Chickenpox Counts in New York City from Jan 1931 to Jun 1972

Stat 248 Final Project

Hector Roux de Bézieux

April 2018

1) Introduction

Chickenpox is a highly contagious airborne disease and no vaccine was available until the end of the 1980's. Before that, adults and children were affected with no known immunization. While the disease is relatively harmless in children, it can much more often be lethal in adults, especially in pregnant women where it was linked to pneumonia. Therefore, correctly predicting future trends in disease propagation is very useful. While chickenpox is now lesser threat, the methods developed here are also relevant to any other contagious disease.

Here, we focus on chickenpox cases in New York City in the middle of the 20^{th} century, for 498 consecutive months. The data is available at https://datamarket.com/data/set/22v7/%20monthly-reported-number-of-chickenpox-new-york-city-1931-1972/#\%20protect\%20kern-.1667em\%20relaxds=22v7&display=line

2) Exploratory Data Analysis

The first thing we can do is plot the time series: We can clearly see annual periodicity in the plot, as well as a downward trend starting from the 1950's.

```
y <- ts(chickenpox, frequency = 12, start = 1931)
plot(y)
```

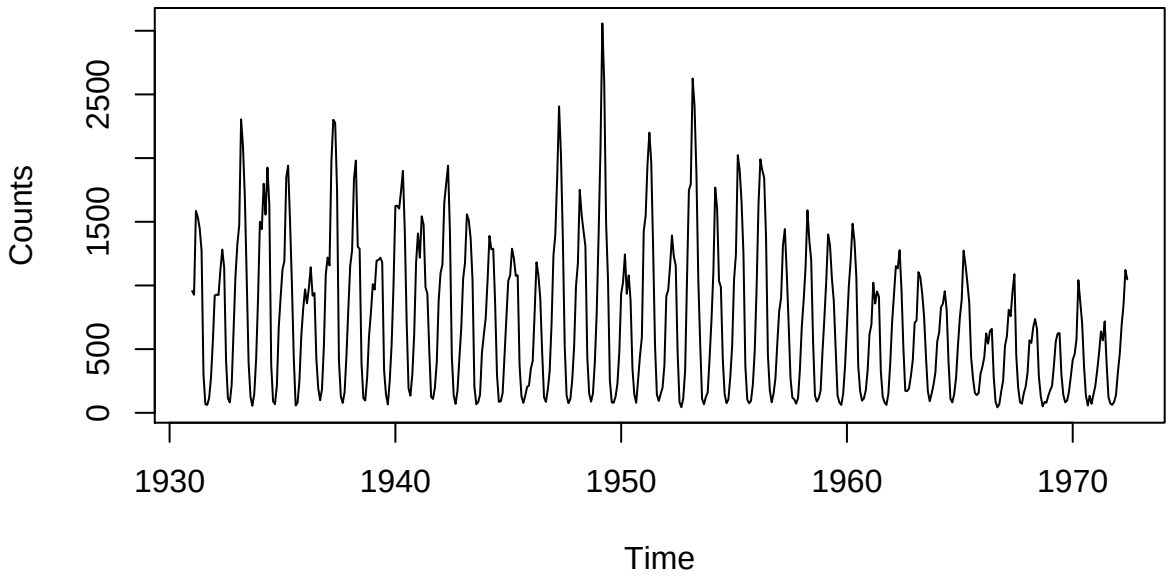


Figure 1: Visualization of the time series

We can then plot the spectral representation of the time series: as we can see, we have especially high peaks at 1 year, and the harmonic frequencies of 1 year, as could be observed already from the raw plotting of the time series.

```
t <- spectrum(y)
abline(v = c(1, 2, 3, 4, 5), col = "red")
```

