JOCLAD 2024

XXXI Jornadas de Classificação e Análise de Dados

ESTG - Politécnico de Leiria

18 a 20 abril

Air quality prediction for work environments using ARDL models

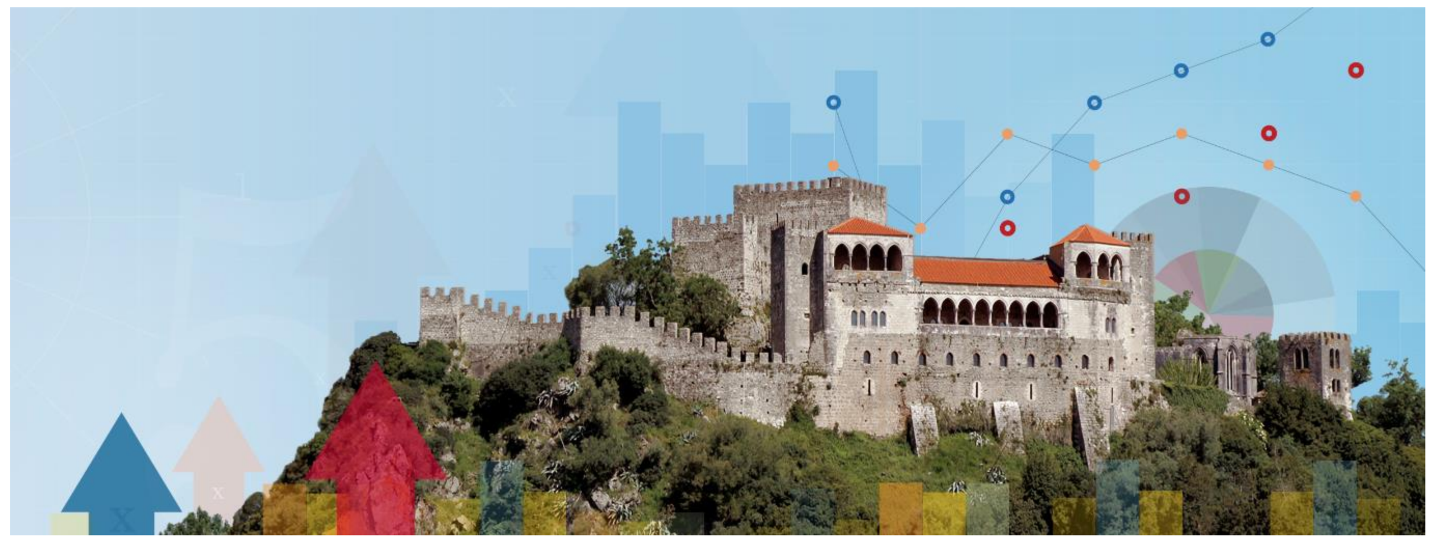
Fernando Batista[1,2*] e Jorge Siopa[1]

1: School of Technology and Management, Polytechnic of Leiria, Portugal

2: CDRSP - Centre for Rapid and Sustainable Product Development, Polytechnic of Leiria

e-mail: fernando.batista@ipleiria.pt, jorge.siopa@ipleiria.pt,

*corresponding author



Introduction

Air quality in urban areas affects the health and well-being of the population. Knowing the importance of this problem, it is necessary in the first phase to monitor all the atmospheric parameters that can affect air quality and in the second phase, to create predictive models. With these predictive models it is possible to analyze the importance of the atmospheric parameters that influence the gases that degrade air quality. Using a Python package for an AutoRegressive Distributed Lag (ARDL) analysis of time series, it was possible to characterize this importance.

Aim

In an attempt to assess well-being at Campus 2 of the School of Technology and Management of the Polytechnic of Leiria, air quality monitoring data obtained by the on-site Mobile Air Quality Monitoring Unit was collected and a study was carried out using this data to characterise the importance of certain atmospheric parameters using a predictive time series analysis tool, ARDL.

Method

This ARDL method is widely used in economics to estimate forecasts. The time series in the model are divided into endogenous and exogenous variables. The parameters that characterize the atmospheric conditions are always assumed to be exogenous variables. The gases in the atmosphere that influence air quality are assumed to be endogenous variables or exogenous variables.

