# Weekly U.S. Influenza Case Counts

Course Project - Advanced Statistical Modelling: Time Series

**Enric Reverter**
**Louis Van Langendonck**


Facultat d'Informàtica de Barcelona
Universitat Politècnica de Catalunya
01/01/2023

# Contents

**Abstract**

This time series analysis of USA flu data aims to identify patterns and trends in the data, estimate the underlying model of the data, validate the model using statistical techniques, use the model to make predictions about future flu trends, and identify and treat any outlying data points that may impact the analysis. The goal of this analysis is to better understand and forecast flu activity in the USA.

# 1 Problem Description

No disease pandemic has killed more people in absolute numbers than the Influenza virus in 1918-19 [1]. This marks one of many lethal outbreaks of the virus throughout history. The last century however, humankind has greatly improved its knowledge and its vaccine technology, reducing the fatality of the often occurring disease. However, as the world is clearly still vulnerable for pandemics, it is important to track this virus and how it spreads [1]. Therefore, since 1997, the National Respiratory and Enteric Virus Surveillance System (NREVSS) has been tracking Influenza infection across the USA on a weekly basis [2]. The corresponding data can be downloaded from `https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`

The main quantity of interest for this report is the number of specimens across the USA that tested positive for the Influenza virus, regardless of its sub-type. The resulting time series is used for analysis, model building and prediction of the presence of this typically seasonally occurring virus. All relevant code can be found in the R markdown file named `influenza.rmd`.

Finally, two things are important to note. Firstly, that the number of tests has significantly increased over time such that a global increase in positive cases is expected and this trend incorporated in model analysis and forecasting [2]. Secondly, that the scope is restricted to the pre-COVID period as the time during the pandemic does not represent typical viral evolution, testing strategy and human interchange necessary for reliable forecasting.

# 2 Pre-processing

To obtain a time series dataset ready for analysis, a few pre-processing steps are made:

- The data source provides three seperate .csv files: a file concerning Clinical Labs data from after 2015, one from Public Health Labs after 2015 and a combined one from before 2015. The two sources from after 2015 are joined vertically after which this is horizontally joined to the pre-2015 dataset.
- Week and year are extracted for each datarow and saved in "Year_week" format, the time indicator used for time serie analysis.
- Total positive cases is calculated as $total\_specimens * percent\_positive/100$.
- Every four years there is an extra day in the year (leap years), sometimes resulting in one extra week present in the dataset (53 instead of 52). In order to have even amount of weeks, it is chosen to always filter out this last week which only represents a single day difference at the worst.

# 3 Time Series Analysis

## 3.1 Identification

### 3.1.1 Exploratory Analysis

Plots of the untransformed data and the logarithmically transformed data are shown in Figure 3.1. It can be observed that a clear seasonal pattern is present as well as a global rising trend, probably due to increased testing. A relative rise in infections can be observed in 2009 and might be considered

as an outlier. This is probably due to the arrival of the swine flu pandemic [3]. Note that instead of the direct natural logarithm, the ln(x+1)-tranfsormation is used in order to circumvent infinity values for zero values. A small manipulation of the data that should not significantly impact any of the coming results.

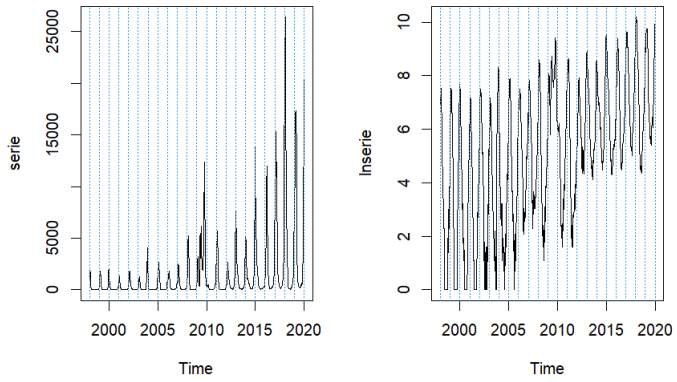



Figure 3.1: Plots of the untransformed time series data (left) and its logarithmic transformation (right).

To see if the variance can be considered constant in time, boxplots and mean-variance plots for both the untransformed and logarithmic data are shown in Figures 3.2 and 3.3. Although not being perfectly constant, the logarithmic transformation seems to improve the homoscedasticity. The `BoxCox.lambda` method yields a value of lambda close to 0 (0.12), for which the logarithmic transformation is further supported. Therefore, it is chosen to work with such transformed serie from now on.

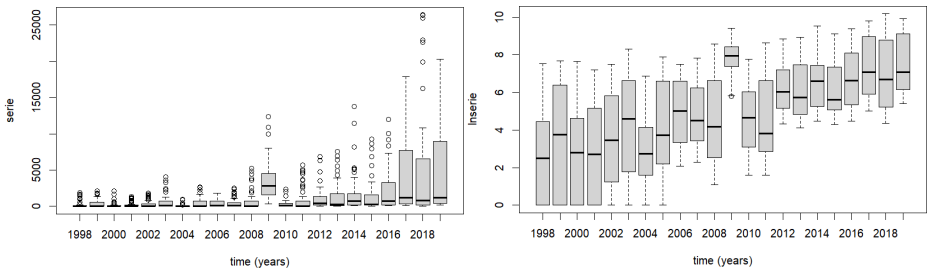



Figure 3.2: Boxplots of the untransformed per-year time series data (left) and its logarithmic transformation (right).

### 3.1.2 Stationarity

First, seasonality is looked at by plotting a subseries for each week, called a 'monthplot' in R, although not restricted to per-month treatment. Moreover, the time series of 52 weeks for each separate year are drawn in the same plot. The results can be found in Figure 3.4. From these, a clear seasonal pattern comes to light: the number of positive cases peaks in the first 10 weeks of the year (january and february), goes down afterwords and starts to rise again around week 40 (in october). A seasonal pattern that is expected for the Influenza virus [1]. To remove the seasonal component a seasonal difference transformation is applied. The plot of this series and the new monthplot is shown in Figure 3.5. These display significant steps in the direction of stationarity. The mean of the series is closer to being constant and the per-week subseries plot indicates that the seasonal pattern has disappeared. Moreover, the total variance of the series has dropped from 6.50 in the original series to 3.25 with seasonal difference. However, The variance (and in a lesser degree the mean) of the series does not seem constant. Therefore, a regular difference is now applied in

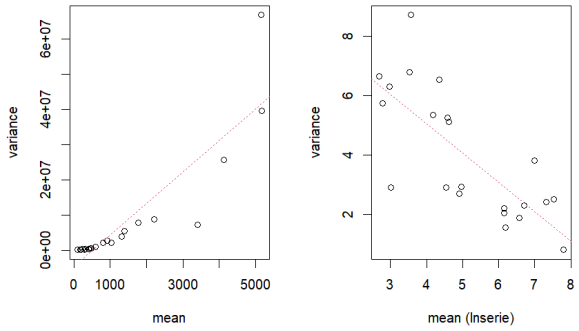



Figure 3.3: Variance vs. mean plots of the untransformed time series data (left) and its logarithmic transformation (right).

order to see if the variance of the series drops even more and a more constant variance and mean can be obtained. The result is plotted on the left of Figure 3.6. This regular difference lowers the total variance even more down to 0.30 and the results seem centered around a more constant mean and a slightly more constant variance. Although some heteroskedasticity is still present, this series can be considered approximately stationary. Finally, a last regular difference is applied in order to see if it lowers variance even further and its resulting series plotted on the right in Figure 3.6. This, however, leads to over-differentiation as variance rises again to 0.68. Therefore, the series with one seasonal and one regular difference is the one that can be considered approximately stationary and will be used for model building.

