. Each of the 168 weekly subseries is modelled separately with a seasonal ARIMA_{52}. The regular component of this ARIMA_{52} captures the weekly seasonality while the seasonal one captures the annual seasonality.
For any given h and d, we can write each of the 168 ARIMA_{52} as in Equation 4

$$\Phi_{(h,d)}(L)Y_w^{(h,d)} = \Theta_{h,d}(L)a_w^{(h,d)}, \tag{4}$$

where $Y_w^{(h,d)}$ is the pollution subseries, $a_w^{(h,d)}$ is the error term. Polynomials $\Phi_{(h,d)}(L)$ and $\Theta_{h,d}(L)$ are the result of multiplying regular and seasonal polynomials, therefore, their order will be $p + P * 52$ and $q + Q * 52$, respectively. We can re-write Equation 4 as an autoregressive process, as shown in Equation 5

$$\Pi_{(h,d)}(L)Y_w^{(h,d)} = a_w^{(h,d)}, \tag{5}$$

where $\Pi_{(h,d)}(L) = \Theta_{h,d}(L)^{-1}\Phi_{(h,d)}(L)$. We can substitute the notation (h,d), with $h \in \{0,1,\ldots,23\}$ and $d \in \{1,2,\ldots,7\}$, as $s \in \{1,2,\ldots,168\}$ so that $s = 24 * d + h$ obtaining Equation 6.

$$\Pi_{(s)}(L)Y_w^{(s)} = a_w^{(s)}. \tag{6}$$

Note that the lag operator L used in each of these 168 subseries $Y_w^{(s)}$ corresponds to applying a weekly lag operator L^{168} to the original series Y_t. Similarly, a seasonal lag operator L^{52} corresponds to a quasi-annual lag L^{8736} in Y_t.

Second Step. Secondly, we consider the 168 residuals subseries $a_w^{(s)}$ from the first step in their natural order, denoted as a_t, and fit a seasonal ARIMAX$_{24}$ (Equation 7). The seasonal component of this ARIMAX$_{24}$ model will capture the daily seasonality while the regular component will capture the dependency between an observation and the immediately preceding ones. Finally, the exogenous component of the ARIMAX$_{24}$ captures the current average wind speed effect, the average temperature effect and the holiday's effect (encoded as a dummy binary variable).

$$\Phi(L)a_t = \Gamma Z_t + \Theta(L)\epsilon_t, \tag{7}$$

where Z_t are the covariables and the polynomials $\Phi(L)$ and $\Theta(L)$ are the result of multiplying regular and seasonal polynomials, therefore, their order will be $p + P * 24$ and $q + Q * 24$, respectively. As before, we can re-write Equation 7 as an autoregressive process, as shown in Equation 8

$$\Pi(L)a_t + \Omega(L)Z_t = \epsilon_t, \tag{8}$$

where $\Pi(L) = \Theta(L)^{-1}\Phi(L)$ and $\Omega(L) = -\Theta(L)^{-1}\ \Gamma$.

Note that, since a_t is the ordered combination in time of all the residuals from the 168 subseries $a_w^{(h,d)}$, the lag operator L corresponds to applying a lag operator L to the original series. Similarly, a seasonal lag operator L^{24} corresponds to a daily lag L^{24}.

4.1.2 Compact formulation

The two steps procedure in the previous section can be rewritten in a compact formulation. For simplicity in the notation, we will assume that $t = 1$ corresponds to the hour $h = 0$, the day $d = 1$, and the week $w = 1$. To stratify Y_t into the 168 subseries we just need to define the relationship between t, w and s: $t = (w-1)*168+s$, $w = (t \operatorname{div} 168)+1$ and $s = (t \mod 168)+1$. Therefore, the Equation 6 of the first step can be compacted formulated as Equation 9.

$$\Pi_{s(t)}(L^{168})Y_t = a_t. \tag{9}$$

Now, we substitute a_t from Equation 9 into Equation 8 (from the second step) to obtain Equation 10:

$$\Pi(L)\Pi_{s(t)}(L^{168})Y_t + \Omega(L)Z_t = \epsilon_t. \tag{10}$$

Finally, we can re-write the autoregressive process defined in Equation 10 in its most common representation

$$\Phi(L)\Phi_{s(t)}(L^{168})Y_t = \Gamma Z_t + \Theta_{s(t)}(L^{168})\Theta(L)\epsilon_t, \tag{11}$$

where it should be noted that the polynomials $\Phi_{s(t)}$ and $\Theta_{s(t)}$ depend on t. The joint estimation of model defined by Equation 11 is complex while the two-step estimation is feasible and simple.

4.2 Set ARIMA orders

When determining the order of the ARIMA's in both steps we propose two approaches:

- Fixed order. A fixed ARIMA order for all models at each step. The order of the 168 ARIMA models at the first step will be ARIMA(0,1,1)(0,1,1)[52] and at the second step ARIMAX(3,1,1)(1,0,0)[24]. This approach reduces the estimation time since there is no need to adjust and compare multiple orders for each one of the 168 models at the first step nor for the single model at the second step. On the other hand, this approach is more rigid than having different orders for each estimated model.

- Variable order. The ARIMA orders are selected for each time series. The orders of each of the 168 ARIMA models at the first step and the single model at the second step are determined each time we train and re-estimate a model with the function *auto.arima*, available in the R package *forecast* developed by Hyndman and Khandakar (2008). This approach increases significantly the model flexibility at the expense of increase the computation time.