The time series obtained for the gases were ozone O_3, carbon monoxide CO, hydrocarbons (NO, NO_X and NO_2) and particles $PM10$ and $PM2.5$. For the atmospheric conditions we have precipitation (mmH_2O), global radiation ($Rglobal$), temperature ($Temp$), humidity ($Humid$), atmospheric pressure ($Press$), wind speed ($Vwind$) and direction ($Dwind$). To try to characterize car traffic, a time series of rush hour was estimated ($Rhour$). This time series consisting of "0" and "1" was suggested in the work done in this field. In order to draw more robust conclusions with time series, we need a lot of data from many days. Although in this study we only have data from 17 days, it was possible to reach some conclusions.

Various statistical tests were carried out on the time series to assess their stationarity, such as ADF and KPSS. The relationship importance between the two types of variables was also assessed.

The ARDL method uses previous data on the variables and manages to capture the seasonal effect of the series. Equation shows the various components of the iterative process characterized by lags.

$$Y_t = \delta + \sum_{i=1}^{P-1} \gamma_i S_{[(mod(t,P)+1)=i]} + \sum_{p=1}^{A} \phi_p Y_{t-p} + \sum_{k=1}^{M} \sum_{j=0}^{Q_k} \beta_{k,j} X_{k,t-j} + \epsilon_t$$

where δ is constant, $\gamma_i S_{[(mod(t,P)+1=i)]}$ capture seasonal shifts, P is the period of the seasonality, A is the lag length of the endogenous variable, M is the number of exogenous variables X_k, Q_k is included the lag length of X_k and ϵ_t is a white noise shock.

Statistical metrics such as the "adjusted square error", $Adj.R^2$, the "root-mean-square error", $RMSE$ and stationarity methods were used to reach conclusions.

Three types of ARDL (3 models) were carried out, the first with atmospheric conditions as exogenous variables, the second with atmospheric gases as exogenous variables and the third was a combination of atmospheric gases and the atmospheric conditions identified in the first model. The three models performed well on the training data, but when the RMSE and the stationarity of the predictive residuals were evaluated, model 2, which only has gas variables, had the best fit and was the only one to have stationary predictive residuals.

Tools

IDE: PyCharm, Python package: Numpy, Pandas, Sklearn, Statsmodels and Matplotlib.

Generate descriptive statistics

	Atmospheric gases								Atmospheric conditions						
	O3	CO	NO	NOX	NO2	MP10	MP25	Rhour	mmH2O	Press	Rglobal	Temp	Humid	Vwind	Dwind
count	1632	1632	1632	1632	1632	1632	1632	1632	1632	1632	1632	1632	1632	1632	1632
mean	18.23	0.33	14.52	32.42	18.01	47.28	35.43	0.11	0.00	1015.3	185.52	12.18	71.89	0.94	0.93
std	16.67	0.26	30.63	39.07	12.99	21.60	16.33	0.31	0.03	4.23	269.25	6.34	23.31	0.82	0.07
min	0.65	0.00	0.00	0.23	0.54	9.32	6.23	0.00	0.00	1007.0	0.45	1.96	19.71	0.08	0.20
50%	9.66	0.29	4.55	21.64	14.70	46.06	36.07	0.00	0.00	1015.2	0.81	11.18	80.75	0.62	0.94
max	51.32	1.48	384.35	433.52	79.13	94.21	75.69	1.00	0.62	1025.7	807.47	25.17	99.10	5.60	1.00

Three models of ARDL

ARDL1	Endogenous variables	Exogenous variables
ARDL2	Endogenous variables & Exogenous variables	
ARDL3	Endogenous variables & Exogenous variables	Exogenous variables

February																March
Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	1
TRAIN											TEST					

Stationary test

	O3	CO	NO	NOX	NO2	MP10	MP25	HPonta	mmH20	Press	Rglobal	Temp	Humid	Vwind	Dwind
ADF(c)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
KPSS(c)	Yes	Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes
ADF(ct)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
KPSS(ct)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes

Correlation

	O3	CO	NO	NOX	NO2	MP10	MP25	Rhour	Rglobal	Temp	Humid	Vwind	Dwind
O3	1.000	-0.722	-0.350	-0.435	-0.481	-0.688	-0.717	-0.048	0.662	0.857	-0.787	0.617	0.004
CO		1.000	0.408	0.530	0.629	0.542	0.529	0.068	-0.558	-0.612	0.621	-0.506	-0.034
NO			1.000	0.958	0.521	0.113	0.131	0.269	-0.088	-0.140	0.109	-0.188	-0.003
NOX				1.000	0.743	0.114	0.119	0.367	-0.215	-0.187	0.164	-0.264	-0.024
NO2					1.000	0.073	0.046	0.467	-0.440	-0.232	0.237	-0.349	-0.063
MP10						1.000	0.945	-0.197	-0.333	-0.651	0.605	-0.385	0.065
MP25							1.000	-0.187	-0.352	-0.691	0.641	-0.403	0.066

Cointegration test

	Exogenous variables
O3	True
CO	True
NO	True
NOX	True
NO2	True
MP10	True
MP25	True

Grangers Causation test - p-value < 0.05

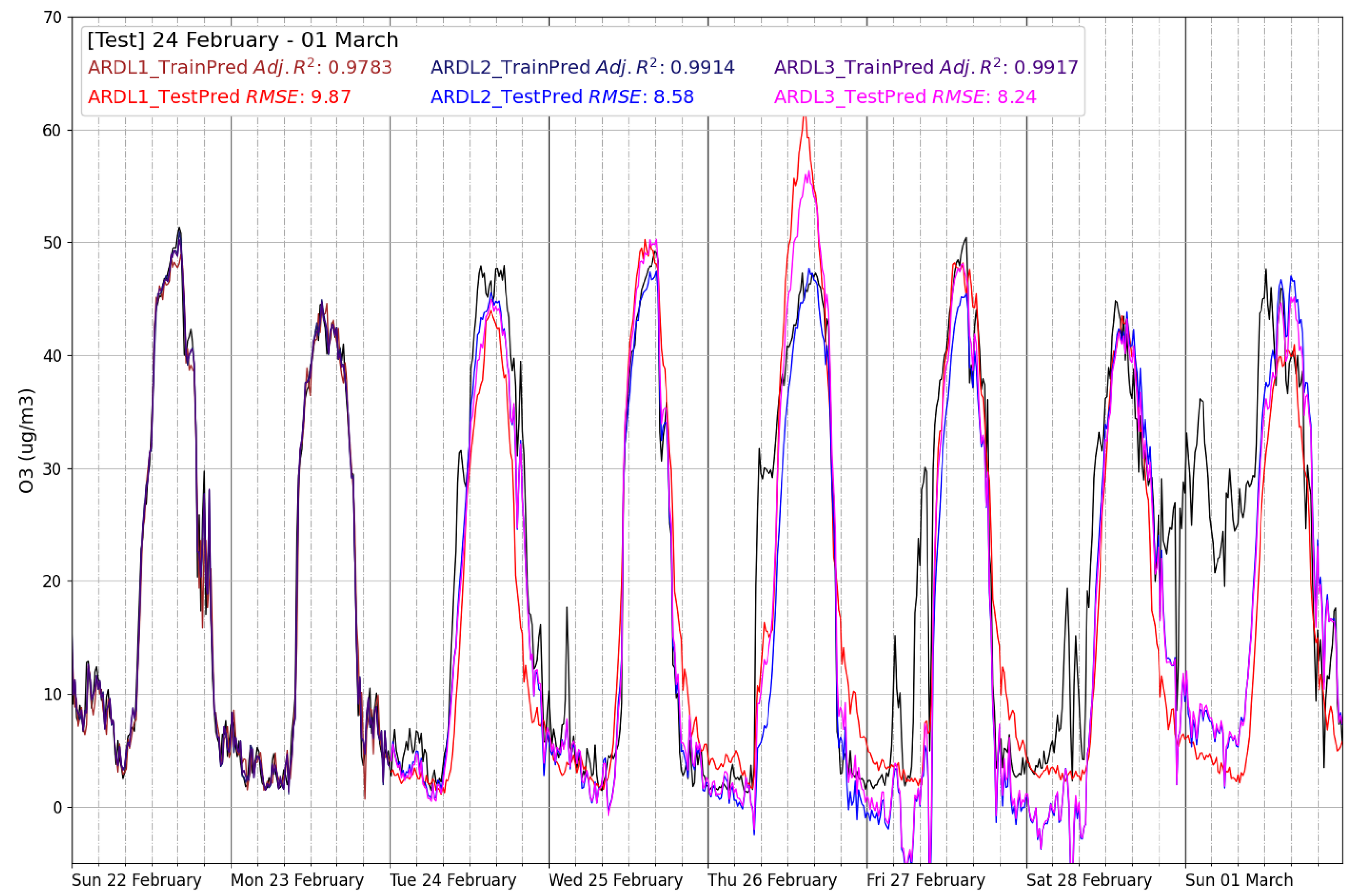
	O3	CO	NO	NOX	NO2	MP10	MP25	Rhour	Rglobal	Temp	Humid	Vwind	Dwind
O3	1	0	0.001	0.014	0	0	0	0.262	0	0	0	0	0
CO	0	1	0.003	0.509	0	0	0	0	0	0	0	0	0.005
NO	0.001	0	1	0.382	0.365	0.702	0.690	0.069	0.001	0	0.001	0	0.087
NOX	0.001	0	0.006	1	0.007	0.284	0.300	0.001	0	0	0	0	0.028
NO2	0	0	0.003	0.003	1	0	0	0	0	0	0	0	0.001
MP10	0	0	0	0	0	1	0	0.003	0	0	0	0	0.158
MP25	0	0	0	0	0	0	1	0	0	0	0	0	0.039

