Time Series Coursework

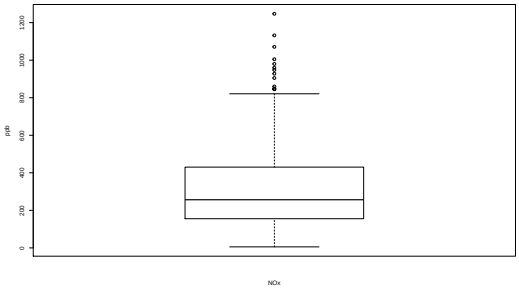
Amaya Syed - ID. 190805496

1. Introduction

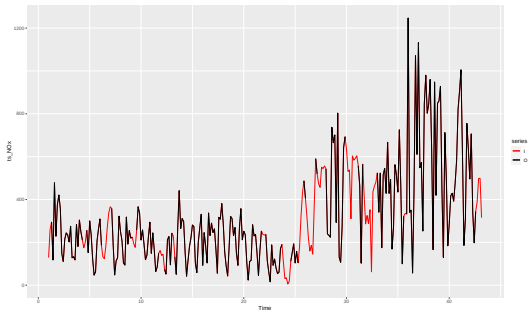
The *AirQuality.csv* dataset is subsetted from an air quality survey containing the responses of gas solid state multisensor device, a certified reference analyser providing ground truth data (GT), as well as temperature and relative humidity sensors [1]. These device were deployed in an Italian city, on a road with heavy traffic, between March 2004 and February 2005 [2]. This subsetted dataset contains the 9 AM hourly GT average for NO_x and NO_2, as well as the temperature, absolute and relative humidity a that time. We will split the data into a training set going from 11/03/04 to 31/12/04/ and a testing set, containing data for the month of January 2005. We will use the training set to build several forecasting models, the accuracy of which we will then verify by comparing the results obtained with our testing set.

2. Exploratory analysis

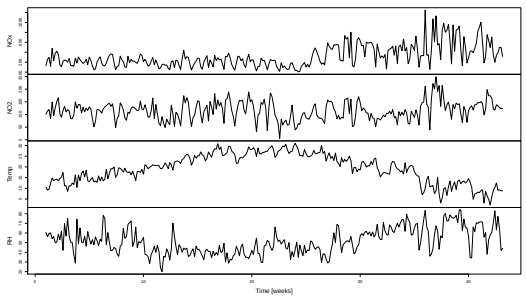
In Figure 1 we see a boxplot of the NO_x training set distribution in (a), the results of the linear interpolation used to input missing data in (b), NO_x, NO_2, Temperature (Temp) and Relative Humidity (RH) in function of time in (c) and the correlation between and distribution of these variables in (d).



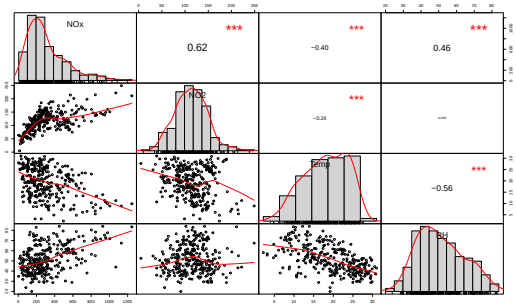
(a) NO_x boxplot, original data on left, modified on right



(b) Linear interpolation in red and orginal data in black



(c) Times Series



(d) correlation and distribution of the variables

Figure 1: Exploratory analysis of the dataset

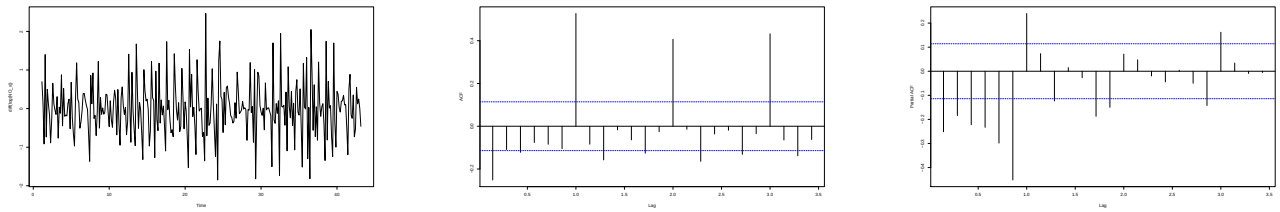
Despite the outlying values (over 1000 ppb) in Fig 1 (a) being extremely high pollution values, we decide to retain these data points without modification. The data was obtained from a certified reference analyser and therefore is likely to be an accurate reading of the air pollution levels at that time. Overall, circa 18 % of values are missing in the training set, corresponding to days the reference analyser might have broken down, as we can see that the missing data tends to cluster in