4.3 Training, re-estimation and prediction

The training dataset used follows an expanding window approach, where the upper bound of the window is rolled forward in time as we move one observation ahead to perform the predictions while the lower bound remains fixed at the first observation. Starting with a training dataset of 17472 observations (2 years $\times$ 52 weeks $\times$ 24 hours), we train a model and obtain predictions for the next 72 hours (observation number 17473, 17474... 17544). Once we have predicted these 72 hours we will move one hour forward, increase the training dataset by one observation (now the training dataset contains 17473 observations), and predict the next 72 hours (observation number 17474, 17475... 17545) and so on and so forth. As we move forward in time, model coefficients will be updated every 8736 hours (364 days $\sim$ one year) using the training dataset that would encompass all observations up to that moment. Moreover, the orders of the ARIMA models are reselected when we use the variable order's approach. The forecast horizon was set at 72 hours as longer horizons are unrealistic given that wind speed and temperature predictions will be needed in future applications.

Notice that we will estimate the parameters of our proposed methodology in two steps, where the first step is trained over the 168 stratified subseries and the second step is trained over the residuals, obtained when fitting the 168 models from the first step, sorted in their natural order. When predicting we will also have two steps, that can be done in parallel, and our final forecast will be a combination of the predictions obtained at both steps. At the first step, all observations will be predicted with a 168 hours horizon. This is the shortest prediction we can perform given that the 168 subseries are composed of observations spaced by a one-week interval (168 hours). Consequently, for a given moment of time t we can obtain predictions for the following 168 hours ($\hat{y}_{t+1}$, $\hat{y}_{t+2}$, ..., $\hat{y}_{t+168}$), one prediction from each of the 168 models, that where forecasted at $t-167$, $t-166$, ..., and t, respectively. Therefore we will only have one prediction per day which was predicted 168 hours ago. Since our final goal is to predict the next 72 hours we will just need the first 72 predictions ($\hat{y}_{t+1}$, $\hat{y}_{t+2}$, ..., $\hat{y}_{t+72}$). For the second step we will predict at time t the following 72 residuals ($\hat{a}_{t+1}$, $\hat{a}_{t+2}$, ..., $\hat{a}_{t+72}$) generated in the first step. In order to get the final forecast for the next 72 hours (see Equation 12), we will add to each of the 72 predictions obtained from the second step its corresponding prediction obtained from the first step, then take the natural exponential of this addition and finally subtract the constant value c used in Section 3.1.4.

$$\hat{Y}^{final} = \hat{y}_{t+1}^{final}, \hat{y}_{t+2}^{final}, \ldots, \hat{y}_{t+72}^{final} = e^{\hat{y}_{t+1}+\hat{a}_{t+1}} - c, e^{\hat{y}_{t+2}+\hat{a}_{t+2}} - c, \ldots, e^{\hat{y}_{t+72}+\hat{a}_{t+72}} - c. \tag{12}$$

5 Prophet model

Prophet designed by Meta, former Facebook, frames the forecasting problem as a curve-fitting exercise (Equation 13). Prophet is a modular regression time series model Harvey and Peters (1990) with three main components: trend, seasonality, and holidays;

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \tag{13}$$

where $g(t)$ is the trend function, $s(t)$ accommodates the effect of seasonalities (based on the Fourier series), $h(t)$ represents the effect of holidays or events which occur on irregular schedules over one or more days and ϵ_t is the error term which represents the idiosyncratic changes. Additional regressors can be added to the trend component as exogenous covariables in Prophet.

Unlike ARIMA, Prophet's forecast model has easy-to-interpret parameters, measurements don't need to be regularly spaced and it doesn't need to interpolate missing values. Prophet also claims that can easily accommodate seasonality with multiple periods, unlike (single) seasonal ARIMA models. Our proposed methodology aims to overcome the latter mentioned problem and accommodate multiple seasonalities while relying on ARIMA models.

6 Case study and Results

In order to compare the different methodologies' performance, we will forecast the pollution levels at Barajas station for the next 72 hours and calculate the Root Mean Squared Error (RMSE) for each of these horizons (see Figure 11

and Table 1). We will use our proposed methodology, named as M-SARIMAX, with variable (blue) and fixed order (red) versus the Prophet model (green).

Unlike holidays, which are known in advance, wind speed and temperature covariates must be forecasted prior to using them as input in our models in future applications. Since we lack of hourly wind speed and temperature historical forecasts, we will use the actual values of these measurements as covariables. The RMSE of the predictions obtained when the three covariables are used are represented by dotted lines and can be considered as our best possible scenario when the wind speed and temperature predictions are the most accurate ones. In order to obtain an indisputable performance metric we will work out the RMSE when using the holidays as our sole exogenous covariable (represented by solid lines in Figure 11). We will represent the RMSE obtained when no exogenous covariable was used at all with a plus sign, as in the case of the predictions obtained in the first step of our methodology (left graph of Figure 11).

Prophet is a deterministic model where its predictions don't rely on previous values and, because of this, the RMSE obtained will be the same at each horizon. Likewise, in the first step of our proposed methodology (left graph of Figure 11), we also observe that the RMSE is constant. This is because during the first step all pollution values are actually predicted with a 168 hours horizon as described in Section 4.3. It is surprising how well our methodology, both with variable and with fixed order, outperforms Prophet when using just the holidays covariable. As expected, when Prophet uses the meteorological covariables, it outperforms the first step of our methodology.

