

# Using ARIMA models for cleaning the data and assessing significant correlation on Hospital admissions and in-hospital deaths because of COVID 19 because and environmental temperature

**Manuel Esteban Arias**

## Paper Review

Linking a breath-system disease with the climatic conditions in [Mejdoubi, Kyndt & Djennaoui \(2020\)](#) was clever. The disposal of government's reliable data and the occurrence of the lockdown as a natural “ceteris paribus” was convenient for research purposes. Nevertheless, the authors recognises problems as the limited period of two months during the first wave of the Lockdown in France. This makes it difficult to generalize the findings to other regions or time periods. This short period is added to the lack of control group in this study which makes it difficult to rule out alternative explanations for the observed correlations.

In terms of variables, the study only considers temperature and not other weather variables such as humidity or air pollution which could also have an impact on COVID-19 spreading. The amount of unobserved variables may be creating noise in the correlation identifications. While this study provides interesting insights into the potential relationship between ambient temperature and COVID-19 severity indicators, further research may establish causality and rule out alternative explanations.

However, maybe the more problematic is the negligence of the way a virus works: it spreads contagiously, as a network of agents that transmit the infection. While the temperature and humidity can be a catalyst of the disease spreading, the pattern of the ICU and deaths will respond mainly to the own dynamic of the contamination wave. As so, the pattern followed in the pick of the wage on a warm day will not be the same as the one in the valley. Accordingly, the correlation may be not capturing how the progress of the pandemic depends on the stock of infected agents in the previous period.. This accumulative process cannot be captured by simple correlations, it would require isolating the autoregressive process and checking the correlations on the unexplained residuals of the model.

## Method

This document will focus on correcting the flaws of the paper, rather than in improving it. While most of the problems outlined above require the new data that are beyond the scope of the paper, the accumulative dynamics of contagions can be modeled. This is why a two-part methodology is proposed:

In the first, it is proposed to use the Box-Jenkins methodology to model the autocorrelation process of the target variables (deaths and admissions). This methodology consists of the following steps: First it starts with a verification of the stationary that can be done visually by an

AutoCorrelationFunction graph or through the Dickey-Fuller test. This test has a null hypothesis that the series follows a random walk and as such, a P-value lower than 5% permits to reject the null hypothesis and affirm that the series is stationary. If  $H_a$  is accepted, it will differentiate the series until we get a stationary result. Second, the AutoCorrelation Function and the Partial AutoCorrelation Function on the stationary series permits to identify the order of autoregression and of the Moving average depending on the bars that escape to the bandwidths. Fourth, having this information, we can estimate the ARIMA candidate models and compare their performance with the Akaike Information Criteria. Since the AIC represents the information “lost” in the model estimation, the best model is often the one with the lowest AIC. Finally, it is necessary to verify if the chosen model’s residuals are a white noise with a Ljung-Box test. The test assesses if the variable is iid, so a P Value lower than 5% permits rejection of the null hypothesis so the variable may not be white noise.

In the second, I depart from the assumption that the so-called "random part" of the ARIMA model estimation is a combination of a stochastic process and the omitted variables. Accepting this, it would be correct to expect a significant correlation of the residuals of the ARIMA with the climatic variables. This is why it is proposed to incorporate the authors’s proposed lags (8 days for admissions and 15 for deaths) to the an ARIMAX model for assessing their significance as well as to run an Cross Correlation test for identifying higher correlations of the residuals with the climatic variable lags for proposing new conclusions.