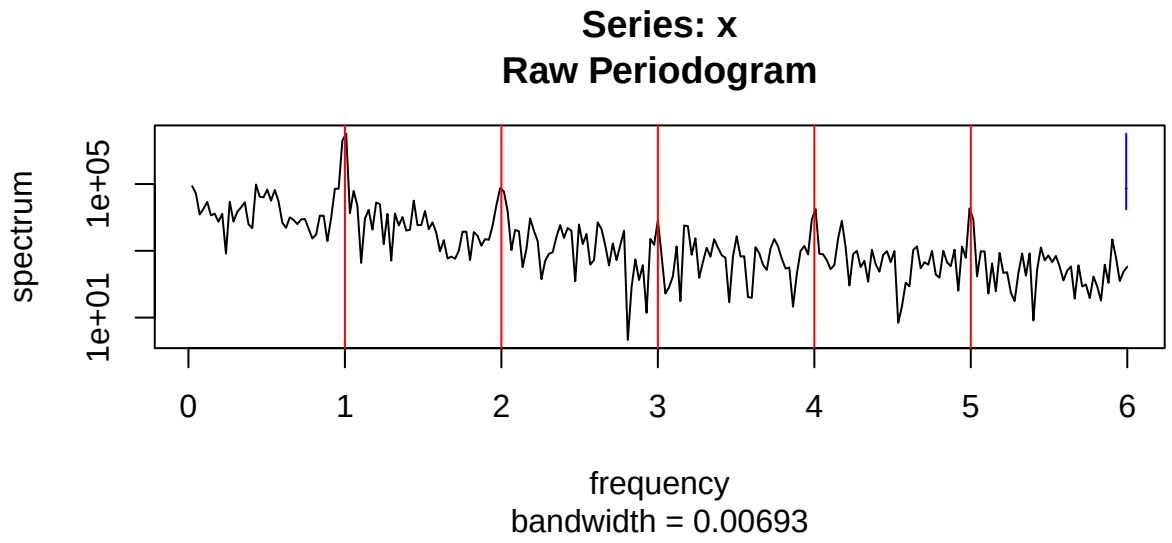


Figure 2: Raw Periodogram

Finally, given the annual periodicity, it ca be especially interesting to look at the lagged difference for this time series. So we perform differentiation with a lag of 12. However, the new time series does not seem to exhibit any particularly interesting behavior so we will stick to modeling the initial time series.

```
plot(y - lag(y, 12))
```

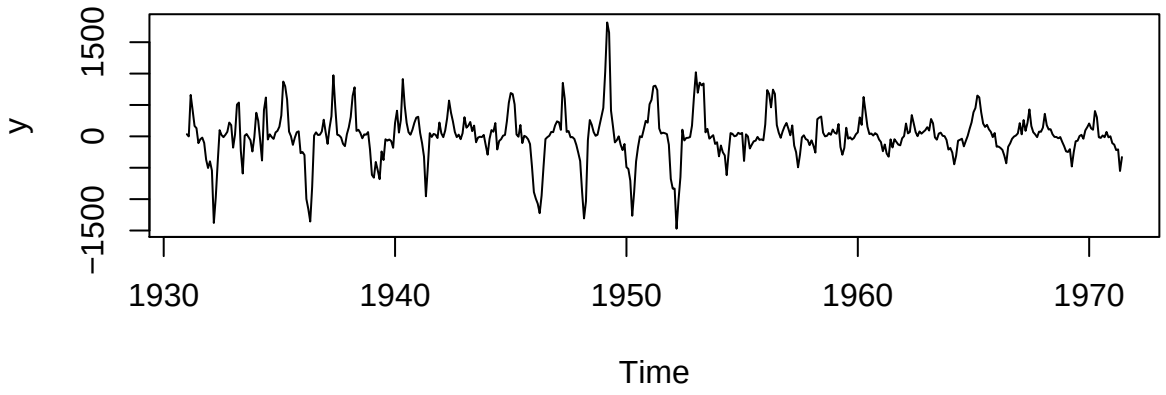


Figure 3: Lagged-difference

3) Modelization

1) General idea

We write Y_t, the realization of our time series measuring the number of occurrences. Our global model is to write $Y_t = f(t) + X_t$ where $X_{t+1} = g(X_{t-k}, X_{t-k+1}, ..., X_{t-1})$. We will try a variety of functions f and g (and values of k) in the next sections. The reasoning behind choosing such a model comes from the graph of the time series where we can see a behavior that only depend on time but not on previous realizations and a part that depend on time.

2) Choice of f

1) First model: linear model

We model $f(t) = \alpha + \beta t + \sum_{i=1}^{5} c_i(cos(\frac{2\pi it}{12}) + sin(\frac{2\pi it}{12}))$ with t in months.

```
c <- s <- list()
for(i in 1:5){
  c[[i]] <- cos(2*i*pi*(1:498)/12)
  s[[i]] <- sin(2*i*pi*(1:498)/12)
}
design <- data.frame(1:498, c, s, y)
colnames(design) <- c("t", paste0("c", 1:5), paste0("s", 1:5), "y")
lm_fit <- lm(y~., data = design)
plot(y)
lines(seq(from=1931,length.out=498, by=1/12), fitted(lm_fit), col = "green")
```