When comparing our proposed methodology's final predictions for the next 72 hours against Prophet (right graph of Figure 11) we can notice that our approach, specifically when using variable order, outperforms Prophet's best predictions for the first 22 and 25 horizons when using the holidays covariable and all the three covariables, respectively, while Prophet uses all three covariables. Our two proposals substantially improve Prophet in small horizons up to eight hours. In addition, the predictions with the second step improve those obtained in the first step in horizons from one up to 30. After that, the second step is counterproductive, which may be due to the deterioration of the prediction with ARIMA models to such long horizons.

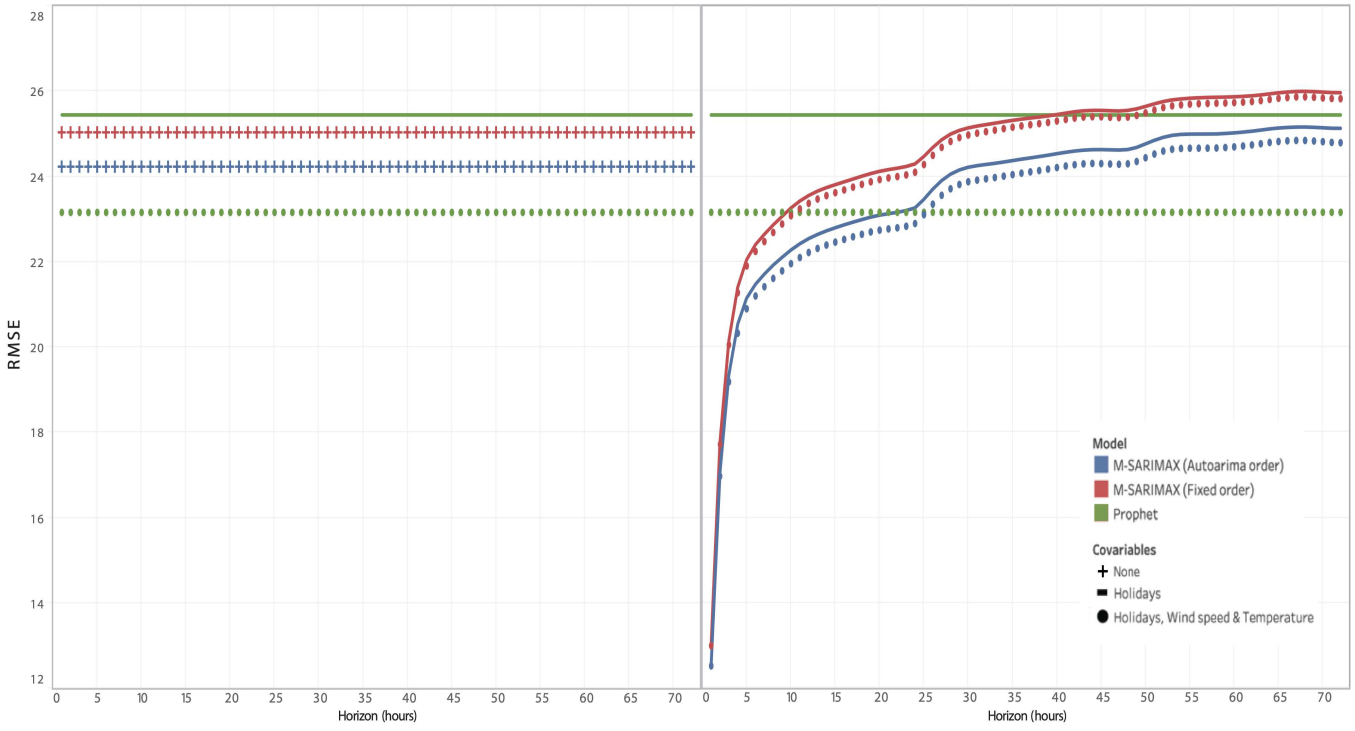


Fig. 11: RMSE for the 72 first horizons for each model and covariates used. (Left) Our methodology first step forecasts vs Prophet forecasts. (Right) Our methodology final forecasts vs Prophet forecasts. Models used: Our methodology with variable order (blue), our methodology with fixed order (red) and Prophet (green). Covariates used: None (represented by addition symbols), the holidays covariable (represented with continuous lines) and the holidays, wind speed and temperature covariables (represented by dots).