## Results

Following a simplified version of the box Jenkins methodology, as reviewed in [Lutkepohl \(2004\)](#). The visualization of the series for admissions seems to suggest a decreasing trend, some convergent to 0 movement that would not be stationary. This would however, when the Dickey-Fuller Test shows results in a P-value of 0,04021 (see the annex) lower than 0,5% and as so we can reject the null hypothesis of non stationarity and affirm that the series is stationary. Checked that it is not necessary to differentiate and the integration order will be 0. Verified that, the exponential decrease of the ACF and the sudden rise in the PACF after lag 3 graphs permits us to think that an optimal AR order would be of 3, and MA would be of 0 ([Nau, 2020](#)). As so 6 models with combinations of 2, 3 and 4 for p, 0 for d, and 0 and 1 for q were tested (see the annex) confirming that the model does requires moving averages, nevertheless, nor the lag 2, nor the 3rd one were significative so at the end a model ARIMA (1,0,0) with an AIC of 566.81 was selected. When the Box-Ljung test was applied (see the annex) the result presente a P value of 0,023 suggesting that the residuals are not white noise. Nevertheless this exercise was repeated with other mother combinations and no value above 5% was found so the select model was kept.

Following a similar methodology to the admissions study, a study was conducted to analyze the correlation between COVID-19 deaths and temperature (max, min). The initial analysis revealed that the series of deaths did not exhibit stationarity, as indicated by the Augmented Dickey-Fuller test with a p-value of 0.3. Therefore, differencing the series once was necessary to achieve stationarity, resulting in a new p-value below the significance threshold of 0.05 (refer to the annex for detailed results).

After ensuring stationarity, the ARIMA model was fitted to the differenced series of deaths. The optimal ARIMA order (4, 1, 1) was determined using the lowest AIC value of 503 (refer to annex) among tested models.

The residuals from the ARIMA (4, 1, 1) model exhibited departures from white noise, as confirmed by the Box-Ljung test with a significantly lower p-value than the 5% significance level.

Once we extract the self explanatory component of the disease spreading, and the associated number of hospital admissions and deaths, we can use the residuals for testing the correlations proposed by the authors and for proposing better lags if necessary.

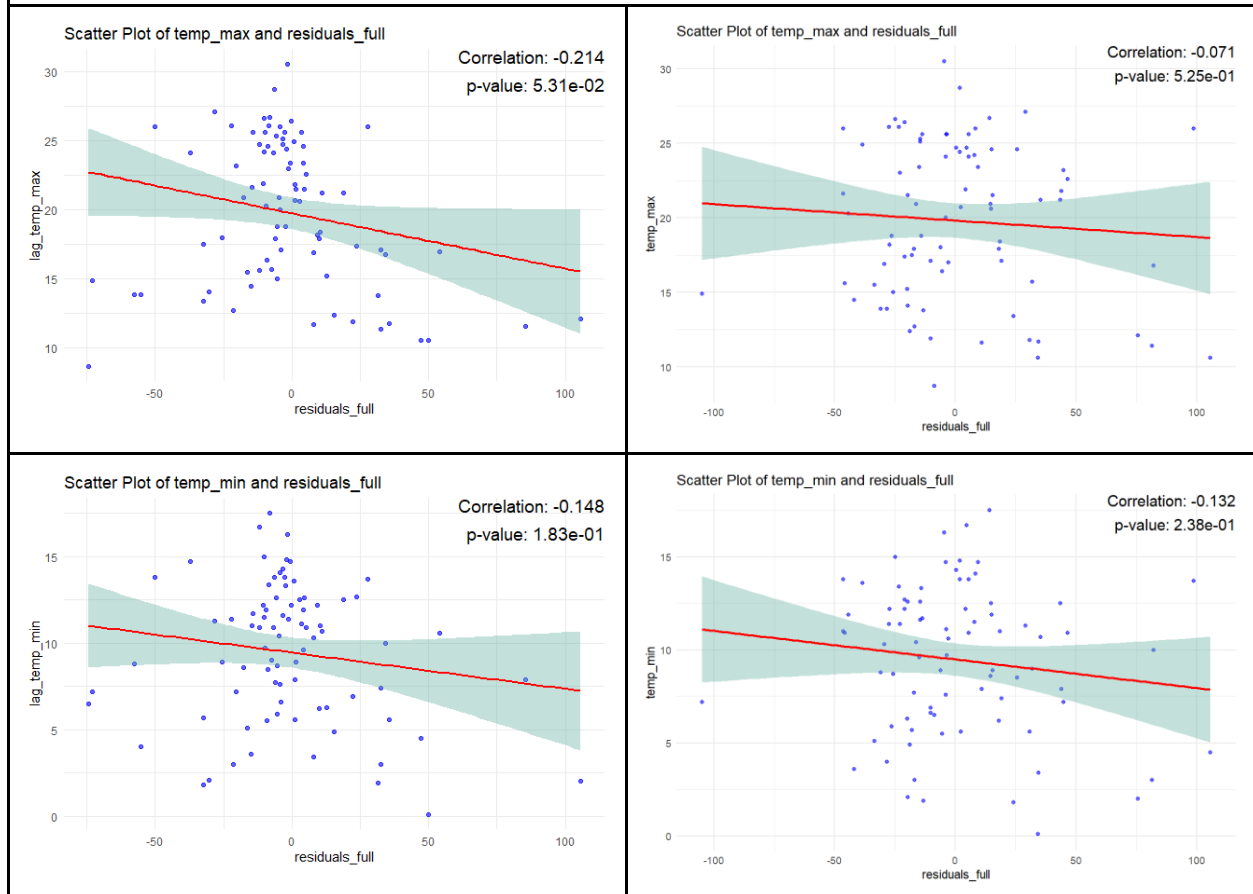
For the admission case, a preliminary exercise was to integrate the maximum and the minimum temperature in an ARIMAX(1,0,0) which, as can be seen below, showed no significant link with the admissions suggesting that the correlation may not exist. To evaluate these correlations individually, a Pearson correlation coefficient was calculated between the residuals of the simple ARIMA (1,0,0) and the maximum and minimum temperatures. While the meanings were negative, as expected, the significance of these relationships presented P values higher than 5% and as such the correlations were found not significant (see figure 1). Nevertheless, the maximum temperature almost passed the test, so the actual lag may not be far from 8.