groups of several days. This complicates inference because move forward methods might obscure any seasonality and linear interpolation accumulates error over time. Possibly a better strategy would be to calculate the $\pm$ two week average of the specific day missing, which has a better chance of respecting seasonality and trend patterns. Nevertheless, because of ease of implementation we performed a linear interpolation to input all missing values, the result of which can be seen in Fig 1 (b). Moreover, after performing a Fourier analysis on the interpolated data, we found the dominant frequency was 7, which corresponds to a weekly "seasonality"pattern whereby there is less air pollution during the weekend when cars are less used and factories tend to be stopped. We input this frequency when creating our times series object in `R`, which were then plotted in (c).

Most notably for NO_x in Fig 1 (c) we see an increase in the overall mean and variance from approx. week 25 onwards, indicating an increase in the overall air pollution. In parallel we see the temperature starts decreasing, until reaching the lowest temperature in December. This dynamic is the result of a natural phenomenon, whereby air pollution increases during colder months because of increased cloud coverage and temperature inversion, therefore acting as a barrier to air dispersal. Variance is also increased as autumn/winter weather patterns oscillate between being overcast, windy, rainy and clear more frequently than spring/summer months. In the correlogram, we verify that Temp is negatively correlated with NO_x, as expected. Additionally, we see relative humidity (RH) is positively correlated with NO_x, which is again logical as higher humidity implies higher cloud covering, hence higher temperatures. Finally, NO_2 is positively correlated with NO_x which is trivial as NO_x is a term which englobes NO and NO_2 and NO_2 derives from NO. This analysis suggests we could leave additional variables out of our analysis, as they are not independent from each other. Overall, we have left Absolute Humidity (AH) readings out of the analysis, as it correlates little to NO_x and highly to temperature, suggesting it is redundant.

To confirm our data is not stationary we perform an Augmented Dickey-Fuller Test, obtaining a p-value of 0.25. We cannot therefore reject the null hypothesis which states the data is non stationary. To examine which might be the optimal transformation parameter for our data, we use the `BoxCox.lambda` function and obtain a value of 0.09. For simplicity, we choose to use a 0 value, corresponding to a log transform of the data.

In Figure 2 we plot the differenced and log transformed NO_x series, as well as its autocorrelation (acf) and partial autocorrelation (pacf) plots.



(a) diff(log NO_x) times series (b) diff(log NO_x) acf (c) diff(log NO_x) pacf

Figure 2: ACF and PACF plots before and after transforming

Visually the data seems stationary, which is confirmed by and ADF of p-value < 0.01. Because both the acf and pacf plots seem to tail off and the data is stationary, a sARIMA model might be a good place to start.

3. Time Series Forecasting

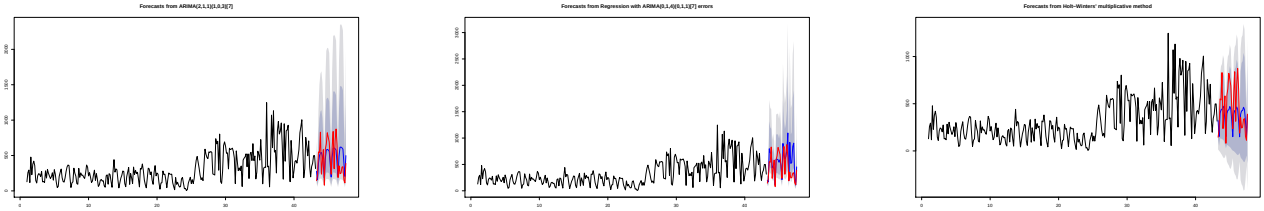
In Table 1 we see the results for several forecasting models. When applicable, we compare their AIC, the RMSE for the fitted model and the RMSE for the forecasted model. The are several types of

models: sARIMA, Dynamic Regression with sARIMA errors, Holt-Winter and finally a autoregressive neural network.

Table 1: We compare different forecasting methods by comparing their AIC and the RMSE for the model and forecast fit.