	Prophet 1 Cov.	Prophet 3 Cov.	M-SARIMAX Auto. 1 Cov.	M-SARIMAX Auto. 3 Cov.	M-SARIMAX Fixed 1 Cov.	M-SARIMAX Fixed 3 Cov.
1	25.44	23.15	12.54 (24.23)	**12.52 (24.23)**	13.01 (25.04)	12.99 (25.04)
2	25.44	23.15	17.06 (24.23)	**16.97 (24.23)**	17.78 (25.04)	17.71 (25.04)
3	25.44	23.15	19.35 (24.23)	**19.18 (24.23)**	20.15 (25.04)	20.05 (25.04)
4	25.44	23.15	20.53 (24.23)	**20.32 (24.23)**	21.40 (25.04)	21.27 (25.04)
5	25.44	23.15	21.14 (24.23)	**20.89 (24.23)**	22.04 (25.04)	21.90 (25.04)
6	25.44	23.15	21.47 (24.23)	**21.19 (24.23)**	22.40 (25.04)	22.24 (25.04)
12	25.44	23.15	22.54 (24.23)	**22.21 (24.23)**	23.55 (25.04)	23.36 (25.04)
18	25.44	23.15	22.99 (24.23)	**22.64 (24.23)**	24.00 (25.04)	23.81 (25.04)
24	25.44	23.15	23.26 (24.23)	**22.89 (24.23)**	24.29 (25.04)	24.09 (25.04)
30	25.44	**23.15**	24.21 (24.23)	23.87 (24.23)	25.14 (25.04)	24.97 (25.04)
36	25.44	**23.15**	24.40 (24.23)	24.07 (24.23)	25.34 (25.04)	25.18 (25.04)
42	25.44	**23.15**	24.59 (24.23)	24.27 (24.23)	25.51 (25.04)	25.36 (25.04)
48	25.44	**23.15**	24.63 (24.23)	24.29 (24.23)	25.54 (25.04)	25.39 (25.04)
54	25.44	**23.15**	24.98 (24.23)	24.65 (24.23)	25.81 (25.04)	25.67 (25.04)
60	25.44	**23.15**	25.02 (24.23)	24.69 (24.23)	25.86 (25.04)	25.73 (25.04)
66	25.44	**23.15**	25.15 (24.23)	24.83 (24.23)	25.97 (25.04)	25.85 (25.04)
72	25.44	**23.15**	25.12 (24.23)	24.79 (24.23)	25.96 (25.04)	25.82 (25.04)

Table 1: RMSE for multiple horizons for Prophet and our proposed methodology with variable order (M-SARIMAX Auto.) and with fixed order (M-SARIMAX Fixed). The values within parentheses represent the RMSE obtained in the first step of our methodology. When the model uses the holidays covariable (1 Cov.), when the model uses the wind speed, temperature and holidays covariables (3 Cov.). We highlight the lowest RMSE values at each horizon in bold.

7 Conclusions

In this paper, we have introduced a methodology that models multiple seasonalities using ARIMA models and allowing the use of external covariables. The performance of this method was tested forecasting the hourly NO_2 pollution levels for the next 72 hours in Barajas, Madrid. In this application, two hourly measurements where considered as features (wind speed and temperature) and one daily measurement (whether the day was a holiday or not). We used Prophet as a benchmark model and proved that our methodology is very competitive, specially when forecasting the first 24 hours, where the RMSE of the predictions was much smaller than that of Prophet's. Not only that, this methodology outperformed Prophet when meteorological features (wind speed and temperature) weren't used as exogenous covariates. We used Prophet as our benchmark model since it is used to model pollutants time series, can accommodate multiple seasonalities and include covariates (unlike models such as TBATS). We have also demonstrated that our methodology outperforms simple decomposition techniques, where time series are decomposed into different components and model each of them separately, when forecasting from one up to 30 horizons. Like simple decomposition techniques, our methodology also decomposes and models each component separately. However, it goes one step further by reconstructing these stratified components and modelling the remaining patterns in the subsequent step, improving the forecasts obtained up to 30 steps ahead but deteriorating the predictions with more distant horizons.

Acknowledgements

The authors would like to thank to the State Meteorological Agency (AEMET) for providing the meteorological data used in this article. The authors gratefully acknowledge the financial support from the Spanish government through Ministry of Science and Innovation projects PID2019-108311GB-I00 and PID2019-109196GB-I00.

References

E. Agirre-Basurko, G. Ibarra-Berastegi, and I. Madariaga. Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. Environmental Modelling and Software, 21(4):430–446, 2006. ISSN 1364-8152. doi:https://doi.org/10.1016/j.envsoft.2004.07.008.

H. Allende, C. Moraga, and R. Salas. Artificial neural networks in time series forecasting: A comparative analysis. Kybernetika, 38(6):685–707, 2002. ISSN 2589-7918. URL http://eudml.org/doc/33612.

R. Aron. Mixing height—an inconsistent indicator of potential air pollution concentrations. Atmospheric Environment (1967), 17(11):2193–2197, 1983. ISSN 0004-6981. doi:https://doi.org/10.1016/0004-6981(83)90215-9.

A. Baklanov and Y. Zhang. Advances in air quality modeling and forecasting. Global Transitions, 2:261–270, 2020. ISSN 2589-7918. doi:https://doi.org/10.1016/j.glt.2020.11.001.

K. Bandara, C. Bergmeir, and H. Hewamalage. LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. IEEE Transactions on Neural Networks and Learning Systems, 32(4):1586–1599, 2021. ISSN 2162-2388. doi:10.1109/TNNLS.2020.2985720.

R. Borge-Garcia and D. De la Paz-Martí. Estudio de dispersión de contaminantes atmosféricos de la planta de valorización energética de residuos Las Lomas en el parque tecnológico de Valdemingómez. Technical report, Grupo de Investigación de Tecnologías Ambientales y Recursos Industriales de la Universidad Politécnica de Madrid, 2017. URL https://diario.madrid.es/wp-content/uploads/2019/01/Anexo-IV-Informe_estudio_Valdemingomez_V3-firmado_REV.pdf. Trabajo realizado para la Dirección General del Parque Tecnológico de Valdemingómez del Ayuntamiento de Madrid (contrato 133/2017/00212).

G.E. Box and G.M. Jenkins. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, 1976. ISBN 9780816210947.

E. Caro, J. Juan, and J. Cara. Periodically correlated models for short-term electricity load forecasting. Applied Mathematics and Computation, 364:124642, 2020. doi:10.1016/j.amc.2019.124642.