Similar to the admissions study, an ARIMAX(4,1,1) model was employed to analyze the correlation between COVID-19 deaths and maximum/minimum temperature. This model incorporated both max and min temperature variables but showed no significant correlation with the variables data. Individually, the Pearson correlation coefficient was calculated to assess the correlations between the deaths and maximum/minimum temperature. The analysis revealed that neither the maximum nor the minimum temperature exhibited statistically significant correlations with the COVID-19 deaths. Although the correlation coefficients were calculated, the p-values associated with these correlations were above the established significance threshold.

Figure 1 Testing Mejdoubi, Kyndt, and Djennaoui (2020) correlations	
On the series	
Admissions Correlations	Deaths Correlations

Arimax Model Admissions - temperatures lagged 8		Arimax Model Admissions - temperatures lagged 8	
Dependent variable:		Dependent variable:	
----- admis		----- deces	
ar1	0.932*** (0.037)	ar1	0.860*** (0.055)
intercept	114.629*** (43.502)	intercept	131.590*** (35.143)
lag_temp_max	-1.386 (1.023)	lag_temp_max	-1.259 (1.285)
lag_temp_min	-0.438 (1.341)	lag_temp_min	-2.484 (1.711)
-----		-----	
Observations	82	Observations	82
Log Likelihood	-389.441	Log Likelihood	-405.981
sigma2	762.004	sigma2	1,150.219
Akaike Inf. Crit.	788.882	Akaike Inf. Crit.	821.962
=====		=====	
Note: *p<0.1; **p<0.05; ***p<0.01		Note: *p<0.1; **p<0.05; ***p<0.01	