Forecasting method	AIC	model RMSE	forecast RMSE
sARIMA models			
sARIMA(0,1,3)(2, 0, 0)7	454	9.09	41.11
sARIMA(2,1,1)(1, 0, 3)7	419	8.23	36.42
sARIMA(2, 1, 1)(1, 1, 1)7	409	8.12	36.46
Regression models with ARIMA errors			
sARIMA(1,1,4)(2,0,0)7	227	6.23	41.26
sARIMA(0, 1, 4)(0,1,1)7	192	5.48	42.91
sARIMA(0, 1, 3)(0,1,1)7	196	5.58	46.41
Triple exponential smoothing (Holt-Winter)			
multiplicative method	NA	7.75	39.12
Neural network autoregression			
nnar(18, 1, 10)	NA	NA	39.41

As we can see from the table, the model with the smallest forecasted rmse is a sARIMA(2, 1, 1)(1, 0, 3)7 (cf. fig 3 (a)), whereas the model with the smallest AIC is a dynamic regression model, which uses NO_2, Temp and RH as regressors, with error structure sARIMA(0, 1, 4)(0, 1, 1)7 (cf. fig 3 (b)). This model has forecasted rmse of 42.91. Overall, there is not much variance in the forecasted rmse - which goes from 36.46 to 42.91. We will plot three models in Figure 3 to better understand this result:

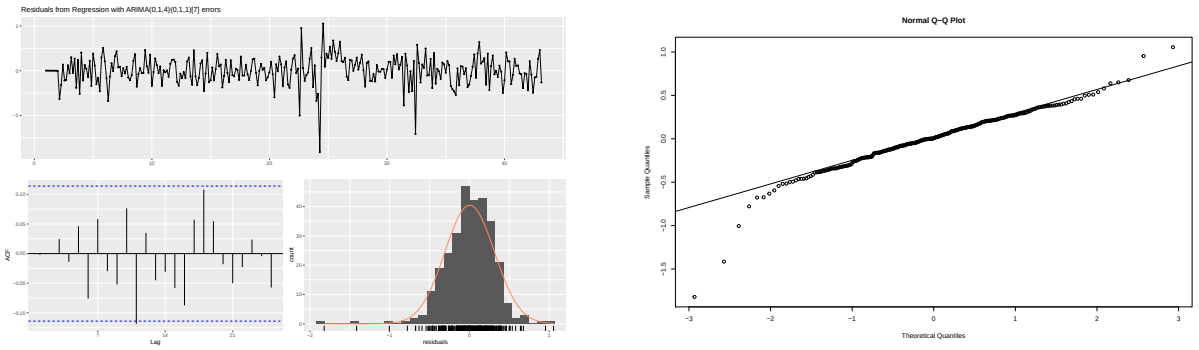


(a) sARIMA) times series (b) Dynamic regression model (c) Holt-Winter

Figure 3: Three different forecasting models

The sARIMA model on average follows the fluctuations in NOx values, without going to extremes. This strategy gives good results, because on average the rmse will yield better results this way. ON the other hand the dynamic regression model seems to forecast the model in more detail, with more extreme values. This might have yielded better results if it were not for the last three data points in the testing set which seem to be much lower than expected given the Temp, RH and day of the week for that month. This might be an instrumental error or it might correspond to a random weather fluctuation not accounted for by the variables considered. The Holt-Winter model yields an rmse result between the last two models, with a similar look to the sARIMA forecat, but is comparatively very easy to fit, with only the seasonal component chosen by hand, and gives good results immediately. Again, the overall fit of the model might be made better by dealing with the atypical datapoints in the testing set.

Finally, we will look at the residuals generated by the sARIMA and dynamic regression models in Figure 4



(a) residual analysis (b) qqplot

Figure 4: Residual analysis for the sARIMA model

The residuals for a sARIMA model should be $\epsilon \sim N(0, \sigma_t^2)$. We see from the plotted residuals that they seem mostly alright, except for a spike around 24, which seems a bit far off from the mean. The ACF plot has one lag between 7-14 which is barely significant, but something to look into. Looking at the distribution of the residuals, one sees it is slightly shifted to the right. Moreover, the extreme values of qqplot are quite far from the norm. This sARIMA model seems adequate, but definitely needs some more work to obtain better overall residuals. Potentially working on the origin of the outlier values in the dataset (instrumental error, weather patterns) might help to do so. Exploring further data transforms might also be valuable.

Referencias

[1] Saverio De Vito *http://archive.ics.uci.edu/ml/datasets/Air+Quality* ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development, (saverio.devito '@' enea.it).

[2] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia *On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators* B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005.

[3] Hyndman, R.J., Athanasopoulos, G. *Forecasting: principles and practice* 2018, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 28/04/2020