S. Garg and H. Jindal. Evaluation of time series forecasting models for estimation of PM2.5 levels in air. In 2021 6th International Conference for Convergence in Technology (I2CT), pages 1–8, 2021. doi:10.1109/I2CT51068.2021.9418215.

J. González-Enrique, J.J. Ruiz-Aguilar, J.A. Moscoso-López, D. Urda, L. Deka, and I.J. Turias. Artificial Neural Networks, sequence-to-sequence LSTMs, and exogenous variables as analytical tools for NO2 (Air pollution) forecasting: A case study in the bay of Algeciras (Spain). Sensors, 21(5):5–30, 2021. ISSN 1424-8220. doi:https://doi.org/10.3390/s21051770.

A. Harvey and S. Peters. Estimation procedures for structural time series models. Journal of Forecasting, 9:89–108, 1990. ISSN 0277-6693. URL https://doi.org/10.1002/for.3980090203.

R.J. Hyndman and Y. Khandakar. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software, 27(3):1–22, 2008. ISSN 1548-7660. doi:https://doi.org/10.18637/jss.v027.i03.

R. Kurtenbach, J. Kleffmann, A. Niedojadlo, and P. Wiesen. Primary NO2 emissions and their impact on air quality in traffic environments in Germany. Environmental Sciences Europe, 24(6):21, 2012. ISSN 2190-4715. doi:https://doi.org/10.1186/2190-4715-24-21.

H. Kwak, J. Ko, S. Lee, and C. Joh. Identifying the correlation between rainfall, traffic flow performance and air pollution concentration in Seoul using a path analysis. Transportation Research Procedia, 25(3552-3563), 2017. doi:https://doi.org/10.1016/j.trpro.2017.05.288.

C. Lee and C. Ko. Short-term load forecasting using lifting scheme and ARIMA models. Expert Systems with Applications, 38(5):5902–5911, 2011. doi:10.1016/j.eswa.2010.11.033.

A.M. De Livera, R.J. Hyndman, and R.D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association, 106(496):1513–27, 2011. ISSN 0162-1459. doi:https://doi.org/10.1198/jasa.2011.tm09771.

R. De Medrano, V. De Buen Remiro, and J.L. Aznarte. SOCAIRE: Forecasting and monitoring urban air quality in Madrid. Environmental Modelling & Software, 143:105084, 2021. ISSN 1364-8152. doi:https://doi.org/10.1016/j.envsoft.2021.105084.

N. Mohamed, H. Ahmad, and Z. Ismail. Improving short term load forecasting using double seasonal ARIMA model. World Applied Sciences Journal, 15:223–231, 2011. ISSN 1818-4952. URL http://www.idosi.org/wasj/wasj15(2)11/12.pdf.

M. Nelson, T.R. Hill, W. Remus, and M. O'Connor. Time series forecasting using neural networks: should the data be deseasonalized first? Journal of Forecasting, 18(5):359–367, 1999. doi:https://doi.org/10.1002/(SICI)1099-131X(199909)18:5<359::AID-FOR746>3.0.CO;2-P.

F.T.M. Nieuwstadt and H. van Dop. Atmospheric Turbulence and Air Pollution Modelling. Springer Dordrecht, 1982. ISBN 978-94-010-9112-1. doi:https://doi.org/10.1007/978-94-010-9112-1.

D. Peña, R.S. Tsay, and R. Zamar. Empirical dynamic quantiles for visualization of high-dimensional time series. Technometrics, 61(4):429–444, 2019. doi:https://doi.org/10.1080/00401706.2019.1575285.

X. Querol. La calidad del aire en las ciudades. Un reto mundial. Fundación Gas Natural Fenosa, 2018. ISBN 9788409019052. URL http://www.fundacionnaturgy.org/wp-content/uploads/2018/06/calidad-del-aire-reto-mundial.pdf.

R. Ramanathan, R. Engle, C.W.J. Granger, F. Vahid-Araghi, and C. Brace. Short-run forecasts of electricity loads and peaks. International Journal of Forecasting, 13(2):161–174, 1997. ISSN 0169-2070. doi:https://doi.org/10.1016/S0169-2070(97)00015-0.

J. Shen, D. Valagolam, and S. McCalla. Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea. PeerJ, 8:1–18, 2020. ISSN 2167-8359. doi:https://doi.org/10.7717/peerj.9961.

R.B. Stull. An Introduction to Boundary Layer Meteorology. 1383-8601. Kluwer Academic Publishers, 1988. ISBN 9789027727688. doi:https://doi.org/10.1007/978-94-009-3027-8.

J.W. Taylor. Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society, 54(8):799–805, 2003. ISSN 0160-5682. doi:https://doi.org/10.1057/palgrave.jors.2601589.

S.J. Taylor and B. Letham. Forecasting at scale. The American Statistician, page 37–45, 2017. ISSN 0003-1305. doi:https://doi.org/10.1080/00031305.2017.1380080.

P.R. Winters. Forecasting sales by exponentially weighted moving averages. Management Science, 6(3):324–342, 1960. doi:https://doi.org/10.1287/mnsc.6.3.324.

World Health Organization. Air quality guidelines of WHO for particulate matter, ozone, nitrogen dioxide and sulfurdioxide, Global update 2005: Particulate matter, ozone, nitrogen dioxide and sulfur dioxide. International Series of Monographs on Physics. WHO Press, 2006. ISBN 9789289021920. URL https://apps.who.int/iris/handle/10665/107823.