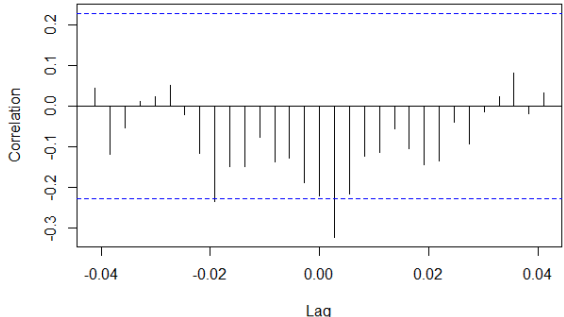
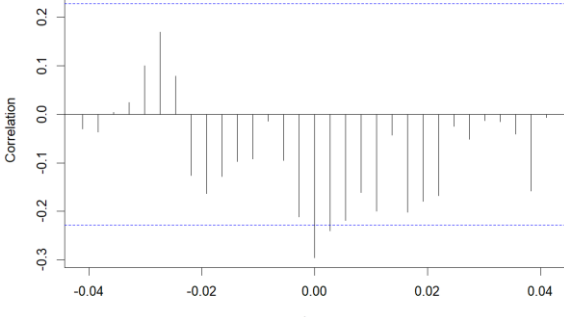
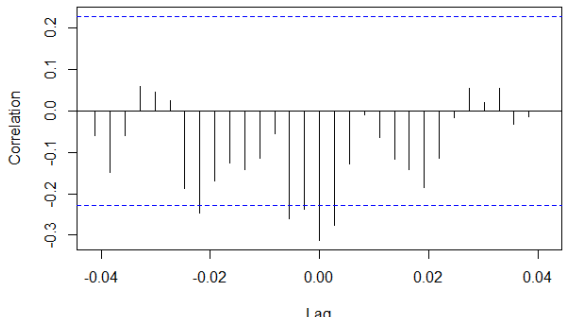
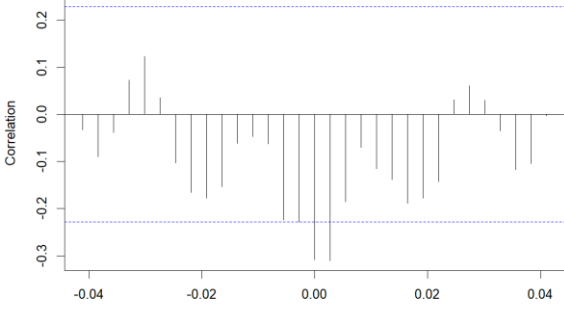
## On the residuals



Seeking for the accurate lags, a Cross Correlation Function was run between the residuals of the ARIMA models and the maximum and minimum temperatures.

In the case of the Admissions, the correlation with the maximum temperature presented significant lags in the first and 7th lag. The algorithm of selection may suggest taking the first of them, but it would make no sense to think that the contamination happened depending on future weather, so we adopt the 7th lag, not far from the 8th day proposal of the paper. Similarly, the calculation for the Minimum temperature suggests a significant lag in the -12 day. Both of these correlations were not significant (see figure 2).

**Figure 2 Scheme for finding optimal lags**

Admis	Deaths
<p><b>Cross-Correlation: admis_residuals vs temp_max</b></p> 	<p><b>Cross-Correlation: deces_residuals vs temp_max</b></p> 
<p>Proposed by the machine:  Lag with maximum correlation: 1  Maximum correlation: -0.3241274  Manually proposed:  Lag with maximum correlation: -7  Maximum correlation: 0.23451397  <b>p-value = 0.05313</b></p>	<p>Proposed by the machine:  Lag with maximum correlation: 0  Maximum correlation: -0.2954058  Manually proposed:  Lag with maximum correlation: -12  Maximum correlation: 0.2745699  <b>p-value = 0.06061</b></p>
<p><b>Cross-Correlation: admis_residuals vs temp_mmin</b></p> 	<p><b>Cross-Correlation: deces_residuals vs temp_mmin</b></p> 

Lag with minimum correlation: -12 Maximum correlation: 0.05852765 <b>p-value = 0.1834</b>	Lag with minimum correlation: -11 Maximum correlation: 0.1227911 <b>p-value = 0.1834</b>
---	--

## Discussion

About the results of the deaths there are some things to be said. First, it is important to remember that the selected ARIMA(1,0,0) filled all the expected conditions expecting the white noise residuals and that can explain the high AIC and the fail divergence in the predicted trend. Second, it is interesting to see that there is actually a correlation between the variables, even if it is not strong enough, given that when the temperature variables with the lags proposed by the authors, was included in an ARIMAX (1,0,0), it actually improved the fit in comparison with the ARIMA (1,0,0), even when the model was still far from expectations(check the annex). Said the, a third remark would be that the results shows that indeed the correlation for the maximum temperature was not far and with one day less, it plays a non significative negative role in the COVID 19 spreading.