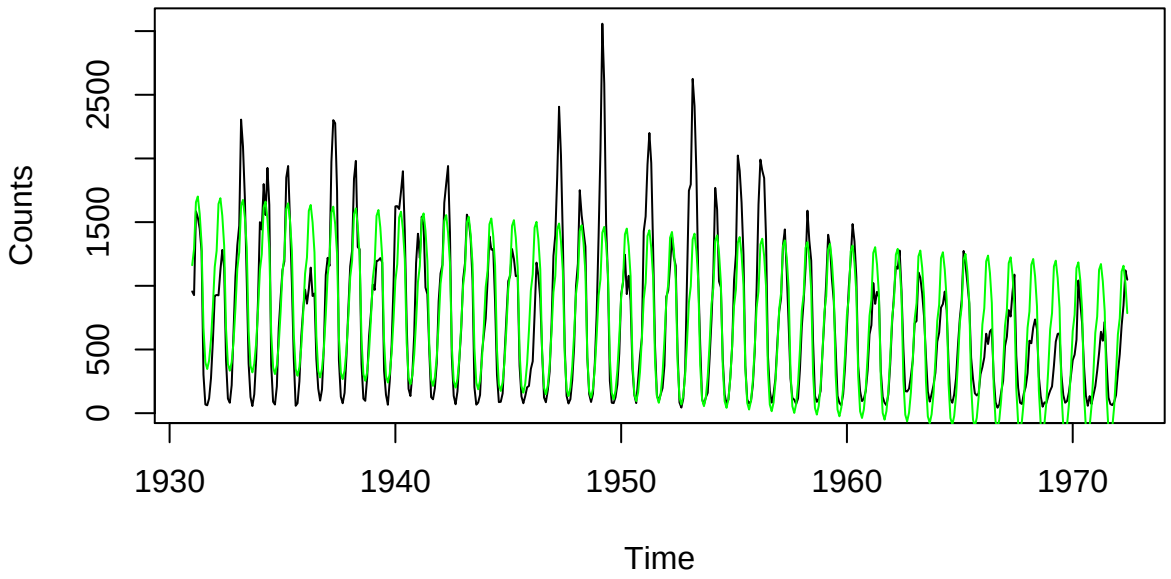
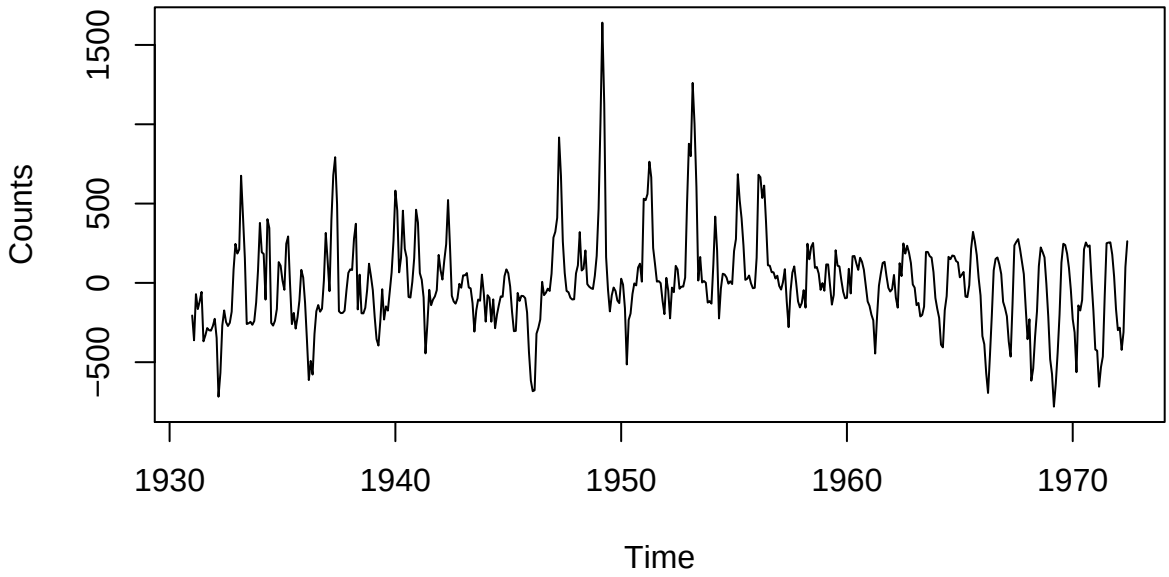


Figure 4: Time series and fitted value of the linear model

As we can see, the fitted values are off from the real values. If we plot the residuals, we see a clear trend.

```r
plot(y - fitted(lm_fit))
```



2) Second model: GLM

Plotting the log of counts seem to lead to a more uniform trend so we decide to update the model to $log(X_t) = f(t) + \epsilon_t$.

```r
plot(log(y))
```