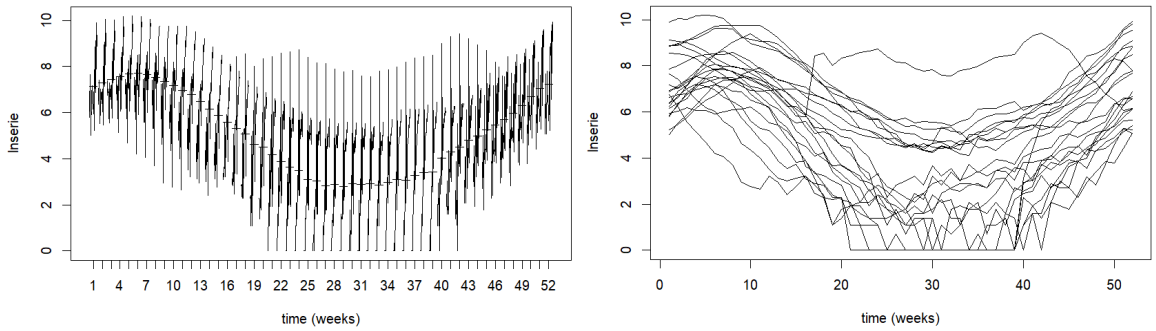



Figure 3.4: The per-week subseries plot (left) and time series for each year plotted together (right).

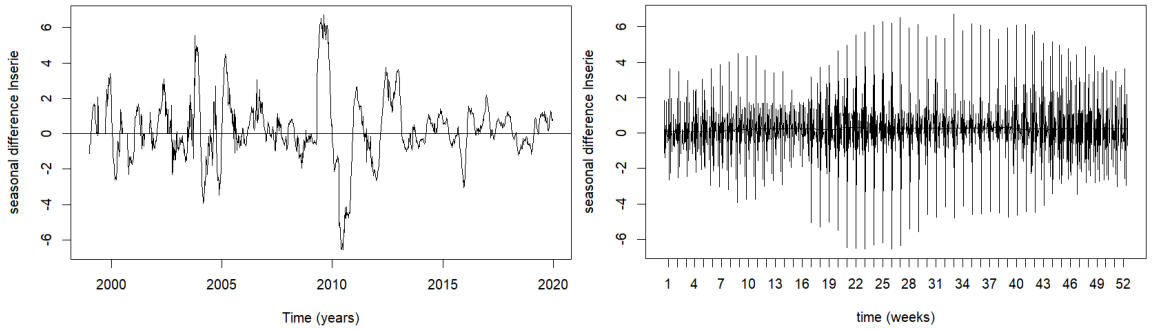



Figure 3.5: Plot of the seasonal difference time series (left) and its per-week subseries plot (right).

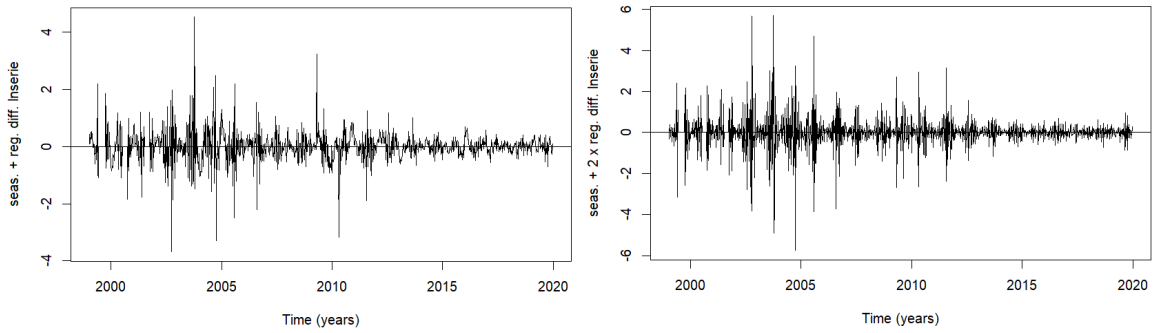



Figure 3.6: Plot of seasonal and one regular difference time serie (left) and plot of seasonal and double regular difference time serie (right).

### 3.1.3 Auto-correlation Analysis

In order to build an arima model, a choice of parameters has to be made using the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF). Because the seasonal lags (in red) are relatively far apart (every 52 lags), these two functions are for readability purposes each plotted twice: for the seasonal pattern with 400 lags on top in Figure 3.7 and for the regular one with 80 lags on the bottom in Figure 3.7.

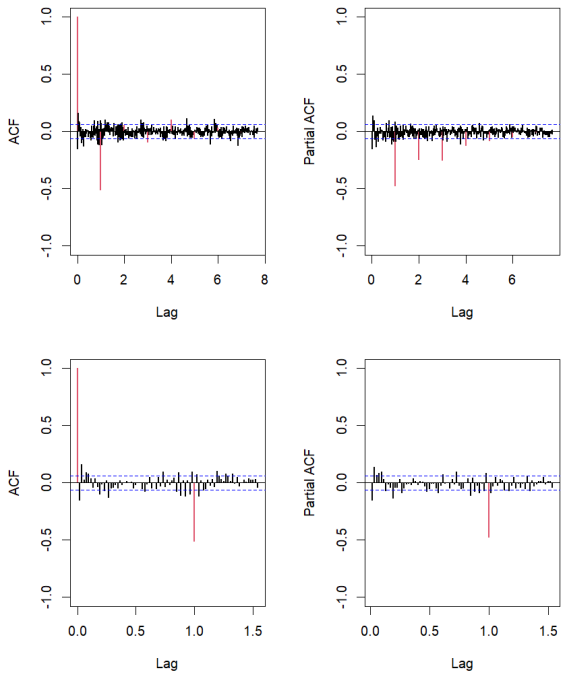



Figure 3.7: Autocorrelation Function and Partial Autocorrelation Function plots with 400 lags (top) and 80 lags (bottom).

First the seasonal pattern is looked into. The following candidate models are decided upon based on the the ACF and PACF plots, holding into account the principle of :

- `ARMA(P=5,D=1,Q=0)_52`: Assuming that after the fifth PACF lag, all later lags are zero while the ACF lags are in a infinitely occurring sinusoidal pattern different from zero.
- `ARMA(P=0,D=1,Q=4)_52`: Assuming that after the fourth ACF lag (without counting the first one), all later lags are zero while the PACF lags are in a infinitely occurring exponentially decreasing pattern different from zero.

- `ARMA(P=0,D=1,Q=1)_52`: Assuming that after the first ACF lag (without counting the first one), all later lags are zero while the PACF lags are in a infinitely occurring exponentially decreasing pattern different from zero. This option is included because the second seasonal lag is already considered zero by the ACF model such that this low complexity model should be tried.