These findings indicate that, based on the available data, there is no significant correlation between COVID-19 deaths and either the maximum or minimum temperature. While the correlations did not reach statistical significance, it is important to consider other potential factors that may influence mortality rates, as well as conduct further investigations to gain a comprehensive understanding of the relationship between temperature and COVID-19 deaths.

## Conclusion

While the where can imply impact in the infectious diseases like COVID 19, the data suggests that its role is despicable when it comes to hospital admissions. While preliminary correlations suggest a active role of the environment temperature, our research showed that this is not one of the main factors for explaining the propagation process. While the infection seems to be an autoregressive process, the straightening of the public health policies in the winter and the relaxation in the summer, as well as the intermittent lock downs and the pattern of social immunity may be playing major roles in the data generation process and biasing the statistical observation of the phenomenon.

## Bibliography

- Lutkephol, H. (2004). Applied time series Econometrics. Chapter 2 Univariate Time Series Analysis. Cambridge University press. Isbn-13 978-0-511-21739-5 ebook.
- Nau, R (2020). *Statistical forecasting:notes on regression and time series analysis*. Section “Summary of rules for identifying ARIMA models”. University of Duke”. Retrieved from: <https://people.duke.edu/~rnau/arimrule.htm>

- Mejdoubi, M., Kyndt X., and Djennaoui, M. 2020. *ICU Admissions and in-Hospital Deaths Linked to COVID-19 in the Paris Region Are Correlated with Previously Observed Ambient Temperature*. Plos One 15(11) : e0242268.

For the coding in R:

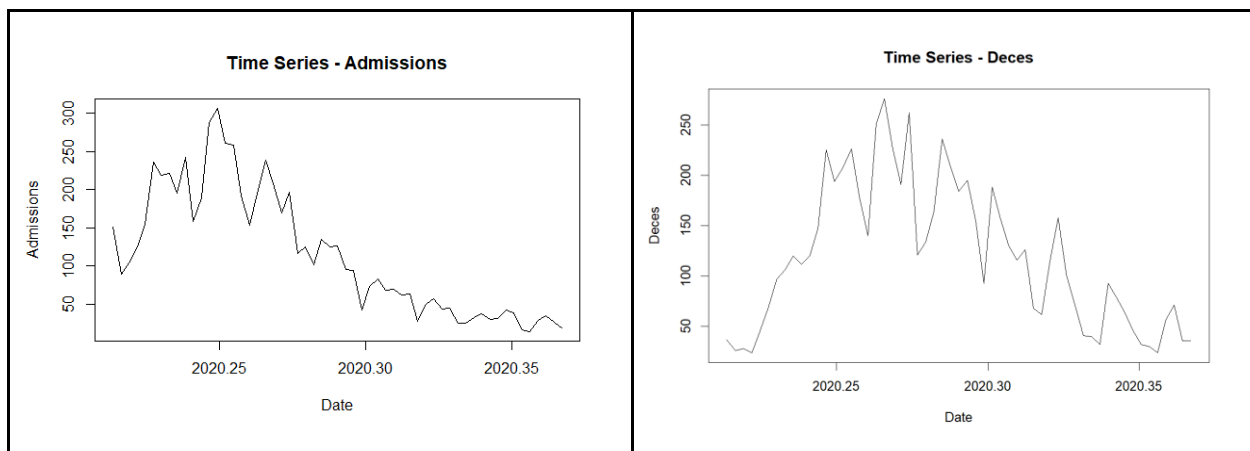
- Sthda.Correlation test between two variables in R. Website. Retrieved from ; <http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r>
- Wickham, H., Hester, J., & François, R. (2018). readr: Read Rectangular Text Data. R package version 1.3.1. Retrieved from <https://CRAN.R-project.org/package=readr>
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(3), 1-22. Retrieved from <https://www.jstatsoft.org/article/view/v027i03>
- Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. Retrieved from <https://CRAN.R-project.org/package=dplyr>
- Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. Retrieved from <https://www.jstatsoft.org/article/view/v040i03>
- Trapletti, A., & Hornik, K. (2019). tseries: Time Series Analysis and Computational Finance. R package version 0.10-48. Retrieved from <https://CRAN.R-project.org/package=tseries>
- Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. Retrieved from <https://CRAN.R-project.org/package=stargazer>
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag. Retrieved from <https://ggplot2.tidyverse.org>
- Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. Retrieved from <https://CRAN.R-project.org/package=ggpubr>