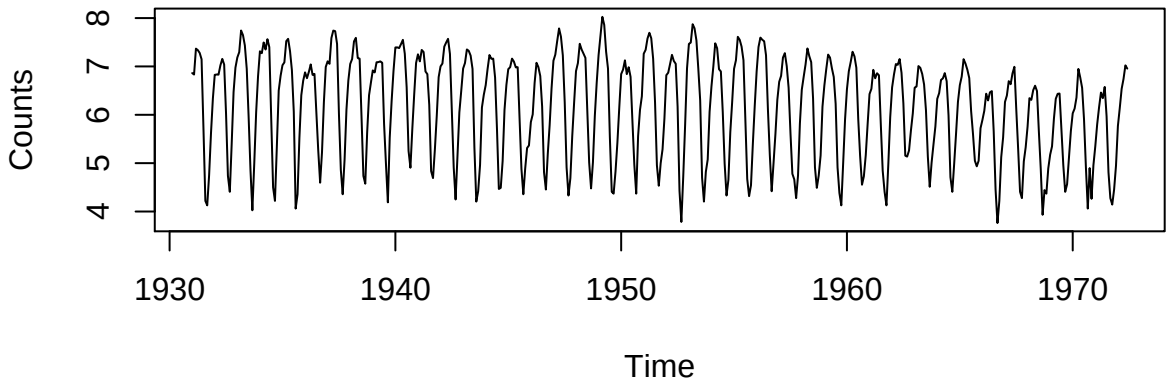


Figure 5: log counts as a function of time

The log representation of the time series incite us to fit a Generalized Linear model with a Poisson assumption, since the log seems a good link function in this case. Moreover, we noticed on Figure 1 that the trend seems to be different from various blocks of time so instead of fitting a polynomial as a function of time, we fit a cubic spline with 3 degrees of freedom, S_t. So we get:

$$f(t) = S_t + \sum_{i=1}^{5} c_i(cos(\frac{2\pi it}{12}) + sin(\frac{2\pi it}{12}))$$

```r
xspline <- ns(c(1:498), df = 4)
Poisson_fit <- glm(y ~ xspline + ., data = design, family = "poisson")
plot(y)
lines(seq(from=1931,length.out=498, by=1/12), fitted(Poisson_fit), col = "green")
```

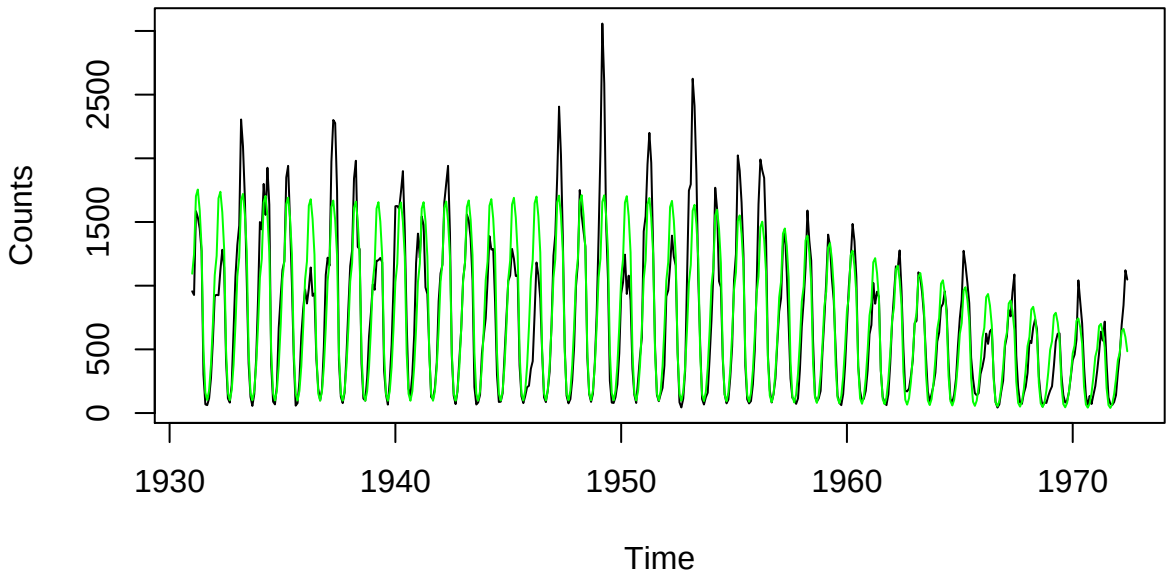


Figure 6: Time series and fitted value of the generalized linear model

If we now look at the residuals from our model, we can see that they have a mean of roughly zero.

```r
plot(y - fitted(Poisson_fit))
```

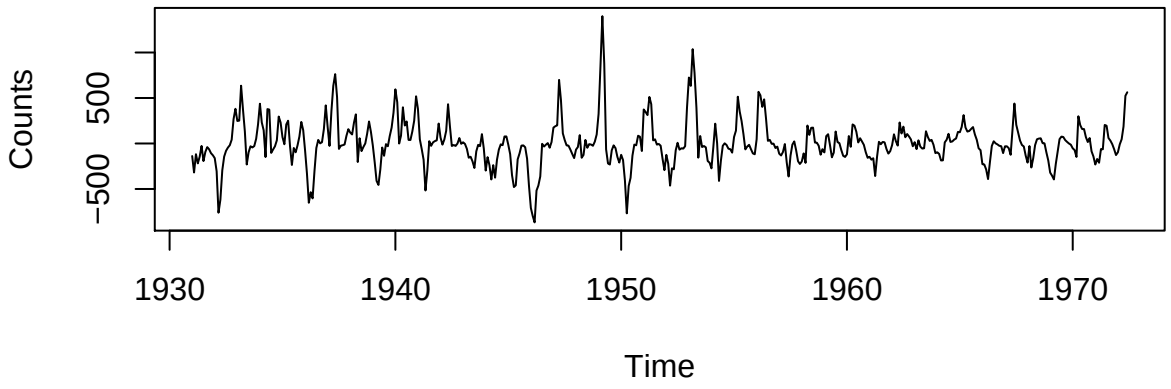


Figure 7: Residual from the generalized linear model as a function of time

This shows that we managed to transform our time series to a stationary time series. We can now focus on fitting a model to the residuals