- `ARMA(P=1,D=1,Q=1)_52 and adding extra parameters one by one`: Given that both sides seem to eventually go to zero and strong assumptions have to be made to warrant using an AR or MA model, this approach should be considered to.

Next the regular pattern is looked into. The following candidate models are decided upon based on the the ACF and PACF plots:

- `ARMA(p=5,d=1,q=0)`: Assuming that after the fifth PACF lag, all later lags are zero while the ACF lags are in a infinitely occurring sinusoidal pattern different from zero.

- `ARMA(p=10,d=1,q=0)`: Assuming that after the tenth PACF lag, all later lags are zero while the PACF lags are in a infinitely occurring sinusoidal decreasing pattern different from zero.

- `ARMA(p=0,d=1,q=5)`: Assuming that after the fifth ACF lag (without counting the first one), all later lags are zero while the PACF lags are in a infinitely occurring sinusoidal decreasing pattern different from zero.

- `ARMA(p=1,d=1,q=1) and adding extra parameters one by one`: same intuition as the seasonal pattern.

## 3.2 Estimation

Using the proposed model parameters from Section 3.1.3, two strategies are maintained for model building. The first is trying all 9 combinations of pure AR(P)_52, MA(Q)_52, AR(p) and ma(q) and selecting the best model based on the AIC criteria. This first method yields as best model an `ARMA(10,1,0)(0,1,4)_52`-model whose parameters are further investigated in Section 3.2.1. The second model is obtained by starting from basic ARMA(1,1) models for both the regular and seasonal components and iteratively adding parameters, which yields the best AIC for `ARMA(2,1,1)(1,1,1)_52`.

### 3.2.1 Parameter Significance

First, the parameters of first model (`ARMA(10,1,0)(0,1,4)_52`-model) are displayed in Equation 3.1. Using the t-statistic for parameter significance, it is clear that 5 of the 14 parameters are clearly not significantly contributing (all significantly lower that |2|). The parameter with the lowest absolute t-ratio is set to zero, the model retrained and the new AIC score calculated. If it is lower than the previous model, it is considered the new best model and the process of removing the worst performing non-significant parameter from the model repeated until no more improvement can be found. Using this approach, the final first model has the following parameters turned off: $X_{t-6}, X_{t-7}, X_{t-8}, Z_{52,t-2}$ and $Z_{52,t-3}$. The first model can be expressed as follows:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4-\phi_5 B^5-\phi_9 B^9-\phi_{10} B^{10})(1-B^{52})(1-B)X_t = (1+\Theta_1 B^{52}+\Theta_4 B^{4*52})Z_t, \tag{3.1}$$

where the values for the parameters $\phi_i$ and $\Theta_i$ are present in Figure 3.8.

| ar1 | ar2 | ar3 | ar4 | ar5 | ar6 | ar7 | ar8 | ar9 |
|---|---|---|---|---|---|---|---|---|
| -0.14037396 | 0.12162453 | 0.09648048 | 0.07711532 | 0.10761240 | 0.00000000 | 0.00000000 | 0.00000000 | -0.07692306 |
| ar10 | sma1 | sma2 | sma3 | sma4 | | | | |
| -0.09947053 | -0.91894830 | 0.00000000 | 0.00000000 | 0.06809500 | | | | |

Figure 3.8: Parameter values for Model 1 with $\phi_i$ represented as ar-i and $\Theta_i$ as sma-i.

The same process of parameter significance is done for the second model. First, the intercept is checked to be non-statistically significant, which is the case. Then, the model is simplified to `ARMA(2,1,1)(0,1,1)_52`, since the T-ratio for the Seasonal AR-parameter is not significant. Nothing else seems to improve the AIC at the same time that parameters are significant. As such, the second model is expressed as depicted in Equation 3.2.

$$(1-\phi_1 B-\phi_2 B^2)(1-B^{52})(1-B)X_t = (1+\theta_1 B)(1+\Theta_1 B^{52})Z_t, \tag{3.2}$$

where:

| ar1 | ar2 | ma1 | sar1 | sma1 |
|---|---|---|---|---|
| 0.46495473 | 0.21350292 | −0.59194396 | −0.04875363 | −0.88944563 |

## 3.3 Validation

### 3.3.1 Residual Analysis

In model building, it was assumed that random noise occurred in the form of "white noise" ($Z_t \sim \mathcal{N}(0, \sigma_z^2)$). Therefore, using Residual Analysis, it is checked if three assumptions regarding this distribution are met. First this is done for the first model (`ARMA(10,1,0)(0,1,4)_52`):

1. $\sigma_t^2$ `constant`: The square root plot in Figure 3.9 shows a near constant pattern with only a few outliers, mostly between 2000-2010. The same observation can be made in the residuals plot, where the residuals can be considered more or less constant except for a few outliers earlier on. The studentized Bresuch-Pagan homoscedascitity test yields a p-value of $3.795e-08$, implying heteroskedasticity. In conclusion, the variance is considered heteroskedastic, but approximating homoskedasticity.

2. $\sigma_t^2$ `normal`: The Q-Q plot in Figure 3.9 shows a mostly straight line with slightly heavy tails and a few outliers. The histogram plot however does not clearly show the presence of heavy tails. The shapiro-Wilk and Anderson-Darling Normality tests both have p-value $< 2.2e-16$, implying non-normality. In conclusion, although the residuals exhibit normal-like behaviour, some outliers and heavy tails are present that are not expected in a true normal distribution, reducing model validity.

3. $\sigma_t^2$ `independence`: From the residual Autocorrelation functions displayed in Figure 3.10, it can be seen that most residual lags can be considered zero. The few non-zero ones can be considered random. The LJung-Box test of independence, shown in Figure 3.11, show high-values (above 0.05) for nearby lags but low p-values (below 0.05) for seasonal lags further away. This implies that independence for these far-away lags is not satisfied, implying the noise in the model can not be considered independent from each other, hence reducing model validity.

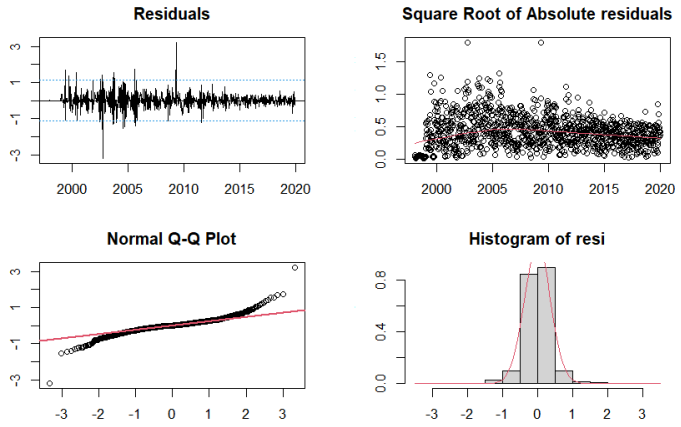



Figure 3.9: Residual Analysis plots for `ARMA(10,1,0)(0,1,4)_52`-model: a residuals plot (top left), a square root of absolute residuals (top right), the quantile-quantile plot (bottom left) and a histogram vs theoretical density plot (bottom right).