## Annex

### Box jenkins methodology

#### 1. Series Evaluations:

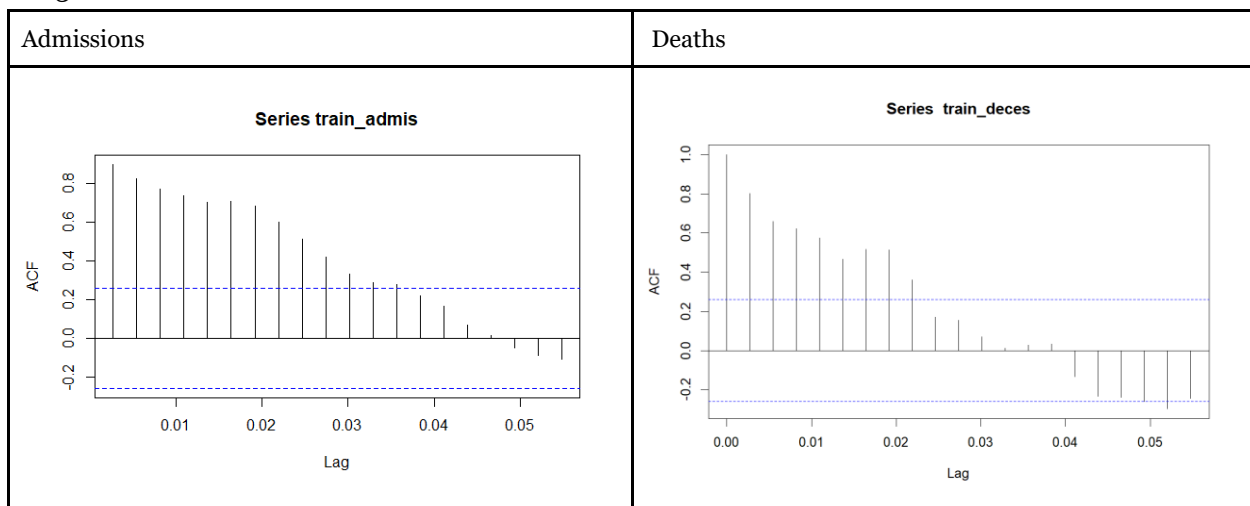
Admissions	Deaths
Train- Test design	Whole series approach



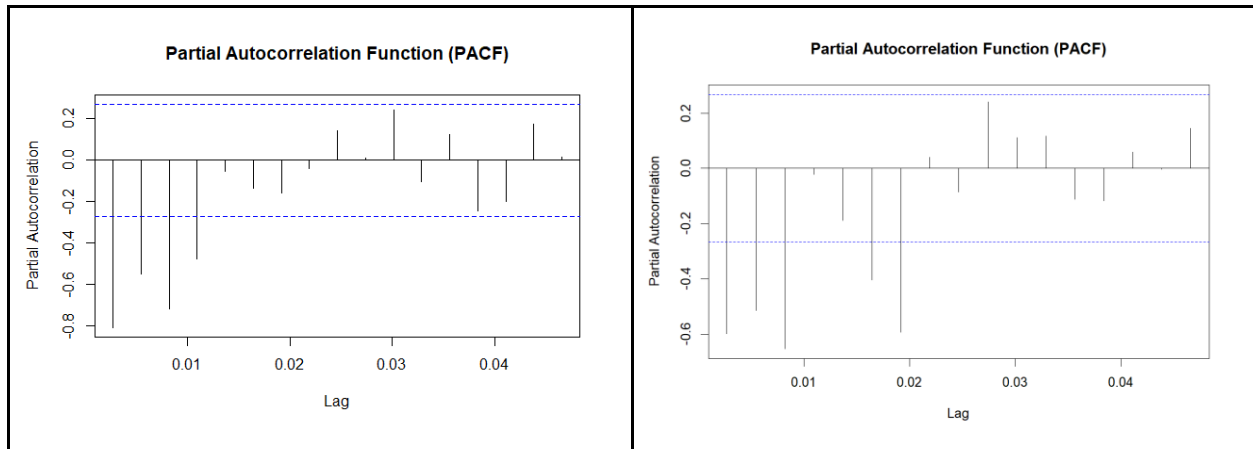
## 2. Stationarity assessment:

Augmented dickey-fuller for admissions	Augmented dickey-fuller for deaths
<p style="text-align: center;">Augmented Dickey-Fuller Test</p> <pre>data: train_admis Dickey-Fuller = -3.6106, Lag order = 3, p-value = 0.04021 alternative hypothesis: stationary</pre>	<p style="text-align: center;">Augmented Dickey-Fuller Test</p> <pre>data: train_deces Dickey-Fuller = -2.6457, Lag order = 3, p-value = 0.3141 alternative hypothesis: stationary</pre>
	<p style="text-align: center;">Augmented Dickey-Fuller Test</p> <pre>data: train_deces_diff Dickey-Fuller = -6.1368, Lag order = 3, p-value = 0.01 alternative hypothesis: stationary</pre>

## 3. Model order selection:







#### 4. Candidate model comparison:

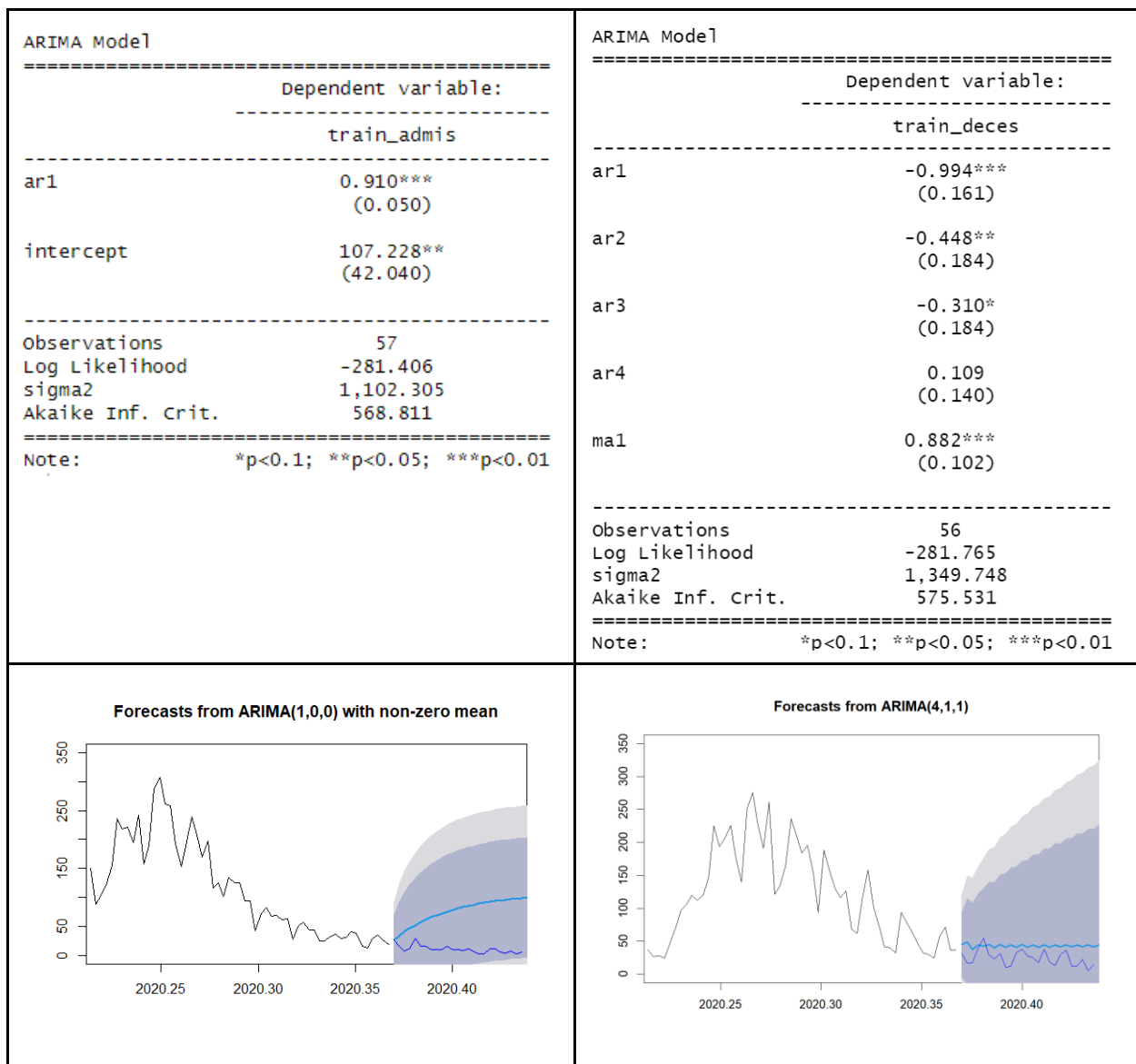
Model comparison			
Admissions		Deaths	
Order (p,d,q)	AIC	Order (p,d,q)	AIC
<b>(1,0,0)</b>	<b>566.81</b>	(1,1,0)	513.59
(1,0,1)	568.42	(1,1,1)	510.18
(2,0,0)	568.51	(2,1,0)	508.94
(2,0,1)	569.26	(2,1,1)	509.71
(3,0,0)	570	(3,1,0)	510.042
(3,0,1)	572.48	(3,1,1)	511.67
(4,0,0)	571.07	(4,1,0)	511.58
(4,0,1)	572.84	<b>(4,1,1)</b>	<b>503.70</b>