G.P. Zhang and M. Qi. Neural network forecasting for seasonal and trend time series. European Journal of Operational Research, pages 501–514, 2005. ISSN 03772217. doi:https://doi.org/10.1016/j.ejor.2003.08.037.

A Appendix Figures

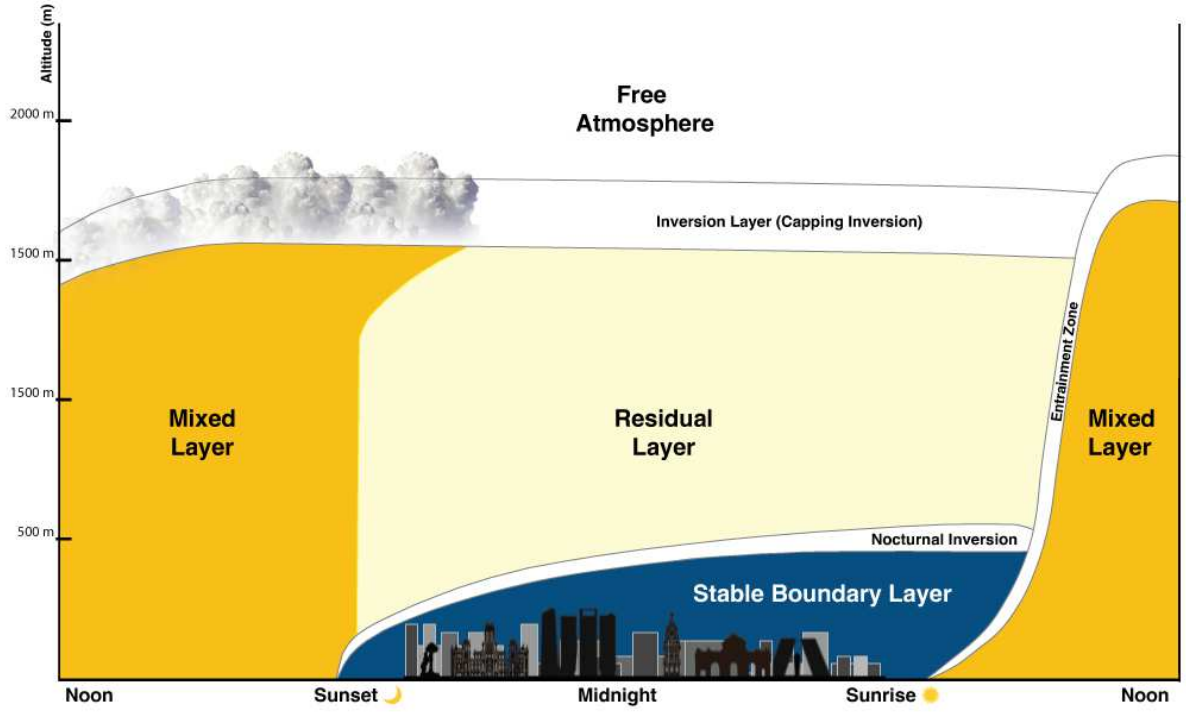


Fig. A.1: Typical evolution of atmospheric Boundary Layer altitude during the day.

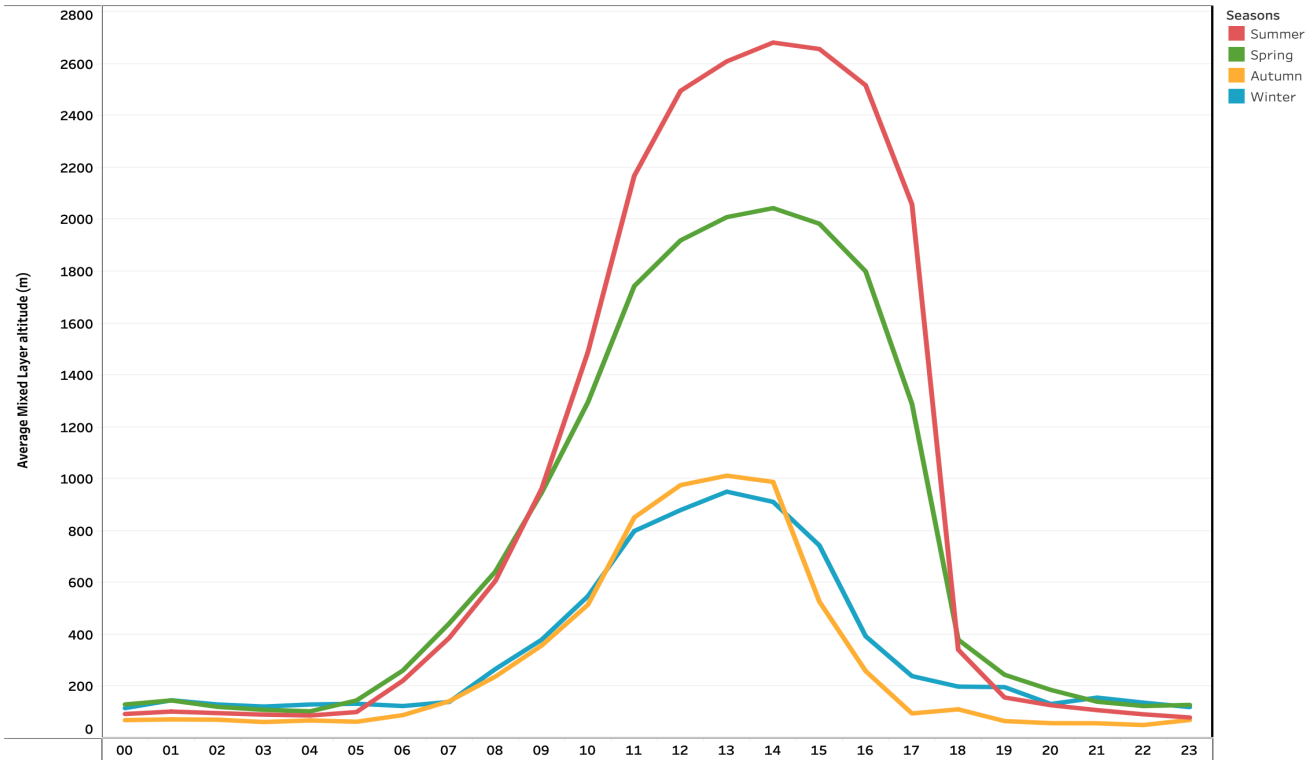
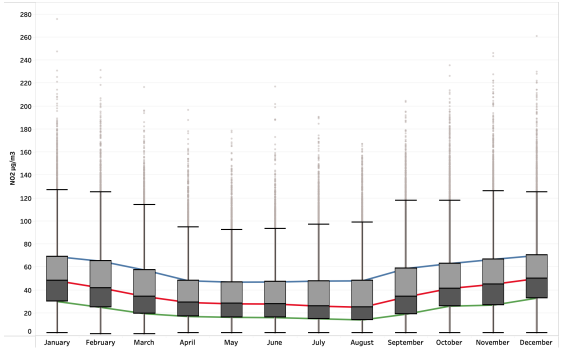
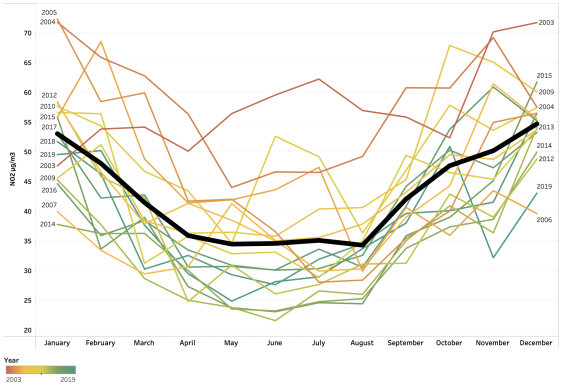