The same assumptions are checked for the second model `ARMA(2,1,1)(1,1,1)_52`:

1. $\sigma_t^2$ `constant`: The mean of the residuals is centered, as observed in Figure 3.12, but there seem to be some outliers between 2000 and 2010. The Bresuch-Pagan test implies heteroscedasticity, which might be explained by the mentioned outliers.

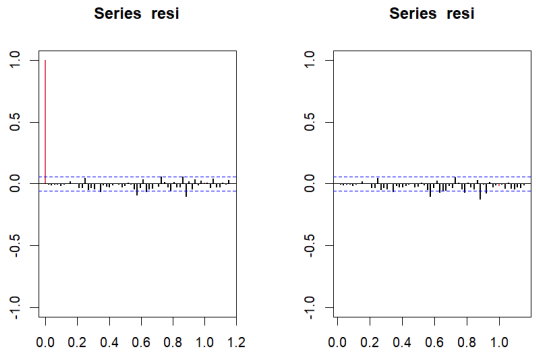



Figure 3.10: Plots of residual Autocorrelation Funtion (ACF) (left) and Partial Autocorrelation Function (PACF) (right) for `ARMA(10,1,0)(0,1,4)_52`-model.

```
Ljung-Box test
      lag.df    statistic     p.value
 [1,]      1   0.02400065 0.876883263
 [2,]      2   0.03679396 0.981771209
 [3,]      3   0.03702620 0.998126026
 [4,]      4   0.03845412 0.999817512
 [5,]      5   0.06795751 0.999937495
 [6,]      6   0.08897556 0.999985806
 [7,]     52  71.87638464 0.035286947
 [8,]    104 144.04917932 0.005722453
 [9,]    156 189.77087991 0.033908533
[10,]    208 246.05130711 0.036338045
[11,]    260 294.66208932 0.068609859
[12,]    312 356.29190615 0.042542894
```

Figure 3.11: LJung-Box test results for `ARMA(10,1,0)(0,1,4)_52`-model.

2. $\sigma_t^2$ `normal`: The Q-Q plot of the same Figure displays tails, which may be related to the outliers, but for the most part follows a straight line. The density estimation looks fine. All statistical tests related to normality are rejected (i.e. Shapiro-Wilk and Anderson-Darling), as expected whenever the sample is large. With proper outlier treatment this issues might be solved.

3. $\sigma_t^2$ `independence between observations`: Some lags are over the confidence bands in the P(ACF) plots, but the Durbin-Watson clearly shows there is no autocorrelation. The Ljung-Box test fails after lag 51.

### 3.3.2 Causality and Invertibility

In order for the models to be considered causal, the roots of the characteristic polynomial of the Autoregressive part $\Phi(B)$ have to lie outside the unit circle or equivalently, the inverse of these roots have to lie inside the unit circle. Similarly, for a model to be invertible, the roots of the characteristic polynomial of the Moving-average part $\Theta(B)$ has to lie outside the unit circle or equivalently, its inverse inside the unit circle. These conditions are tested and plotted for both models. For the `ARMA(10,1,0)(0,1,4)_52`-model, Figure 3.14 shows that all values lie outside the unit circle (modul $> 1$ and inverse within unit circle) for both processes. This confirms that this first model is invertible and causal. Given these properties, the first model can be expressed as both a infinite pure AR process or a infinite pure MA process. These can be expressed as follows:

$$(1 - B^{52})(1 - B)X^t = (\psi_0 + \psi_1 B + \psi_2 B + ...)Z_t \text{ (pure MA}(\infty))$$
$$Z_t = (\pi_0 + \pi_1 B + \pi_2 B + ...)(1 - B^{52})(1 - B)X^t \text{ (pure AR}(\infty))$$

, with the first 25 coefficient values for both processes for the `ARMA(10,1,0)(0,1,4)_52`-model shown in Figure 3.15.

For `ARMA(2,1,1)(1,1,1)_52` the same is true, since all inverted roots lie inside the unitary circle, as depicted in Figure 3.16. That is, it is both invertible and causal. Therefore, the model can be expressed similarly as expressed earlier.

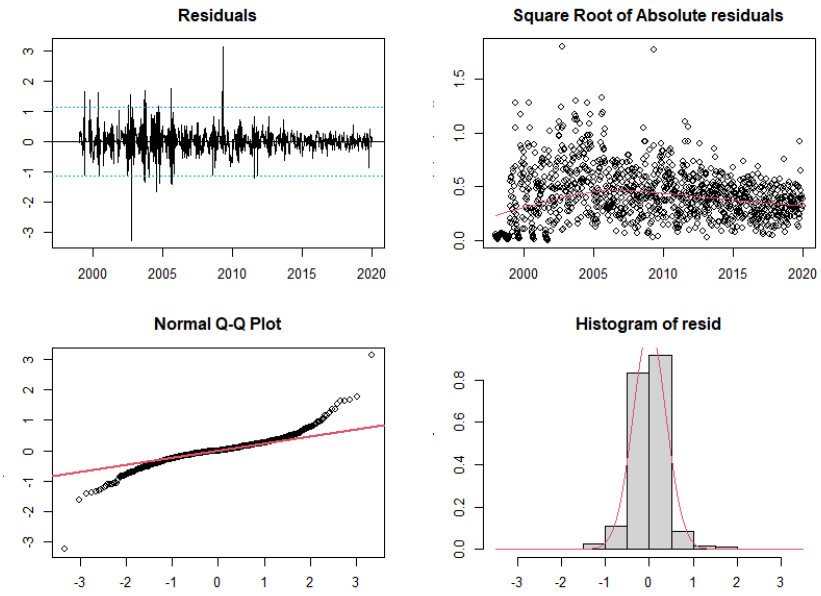



Figure 3.12: Residual Analysis plots for `ARMA(2,1,1)(0,1,1)_52`-model: a residuals plot (top left), a square root of absolute residuals (top right), the quantile-quantile plot (bottom left) and a histogram vs theoretical density plot (bottom right).

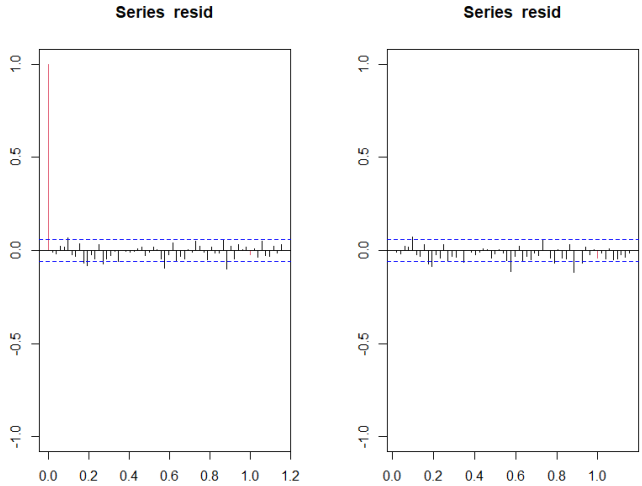
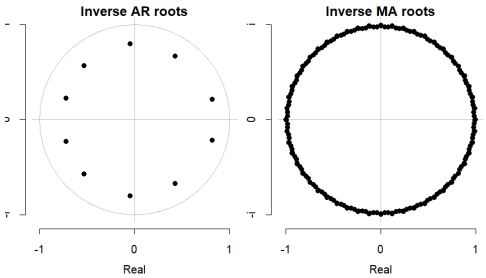



Figure 3.13: Plots of residual Autocorrelation Funtion (ACF) (left) and Partial Autocorrelation Function (PACF) (right) for `ARMA(2,1,1)(0,1,1)_52`-model.