3) Choice of X_t

1) Scaling of the residuals

The first thing we can notice is that higher values have higher variance than smaller values. Therefore, we focused on the scaled-residuals, also known as Pearson's residuals $Z_t = \frac{Y_t - \hat{Y}_t}{\sqrt{\hat{Y}_t}}$. In Figure 8, we can see that this lead to a more homogeneous residuals

```
Z <- (y - fitted(Poisson_fit))/sqrt(fitted(Poisson_fit))
plot(Z)
```

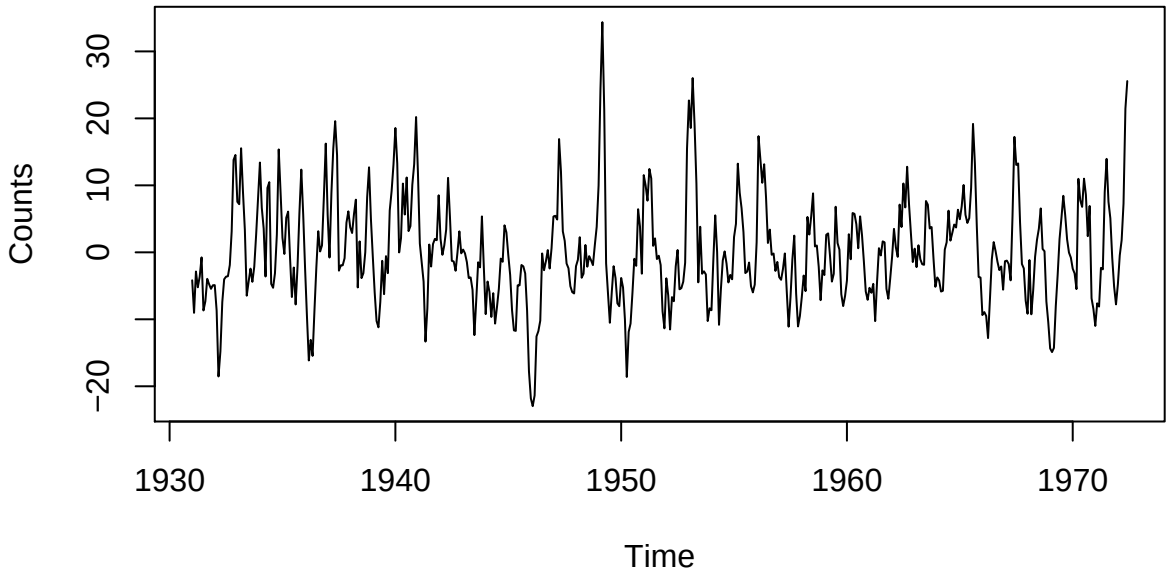


Figure 8: Scaled Residuals

2) Finding the best-model for Z_t

We fit an ARMA model on Z_t such as we have: $Z_t = \sum_{i=i}^{p} \phi_i Z_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$, with ϵ_t white-noise. We try to find the best p and q using the Bayesian Information Criterion (BIC) as a model selection criterion.

```
auto.arima(Z, d=0, ic = "bic", stepwise = F)
```

```
## Series: Z
## ARIMA(2,0,0) with zero mean
##
## Coefficients:
##          ar1      ar2
##       0.9365  -0.2273
## s.e.  0.0437   0.0441
##
## sigma^2 estimated as 24.56:  log likelihood=-1503.24
## AIC=3012.48   AICc=3012.53   BIC=3025.11
```

Using BIC, we select an ARMA(2, 0) model for Z_t so we model $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \epsilon_t$.

```r
epsilon <- Z - fitted(Arima(Z, order = c(2, 0, 0)))
plot(epsilon)
```