ARDL1	Lags		Q1	Q2	Q3	Q4	Q5	Q6
	P	A	Rglobal	Temp	Humid	Vwind	Dwind	Rhour
O3	96	1				3		
CO	96	3			1	0		0
NO	96	1				0		
NOX	96	1				0		
NO2	96	1		2		0		0
MP10	96	2						
MP25	96	4						

NOX ≈ NO + NO2

ARDL2	Lags		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
	P	A	O3	CO	NO	NOX	NO2	MP10	MP25	Rhour
O3	96	1	---		3	1				
CO	96	3		---	2		1		1	
NO	96	2	1	2	---					
NOX	96	3	1	2		---	3			
NO2	96	3	1	2	3		---			
MP10	96	2		1				---		
MP25	96	3		0			2		---	

ARDL3	Lags		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
	P	A	O3	CO	NO	NOX	NO2	MP10	MP25	Rhour	Rglobal	Temp	Humid	Vwind	Dwind
O3	96	1	---		3	1								2	
CO	96	3		---	1		2		1				0		
NO	96	2	1	2	---										
NOX	96	3	1	2		---	3								
NO2	96	3	1	2	3		---		0	0		0			
MP10	96	2		1				---							
MP25	96	3		0			2		---						



Resid - Durbin-Watson test

	ARDL1	ARDL2	ARDL3
O3	2.07	2.04	2.08
CO	2.01	2.02	2.03
NO	2.01	2.00	2.00
NOX	2.00	2.06	2.06
NO2	2.05	2.08	2.06
MP10	2.00	2.01	2.01
MP25	1.97	1.97	1.97

	TRAIN - Adj.R2			TEST - RMSE		
	ARDL1	ARDL2	ARDL3	ARDL1	ARDL2	ARDL3
O3	0.9783	0.9914	0.9917	9.87	8.58	8.24
CO	0.8774	0.9054	0.9065	0.19	0.16	0.14
NO	0.7461	0.8150	0.8150	25.41	18.77	18.77
NOX	0.7686	0.9230	0.9230	32.98	17.00	17.00
NO2	0.8949	0.9745	0.9749	8.00	7.74	5.80
MP10	0.9897	0.9902	0.9902	13.91	10.32	10.32
MP25	0.9931	0.9935	0.9935	10.82	6.06	6.06

TEST Error - Stationarity test

	ARDL1		ARDL2		ARDL3	
	ADF(c)	KPSS(c)	ADF(c)	KPSS(c)	ADF(c)	KPSS(c)
O3	Yes	No	Yes	Yes	Yes	No
CO	Yes	No	Yes	Yes	Yes	No
NO	Yes	No	Yes	Yes	Yes	Yes
NOX	Yes	No	Yes	Yes	Yes	Yes
NO2	Yes	No	Yes	Yes	Yes	No
MP10	No	Yes	Yes	Yes	Yes	Yes
MP25	Yes	Yes	Yes	Yes	Yes	Yes

ARDL2

Conclusion

From the collected data we can infer that there is a stronger link between pollutant gases in the atmosphere than with meteorological data, for short time forecasting.