### 3.3.3 Stability Analysis and Model Selection

In order to see which of the two models is the more suitable, the final 12 observations of the data are kept aside and the model retrained without. Subsequently, these 12 observations are predicted by this model including a confidence interval and compared to the actual values (initially kept aside). For model 1, a plot of the result can be found in Figure 3.17. The predictions for the second model are depicted in Figure 3.18. It can be observed how the predictions do not seem to differ much from each other.

In Table 3.1 performance metrics for prediction for both models are summarized. The second model seems to predict more accurately the future observations, as both the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are lower in comparison to the first model. Since the observations are the same for both models, the Percentage errors are proportional to RMSE




```
Modul of AR Characteristic polynomial Roots:  1.261348 1.280803 1.280803 1.190165
1.249788 1.319448 1.249788 1.190165 1.319448 1.261348

Modul of MA Characteristic polynomial Roots:  1.005051 1.010069 1.018475 1.010069
1.005051 1.005051 1.010069 1.005051 1.005051 1.005051 1.005051 1.005051 1.018475
1.005051 1.010069 1.018475 1.005051 1.010069 1.010069 1.005051 1.018475 1.018475
1.005051 1.010069 1.018475 1.005051 1.010069 1.018475 1.018475 1.018475 1.018475
1.005051 1.005051 1.018475 1.018475 1.010069 1.018475 1.010069 1.018475 1.018475
1.010069 1.005051 1.018475 1.005051 1.005051 1.010069 1.005051 1.005052 1.018475
1.005051 1.018475 1.010069 1.005051 1.018475 1.005051 1.010069 1.005051 1.005051
1.005051 1.005051 1.01007 1.010069 1.005051 1.018475 1.018475 1.005051 1.005051
1.018475 1.010069 1.005051 1.005051 1.010069 1.018475 1.005051 1.005051 1.010069
1.005051 1.010069 1.005051 1.018475 1.010069 1.005051 1.010069 1.010069 1.018475
1.005051 1.010069 1.018475 1.018475 1.005049 1.018475 1.018475 1.005051 1.018475
1.005051 1.005052 1.010069 1.018475 1.005051 1.010069 1.018475 1.005049 1.018475
1.010071 1.00506 1.018475 1.018475 1.010062 1.018475 1.018475 1.018475 1.005051
1.018475 1.018475 1.005052 1.018475 1.018475 1.018475 1.018475 1.018475 1.018475
1.018475 1.018489 1.010069 1.018475 1.005051 1.018475 1.018475 1.010069 1.018475
1.018474 1.018475 1.018475 1.018474 1.018438 1.010069 1.018475 1.01005 1.010155
1.018728 1.010069 1.018475 1.005043 1.018475 1.018475 1.005051 1.018475 1.018475
1.018475 1.018452 1.017949 1.01008 1.018475 1.018475 1.018475 1.010069 1.018475
1.01019 1.010069 1.010069 1.018475 1.018479 1.018475 1.010069 1.005052 1.018475
1.018512 1.018475 1.018475 1.018365 1.018475 1.018552 1.018475 1.018475 1.018475
1.010069 1.018482 1.018476 1.010069 1.018456 1.018402 1.018475 1.010069 1.018377
1.018475 1.010069 1.010069 1.018459 1.018475 1.018475 1.010147 1.010059 1.010069
1.018475 1.018318 1.018475 1.018556 1.01007 1.018469 1.01851 1.010073 1.018475
1.01003 1.018441 1.018475 1.010415 1.005198 1.018349
```

Figure 3.14: Inverse roots plot of Autoregressive part (top left) and Moving-Average part (top right) of the as well as the modul values of the respective roots (bottom).

```
       psi 1        psi 2        psi 3        psi 4        psi 5        psi 6
psi 7
-0.1403739589  0.1413293786  0.0595685947  0.0723992158  0.1075049772 -0.0047454898
0.0405289409
       psi 8        psi 9       psi 10       psi 11       psi 12       psi 13
psi 14
 0.0160991636 -0.0586301580 -0.0633711673  0.0090243647 -0.0276683378 -0.0154158847
-0.0269979044
       psi 15       psi 16       psi 17       psi 18       psi 19       psi 20
psi 21
-0.0172067475 -0.0061636529 -0.0132684061 -0.0013795018  0.0044596740  0.0012084651
-0.0002161344
       psi 22       psi 23       psi 24
 0.0030113856  0.0034732312  0.0044400533
```

```
       pi 1        pi 2        pi 3        pi 4        pi 5        pi 6        pi 7
pi 8        pi 9
-0.14037396  0.12162453  0.09648048  0.07711532  0.10761240  0.00000000  0.00000000
0.00000000 -0.07692306
       pi 10       pi 11       pi 12       pi 13       pi 14       pi 15       pi 16
pi 17       pi 18
-0.09947053  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
0.00000000  0.00000000
       pi 19       pi 20       pi 21       pi 22       pi 23       pi 24
 0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
```

Figure 3.15: Coefficients of pure MA($\infty$) (left) and AR($\infty$) (right) models of the `ARMA(10,1,0)(0,1,4)`$_{52}$-model.

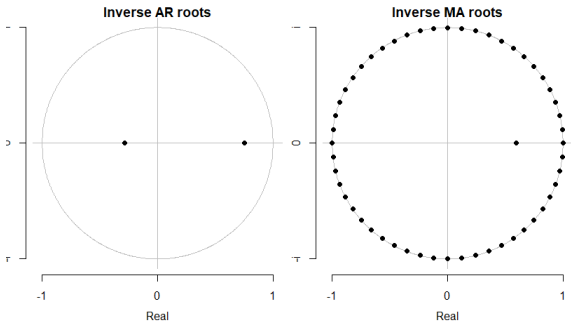



Figure 3.16: Inverse roots plot of Autoregressive part (left) and Moving-Average part (right) of the `ARMA(2,1,1)(0,1,1)`$_{52}$-model.

and MAE differences. The confidence interval bands for the second model, however, are larger. All things considered, the second model is chosen for the long term predictions and further tuning.