Fig. A.2: Average Mixed Layer altitude in meters at each day hour and stratified by seasons: Summer (red), Spring (green), Autumn (orange) and Winter (blue). Data obtained from Borge-Garcia and la Paz-Martí (2017).

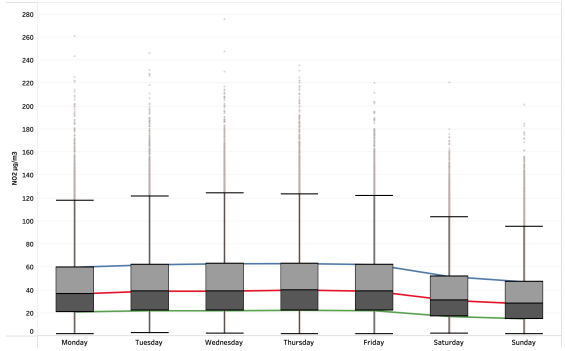


(a) Boxplots: Intra-year distribution of hourly NO_2 $\mu g/m^3$ concentrations by month. **Coloured lines**: .25 (green), .5 (red) and .75 (blue) Timewise Quantiles.

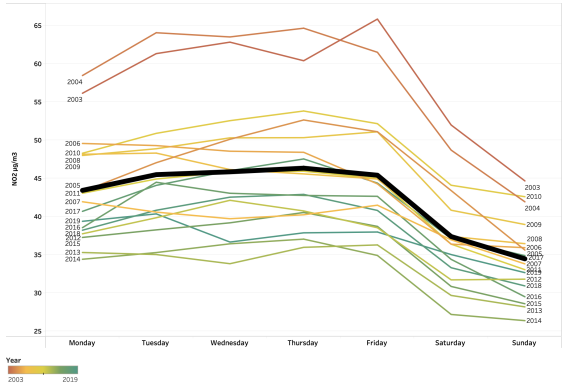


(b) Coloured lines: One line per year with the monthly NO_2 $\mu g/m^3$ average. **Black line**: Historical monthly NO_2 $\mu g/m^3$ average.

Fig. A.3: Annual seasonality.

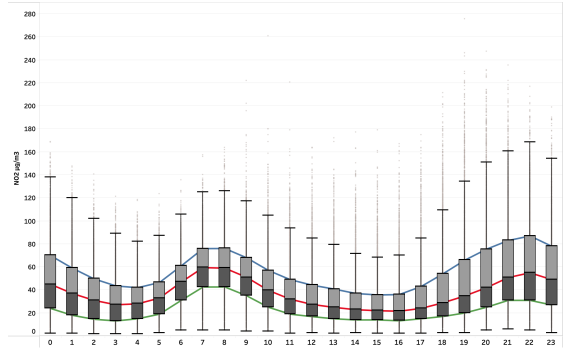


(a) Boxplots: Intra-week distribution of hourly NO_2 $\mu g/m^3$ concentrations by weekday. **Coloured lines**: .25 (green), .5 (red) and .75 (blue) Timewise Quantiles.