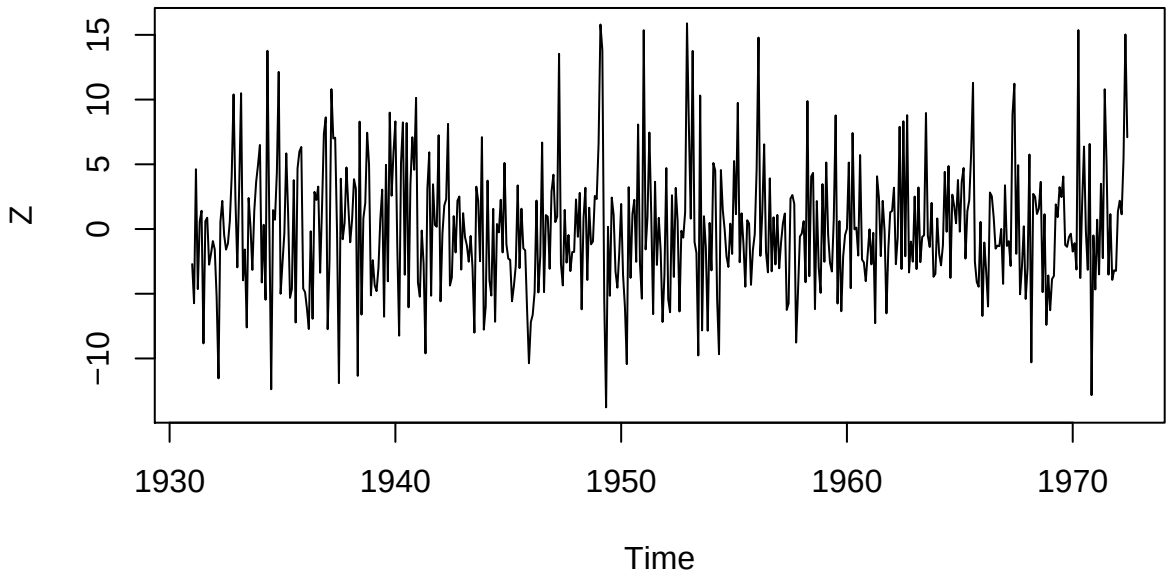


Figure 9: View of the residuals from the ARMA(2, 0) model

```r
spectrum(epsilon)
```

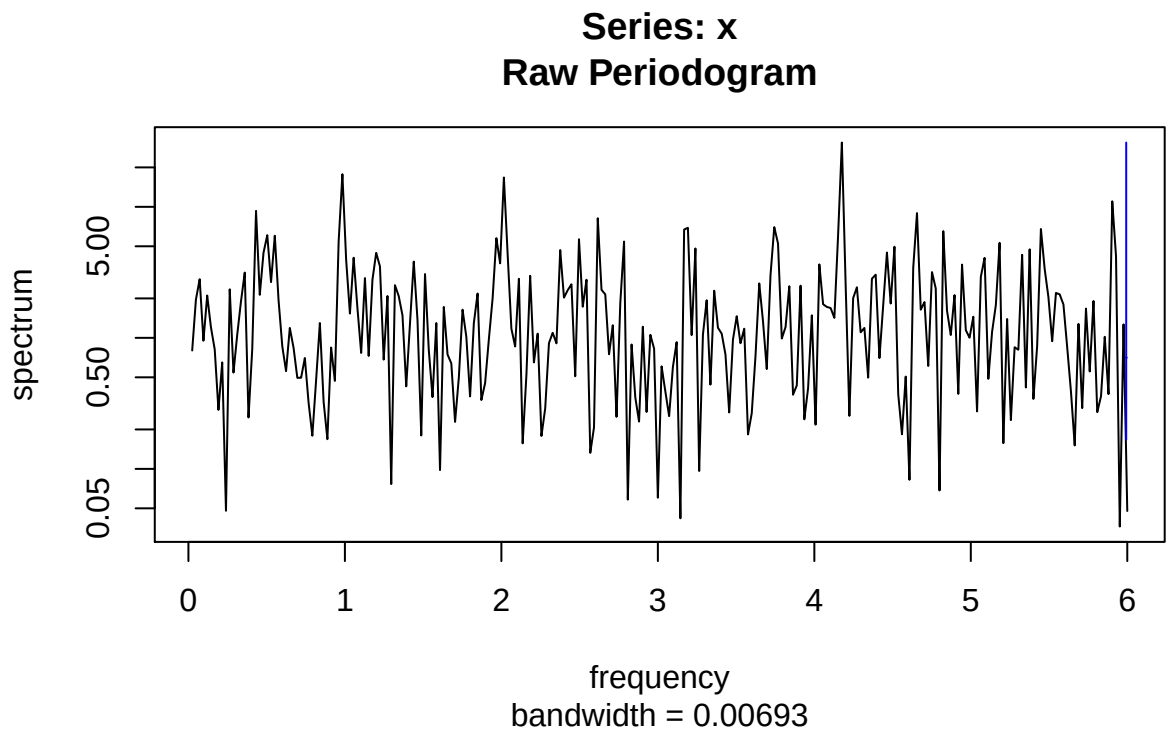


Figure 10: View of the residuals from the ARMA(2, 0) model

We can finally see how well we fit the residuals. Overall, we see no behavior that would lead us to invalidate our model and we therefore now look at its predictive power.

```r
plot(Z)
lines(fitted(Arima(Z, order = c(2, 0, 0))), col = "red")
```