Both models are considered stable. For each, the instance with all data and the one without the last 12 observations are similar in magnitude and direction. This means the correlation structure has not changed significantly. If COVID was to be taken into account however, it would certainly

not hold.

```
> modA

Call:
arima(x = lnserie, order = c(2, 1, 1),
seasonal = list(order = c(0, 1, 1), period = 52))

Coefficients:
          ar1     ar2      ma1     sma1
      0.4644  0.2144  −0.5913  −0.9074
s.e.  0.0787  0.0295   0.0765   0.0254

sigma^2 estimated as 0.1479:  log likelihood = −550.36,  aic = 1110.72
> modB

Call:
arima(x = lnserie2, order = c(2, 1, 1),
seasonal = list(order = c(0, 1, 1),
    period = 52))

Coefficients:
          ar1     ar2      ma1     sma1
      0.4666  0.2164  −0.5950  −0.9073
s.e.  0.0776  0.0298   0.0753   0.0257

sigma^2 estimated as 0.1492:  log likelihood = −549.32,  aic = 1108.65
```

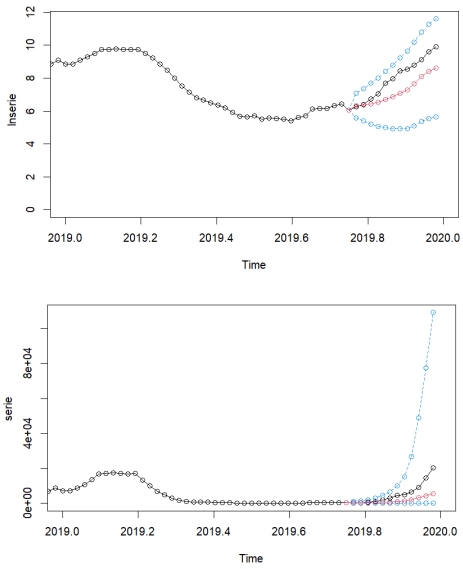



Figure 3.17: Prediction (red line) of last twelve observations (true is black line) with confidence interval (blue lines) by the `ARMA(10,1,0)(0,1,4)_52`-model for both the logarithmically transformed series (top) and the untransformed one (bottom).

## 3.4 Predictions

The `ARMA(2,1,1)(1,1,1)_52`-model is used for long time forecasting. These predictions are depicted in Figure 3.19. Unfortunately, COVID struck during this long term forecast, for which reality looks

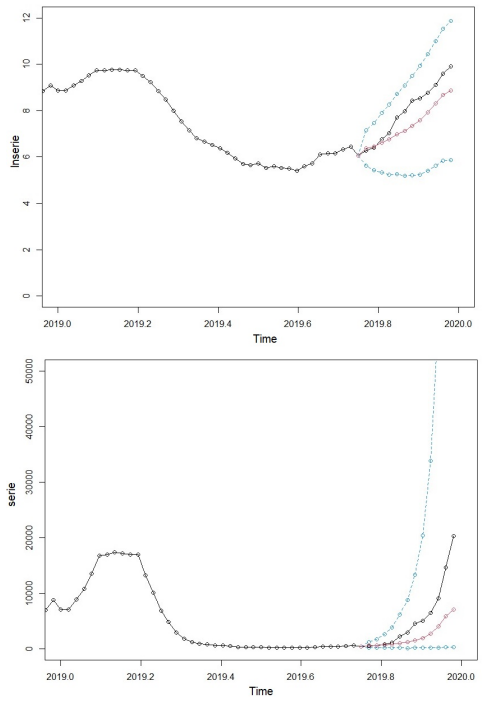



Figure 3.18: Prediction (red line) of last twelve observations (true is black line) with confidence interval (blue lines) by the `ARMA(2,1,1)(1,1,1)_52`-model for both the logarithmically transformed series (top) and the untransformed one (bottom).

Table 3.1: Model performance on last 12 observations

| | **ARMA(10,1,0)(0,1,4)_52** | **ARMA(2,1,1)(1,1,1)_52** | **ARMA(·)_52+OC** |
|---|---|---|---|
| **RMSE** | 5851.561 | 5115.355 | 3612.893 |
| **MAE** | 3854.023 | 3347.569 | 2506.339 |
| **RMSPE** | 0.5783517 | 0.4924453 | 0.4059268 |
| **MAPE** | 0.5182999 | 0.4401799 | 0.3640566 |
| **Mean Length CI** | 23437.63 | 30197.82 | 23488.63 |

quite deviated from the forecast, as shown by the red line. Less influenza cases were recorded since most cases were directly diagnosed as being COVID to not saturate hospitals or because the virus did not have equal chance to spread due to restrictions.

## 3.5 Outlier Treatment

Automatic detection of outliers is applied to the chosen model, `ARMA(2,1,1)(0,1,1)_52`. Many outliers are detected, which are depicted below. Only the most significant of each type will be mentioned. Lots of outliers of all typed are detected. Regarding additive outliers, there is one found in the 39th week of 2003, which could be related to Hurricane Isabel, for example. A level shift is significant during the 17th week of 2009, which coincides with the first 100 days of mandate of President Obama. Some measures for tracking influenza differently might have been applied. However, the used outlier detection method typically used for monthly data, seems ill-suited for the small time unit of weeks that is used for analysis. The short time span allows for stochastic effects to be more present between time units hence resulting in too much detected outliers. Therefore, attempting to identify and link events in the world to these outliers is considered redundant. Although out of scope for the project, in future work, the method used for outlier detection should be reviewed and adapted.

```
   Obs type_detected     W_coeff ABS_L_Ratio    Fecha     perc.Obs
8   73            TC  1.4737734    5.225234 21 1999  436.567733
```