#### 5. White noise verification:

Admissions	Deaths
<p>Box-Ljung test</p> <p>data: residuals</p> <p>x-squared = 23.534, df = 12, p-value = 0.02352</p>	<p>Box-Ljung test</p> <p>data: residuals</p> <p>x-squared = 58.583, df = 12, p-value = 4.087e-08</p>

#### 6. Selected model performance

Admissions	Deaths
------------	--------



## Training an ARIMAX including the Authors' proposed lags

Admissions ARIMAX	Deaths
-------------------	--------

# ARIMAX Model

Dependent variable:	
train_admis	
ar1	0.897*** (0.056)
intercept	146.462*** (45.467)
train_temp_max	-1.703 (1.319)
train_temp_min	-0.694 (1.926)
Observations	57
Log Likelihood	-280.288
sigma2	1,062.265
Akaike Inf. Crit.	570.575

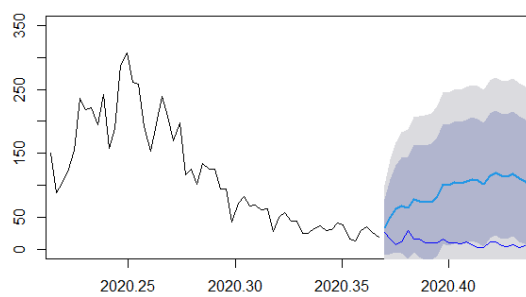
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

# ARIMAX Model full series

Dependent variable:	
deces	
ar1	0.860*** (0.055)
intercept	131.590*** (35.143)
temp_max	-1.259 (1.285)
temp_min	-2.484 (1.711)
Observations	82
Log Likelihood	-405.981
sigma2	1,150.219
Akaike Inf. Crit.	821.962

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Forecasts from Regression with ARIMA(1,0,0) errors



Forecasts from Regression with ARIMA(4,1,1) errors