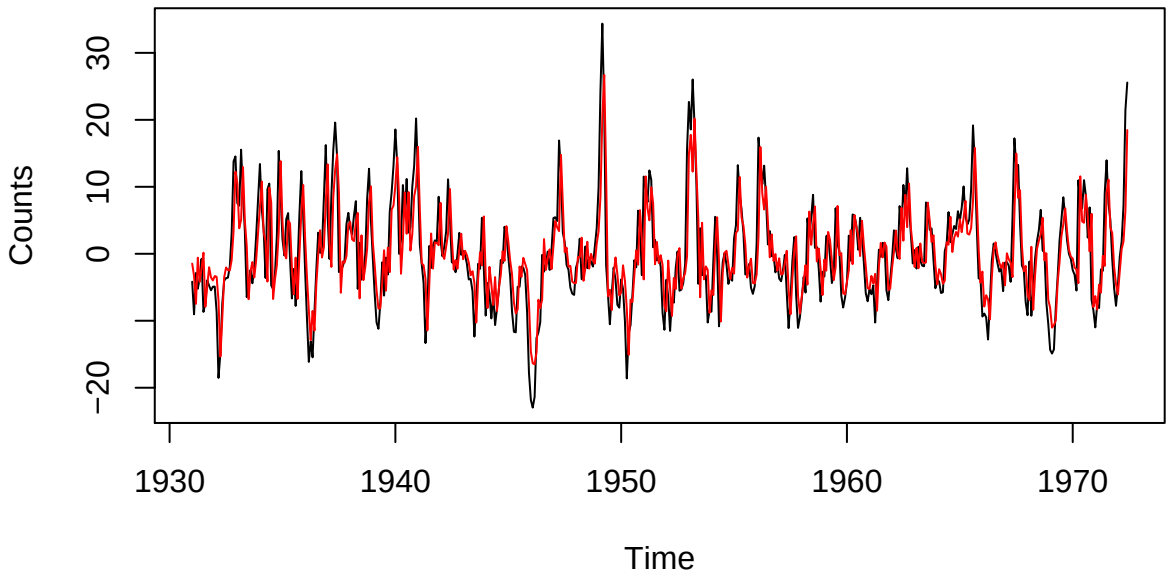


Figure 11: Fitted values versus actual values for X_t

4) Predicite power

1) Residuals: one-step ahead prediction

Starting at the 121^{th} month until the end, we fit the ARMA(2, 0) model on the past residuals and we predict the next value. We can then compare this fitted value to its actual realization.

```r
predict <- matrix(0, ncol = 3, nrow = 498 - 120)
for(i in 121:498){
  val <- forecast(Arima(Z[1:i], order = c(2,0,0)), 1)
  predict[i - 120,] <- c(val$lower[,2], val$mean, val$upper[,2])
}

predict_lo95 <- ts(predict[,1], frequency = 12, start = 1941)
predict_mean <- ts(predict[,2], frequency = 12, start = 1941)
predict_hi95 <- ts(predict[,3], frequency = 12, start = 1941)

plot(Z)
lines(predict_mean, col = "red")
lines(predict_lo95, col = "green")
lines(predict_hi95, col = "green")
```

2) Actual forecasting: prediction of the last 18 months

We re-train the whole model (f and ARMA(2, 0)) without the last year and we try to predict the last year to see how well we perform.