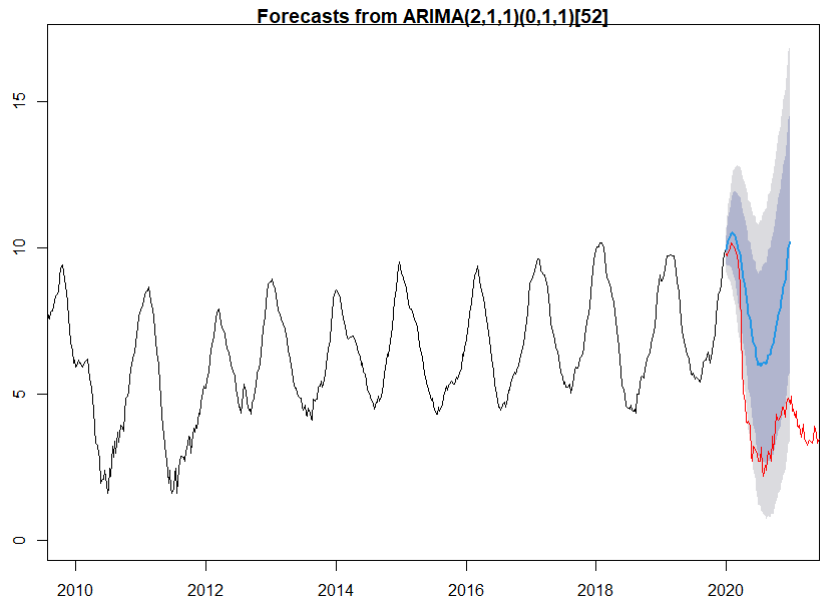



Figure 3.19: Long term forecast. The blue line depicts the prediction, the red one the reality with covid. Confidence bands are also displayed for the forecast.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 74 | 75 | TC | −0.5801093 | 3.000788 | 23 | 1999 | 55.983715 |
| 9 | 92 | LS | 1.5702564 | 5.163024 | 40 | 1999 | 480.788060 |
| 32 | 93 | AO | −0.6531229 | 3.386335 | 41 | 1999 | 52.041803 |
| 65 | 98 | AO | 0.5083372 | 3.081963 | 46 | 1999 | 166.252443 |
| 64 | 107 | LS | −0.6754062 | 3.075786 | 03 | 2000 | 50.894962 |
| 34 | 120 | TC | 0.7899745 | 3.403900 | 16 | 2000 | 220.334032 |
| 53 | 123 | AO | −0.5564332 | 3.203408 | 19 | 2000 | 57.325008 |
| 11 | 125 | LS | 1.5530112 | 5.227319 | 21 | 2000 | 472.567857 |
| 24 | 129 | AO | −0.7558356 | 3.750991 | 25 | 2000 | 46.961806 |
| 67 | 147 | AO | 0.4961351 | 3.032318 | 43 | 2000 | 164.236150 |
| 47 | 174 | AO | −0.6018698 | 3.366636 | 18 | 2001 | 54.778644 |
| 62 | 177 | TC | −0.6225728 | 3.066897 | 21 | 2001 | 53.656219 |
| 44 | 181 | LS | −0.8156681 | 3.390579 | 25 | 2001 | 44.234368 |
| 22 | 196 | AO | 0.8274778 | 4.054932 | 40 | 2001 | 228.754185 |
| 68 | 200 | LS | −0.6685323 | 3.094631 | 44 | 2001 | 51.246016 |
| 25 | 202 | AO | 0.7466225 | 3.727237 | 46 | 2001 | 210.986197 |
| 17 | 232 | LS | −1.2098168 | 4.309166 | 24 | 2002 | 29.825193 |
| 58 | 233 | AO | −0.5402744 | 3.180665 | 25 | 2002 | 58.258838 |
| 26 | 237 | LS | −0.9815058 | 3.717719 | 29 | 2002 | 37.474636 |
| 10 | 238 | AO | 1.1814546 | 5.208603 | 30 | 2002 | 325.911154 |
| 30 | 243 | AO | 0.6663416 | 3.419420 | 35 | 2002 | 194.710106 |
| 12 | 246 | LS | 1.4619132 | 4.972211 | 38 | 2002 | 431.420566 |
| 3 | 248 | LS | −2.4412201 | 7.371379 | 40 | 2002 | 8.705457 |
| 14 | 249 | LS | 1.3486176 | 4.677153 | 41 | 2002 | 385.209653 |
| 5 | 250 | AO | −1.4292760 | 5.927919 | 42 | 2002 | 23.948224 |
| 60 | 253 | TC | −0.6330981 | 3.092277 | 45 | 2002 | 53.094434 |
| 31 | 256 | TC | −0.7951440 | 3.374036 | 48 | 2002 | 45.151620 |
| 61 | 260 | LS | −0.6972614 | 3.135556 | 52 | 2002 | 49.794711 |
| 51 | 283 | TC | 0.7029539 | 3.297849 | 23 | 2003 | 201.971002 |
| 43 | 291 | AO | 0.6432838 | 3.526076 | 31 | 2003 | 190.271879 |
| 6 | 294 | TC | −1.5844921 | 5.476653 | 34 | 2003 | 20.505192 |
| 50 | 295 | AO | −0.5880900 | 3.337295 | 35 | 2003 | 55.538704 |
| 2 | 299 | AO | −1.9146088 | 7.487047 | 39 | 2003 | 14.739948 |
| 54 | 301 | LS | 0.7257596 | 3.164031 | 41 | 2003 | 206.630007 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21 | 302 | LS | 1.1148958 | 4.090470 | 42 | 2003 | 304.925055 |
| 23 | 313 | LS | −1.0331956 | 3.841555 | 01 | 2004 | 35.586793 |
| 46 | 315 | LS | −0.8074124 | 3.388148 | 03 | 2004 | 44.601065 |
| 36 | 324 | LS | 0.8577964 | 3.420771 | 12 | 2004 | 235.795887 |
| 37 | 328 | AO | 0.6487575 | 3.447556 | 16 | 2004 | 191.316219 |
| 19 | 331 | AO | −0.8663316 | 4.151113 | 19 | 2004 | 42.049125 |
| 73 | 333 | LS | 0.6371941 | 3.007544 | 21 | 2004 | 189.116700 |
| 40 | 339 | AO | −0.6274727 | 3.386046 | 27 | 2004 | 53.393951 |
| 7 | 342 | TC | −1.5734266 | 5.508844 | 30 | 2004 | 20.733352 |
| 42 | 343 | AO | 0.6210380 | 3.385037 | 31 | 2004 | 186.085855 |
| 66 | 346 | TC | 0.6173774 | 3.090228 | 34 | 2004 | 185.405914 |
| 70 | 349 | AO | 0.4898108 | 3.028384 | 37 | 2004 | 163.200743 |
| 13 | 351 | AO | 1.0570468 | 4.812188 | 39 | 2004 | 287.785957 |
| 27 | 354 | AO | 0.7163634 | 3.616212 | 42 | 2004 | 204.697568 |
| 45 | 359 | TC | −0.7459060 | 3.397238 | 47 | 2004 | 47.430440 |
| 33 | 392 | TC | −0.7943740 | 3.402422 | 28 | 2005 | 45.186402 |
| 4 | 395 | AO | −1.6878091 | 6.890434 | 31 | 2005 | 18.492424 |
| 15 | 400 | AO | −1.0093138 | 4.681567 | 36 | 2005 | 36.446899 |
| 20 | 402 | AO | −0.8610757 | 4.155835 | 38 | 2005 | 42.270713 |
| 76 | 403 | AO | 0.4571735 | 2.893118 | 39 | 2005 | 157.960296 |
| 28 | 406 | AO | 0.7032010 | 3.568507 | 42 | 2005 | 202.020915 |
| 59 | 440 | LS | −0.6922688 | 3.083868 | 24 | 2006 | 50.043938 |
| 69 | 453 | TC | 0.6023653 | 3.050335 | 37 | 2006 | 182.643375 |
| 71 | 469 | TC | −0.5944565 | 3.033247 | 01 | 2007 | 55.186242 |
| 57 | 490 | LS | −0.7217328 | 3.186506 | 22 | 2007 | 48.590952 |
| 48 | 493 | AO | −0.6004717 | 3.370213 | 25 | 2007 | 54.855284 |
| 18 | 553 | AO | −0.8802296 | 4.180099 | 33 | 2008 | 41.468770 |
| 39 | 558 | AO | 0.6455636 | 3.461212 | 38 | 2008 | 190.706152 |
| 1 | 589 | LS | 3.0963965 | 8.897316 | 17 | 2009 | 2211.810522 |
| 52 | 619 | LS | −0.7651743 | 3.298993 | 47 | 2009 | 46.525281 |
| 55 | 621 | LS | −0.7246167 | 3.165580 | 49 | 2009 | 48.451026 |
| 56 | 623 | LS | −0.7785924 | 3.418648 | 51 | 2009 | 45.905173 |
| 49 | 645 | TC | −0.7287296 | 3.376020 | 21 | 2010 | 48.252158 |
| 41 | 650 | TC | −0.7640019 | 3.399881 | 26 | 2010 | 46.579862 |
| 75 | 699 | AO | −0.4816607 | 3.026717 | 23 | 2011 | 61.775660 |
| 16 | 706 | AO | −0.9840439 | 4.590624 | 30 | 2011 | 37.379645 |
| 63 | 715 | AO | 0.5130371 | 3.071229 | 39 | 2011 | 167.035661 |
| 29 | 717 | TC | −0.8427394 | 3.525937 | 41 | 2011 | 43.052950 |
| 38 | 759 | TC | 0.7876024 | 3.445298 | 31 | 2012 | 219.811999 |
| 72 | 815 | LS | 0.6515406 | 3.044821 | 35 | 2013 | 191.849419 |
| 77 | 819 | AO | 0.4675924 | 2.953312 | 39 | 2013 | 159.614676 |
| 35 | 1132 | LS | −0.8806722 | 3.416644 | 40 | 2019 | 41.450418 |