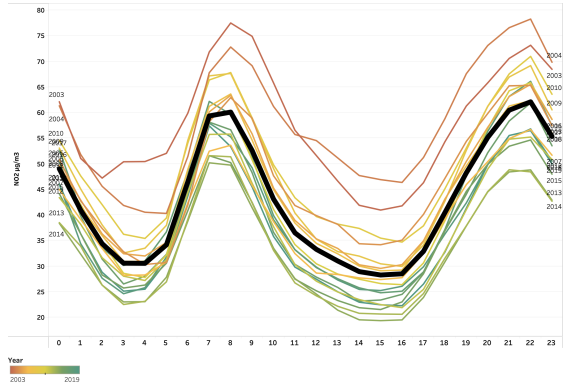


(b) Coloured lines: One line per year with the weekday NO_2 $\mu g/m^3$ average. **Black line**: Historical weekday NO_2 $\mu g/m^3$ average.

Fig. A.4: Weekly seasonality.



(a) Boxplots: Intra-day distribution of hourly NO_2 $\mu g/m^3$ concentrations by day hour. **Coloured lines**: .25 (green), .5 (red) and .75 (blue) Timewise Quantiles.



(b) Coloured lines: One line per year with the day hour NO_2 $\mu g/m^3$ average. **Black line**: Historical day hour NO_2 $\mu g/m^3$ average.

Fig. A.5: Daily seasonality.

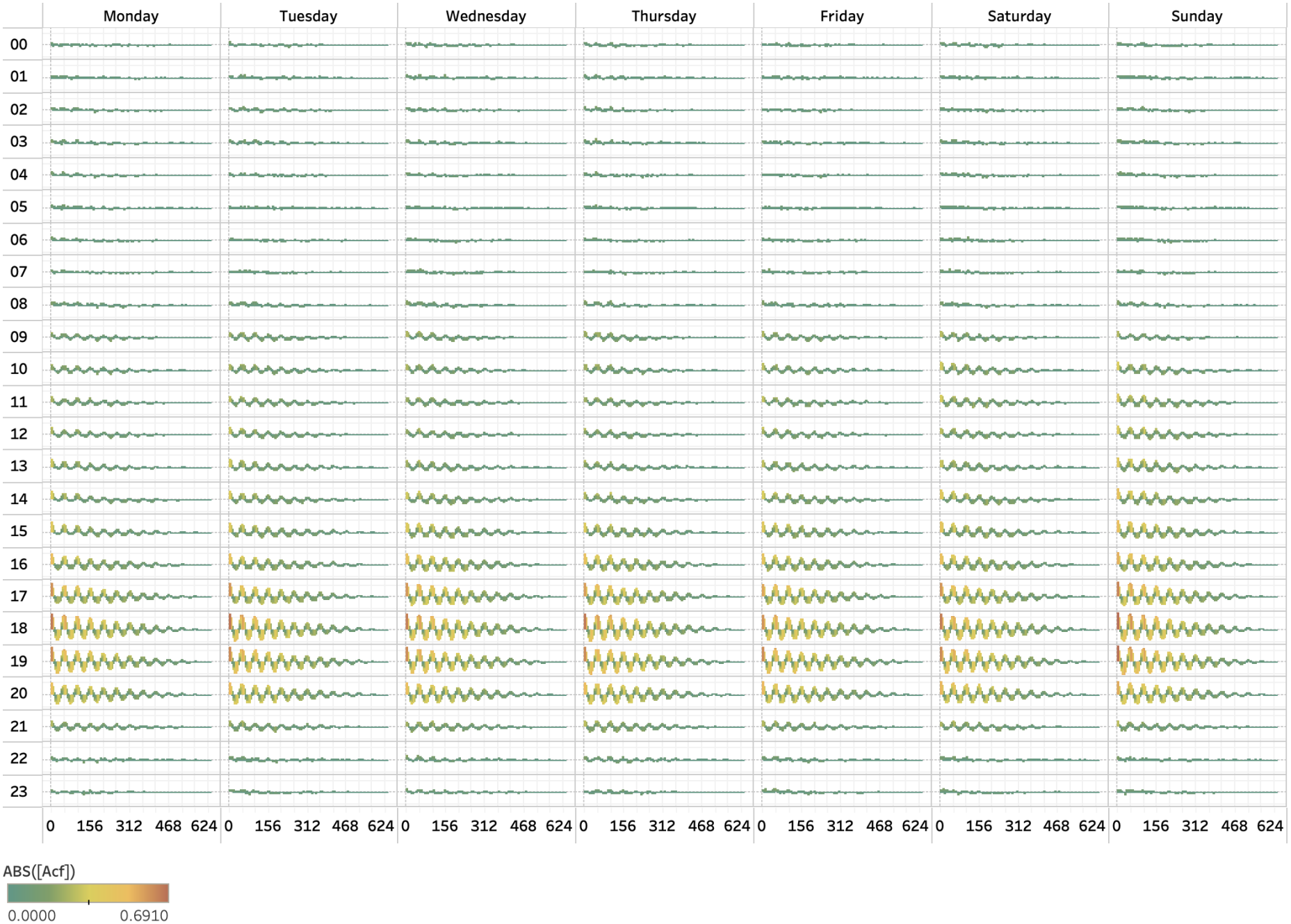


Fig. A.6: One row per hour of a day and one column per weekday. Each cell contains the hourly average NO_2 $\mu g/m^3$ levels. Each cell contains the simple autocorrelogram function (SACF) coloured based on the correlation absolute value. Red being the highest absolute values and green the values closest to 0. SAC values range between -0.6 and 0.76.