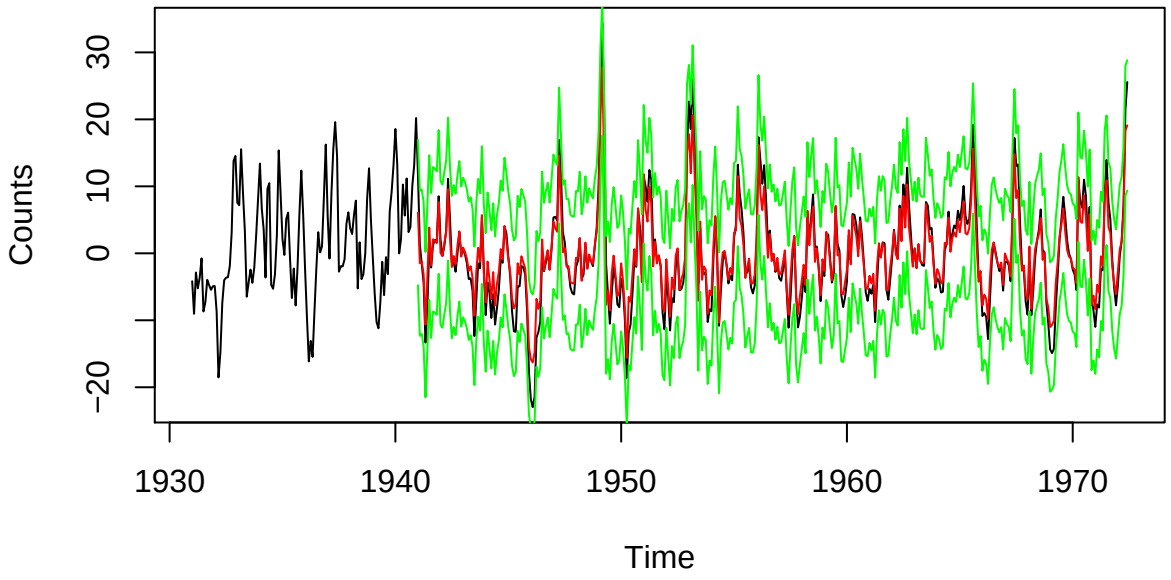


Figure 12: Actual realization (black) versus one-step ahead prediction (red) with a 95 confidence interval (green)

```
Z_train <- ts(Z[1:480], frequency = 12, start = 1931)
Z_test <- ts(Z[481:498], frequency = 12, start = 1971)
Z_predict <- forecast(Arima(Z_train, order = c(2,0,0)), 12)
plot(Z)
lines(ts(Z_predict$mean, frequency = 12, start = 1971), col ="red", lwd = 2)
lines(ts(Z_predict$lower[,2], frequency = 12, start = 1971), col ="green", lwd = 2)
lines(ts(Z_predict$upper[,2], frequency = 12, start = 1971), col ="green", lwd = 2)
```

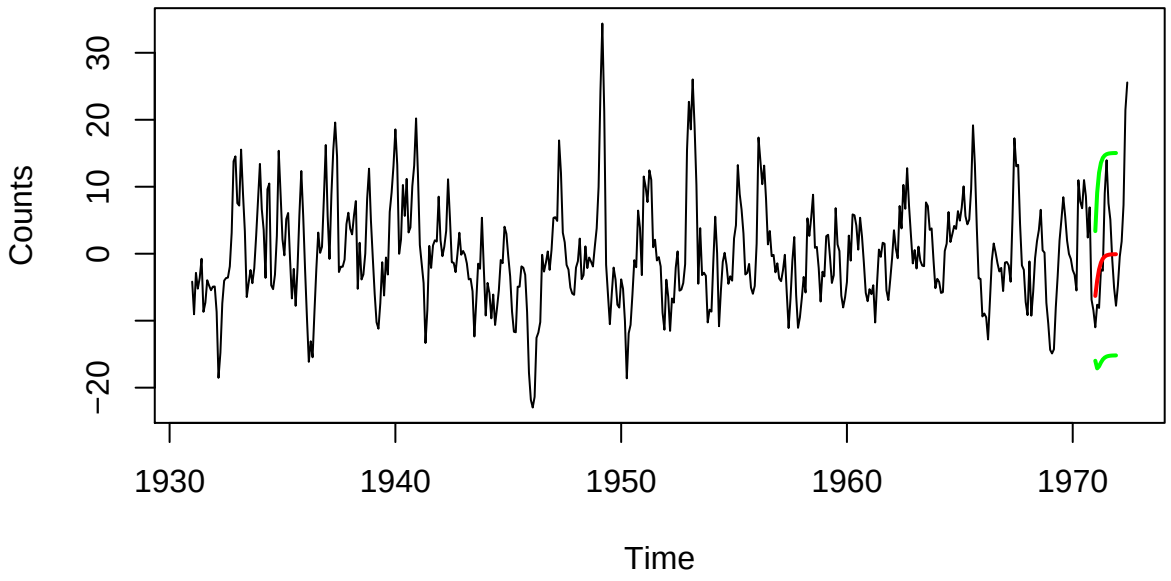


Figure 13: Actual realization (black) versus prediction (red) with a 95 confidence interval (green) over the last 18 months

5) Discussion

We clearly observe a seasonal trend in the data, with chickenpox counts being significantly higher in winter and the beginning of spring, with clear peaks between March and May, consistently with our observations. Moreover, we notice a strong downward trend starting in the mid 1950's. While chickenpox vaccine wasn't invented until much later, other safety measures (quarantines, better monitoring, keeping children out of school during epidemics) can also lower the number of infected individuals and where possiblby enforced by New York Public Health Department.

Chickenpox, furthermore, is a contagious disease, that is passed on by contact or by infected people's sneezes since the disease is airborne. Therefore, the model proposed by our ARMA model is sensible. On top of seasonal behavior, the number of occurrences is heavily linked to the number of occurrences over the previous months. Since symptoms take between 10 to 21 days to appear, it is logical that the number of occurrences is the previous months is positively linked to the occurrences in the next month. This is reflected in our ARMA model where $\phi_1 = 0.94$.

It is harder to explain why $\phi_2 = -0.23$. A fist comment is that $|\phi_2| < |\phi_1|$: the previous month has more impact than the one before that. However, the fact that $\phi_2 < 0$ might seem counter-intuitive. One possible explanation is that, since chickenpox is highly contagious, a high number of occurrences tow months in a row lead to much smaller number of people susceptible to catch the disease in the 3^{rd} month, hence the negative sign. This explanation is in no mean certain and deeper domain-knowledge would be necessary here to get a better understanding of the phenomenom at hand.

Our model is not able to predict very far into the future, as we can see from Figure 13. However, its accuracy over the span of the next months is quite good and could already have help to shape public policies.