Subsequently, the model is linearized, which can be depicted in Figure **??**. There are so many outliers identified that is difficult to analyze them just by looking at its profile. The P(ACF) of the linearized series can be observed in Figure 3.21. It seems like other models could be tried, such as `ARMA(4,1,0)(1,1,0)_52`, or `ARMA(0,1,6)(0,1,1)_52`, but the one that accomplishes the best AIC with all parameters significant is the same `ARMA(2,1,1)(0,1,1)_52`. The same last 12 observations are then predicted with it, the results of which are displayed in Table 3.1. Again, the model is stable, but more importantly, it clearly outperforms the other models. The predicted values are depicted in Figure 3.22.

# 4 Conclusions

The goal of this time series analysis is to understand and forecast pre-COVID (until 2020) flu activity in the USA. The variable of interest is the total amount of positive cases per week across the USA. Due to leap years, a small correction for calendar effects had to be made by omitting the 53th week

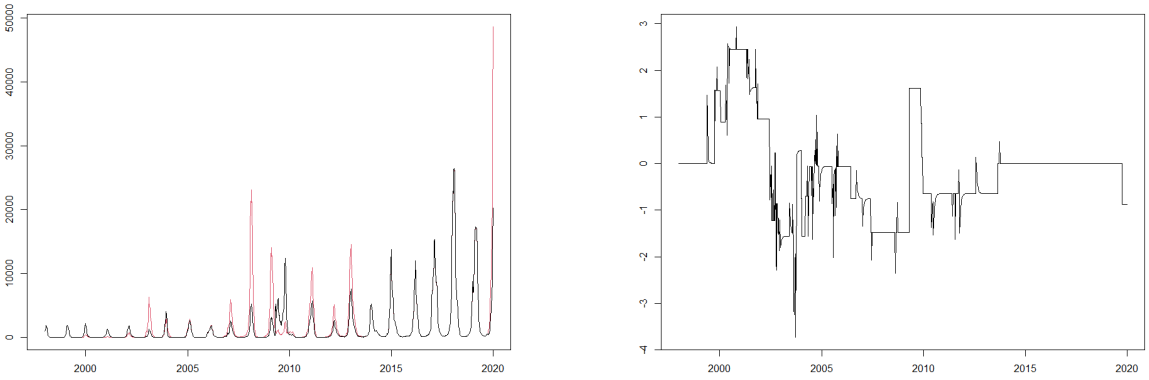



Figure 3.20: Linearized series (red) on top of the serie with the outiers. Effect of the outliers in the log-transformed series.

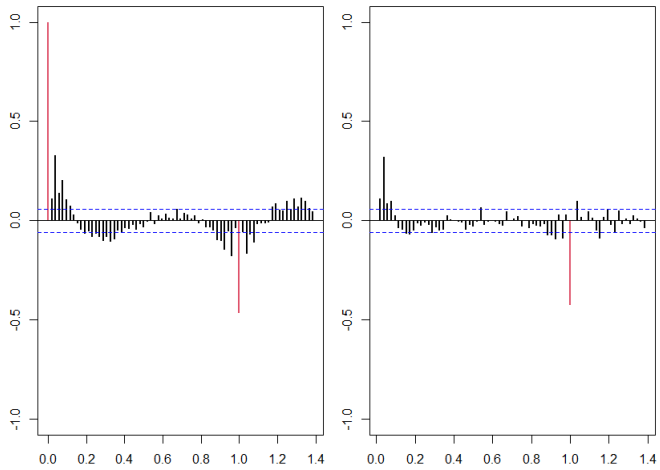



Figure 3.21: (P)ACF of the linearized serie.

every four years. In the identification part, it was found that mulitple transformations are necessary to obtain stationary data: A logarithmic transformation, a seasonal (52-week) difference and a regular difference. Based on the (P)ACF of the transformed series, two candidate arima models where considered: an `ARMA(10,1,0)(0,1,4)_52`-model and an `ARMA(2,1,1)(0,1,1)_52`-model. For each of these, parameters are estimated and, if not significant, removed from the model. Afterwords, both models are extensively validated using analysis of the models residuals, causality, invertibility and stability. Based on the results, the `ARMA(2,1,1)(0,1,1)_52`-model was considered the better performing model. This model was finally used for forecasting. Although the observed cases after 2020 lie on the edge of the confidence intervals of the forecast, they seem to lie significantly lower than prediction. This is due to COVID disrupting the spread of the flu virus and are hence considered non-representative observations. Therefore, the final model and forecast can still be considered a plausible estimation of typical, non-COVID, flu-virus case evolution. Finally, outlier analysis is applied to the model. It seemed to be oversensitive to the stochastic nature of the weekly data, such that it is proposed that the outlier detection method should be revised. However, applying linearization on the currently detected outliers results in a more valid and stable model, hence constituing the final model to be used.

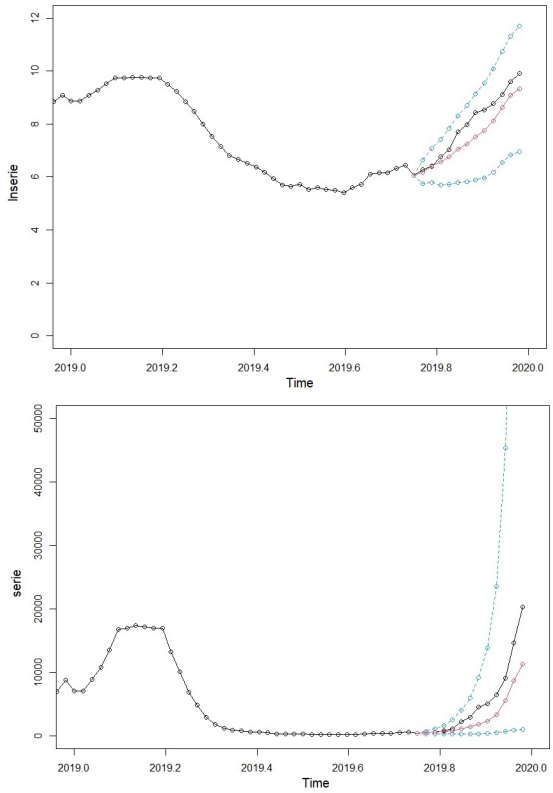



Figure 3.22: Prediction of last 12 observations (stability analysis) of the linearized model for the logarithmically transformed (top) and untransformed (bottom) series.

# References

[1] S. Lemon and A. Mahmoud, "The threat of pandemic influenza: Are we ready?," *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, vol. 3, pp. 70–3, 02 2005.

[2] "U.s. influenza surveillance: Purpose and methods," Oct 2022.

[3] S. Al hajjar and K. McIntosh, "The first influenza pandemic of the 21st century," *Annals of Saudi medicine*, vol. 30, pp. 1–10, 03 2010.