Universität Zürich
Institute for Computational Science
Spring Semester 2021
ESC403

Professor Robert Feldmann (lectures)
Dr. Darren Reed (exercise classes)
Dimakopoulos Vasileios (TA)
Elia Cenci (TA)

INTRODUCTION TO DATA SCIENCE FINAL EXAM 2021

Solutions from David Linder

June 11, 2021

1 Time Series

1.1

To be considered stationary a time series should have the following properties:

- The mean $E[X_t]$ is the same for all times t
- The variance $Var[X_t]$ is the same for all times t
- The covariance between X_t and X_{t-1} is the same for all t, n

That means we want

- no obvious trends
- constant variance with time
- constant autocorrelation structure over time
- no periodic fluctuations (no seasonality)

I found some examples here: In figure 1 we see that most of the series are non-stationary except series (b) and (g). In series (g) there are cycles but they are not periodic.

1.2

- By looking at the data we can clearly see that this time series is not stationary. After log transformed the X-data we can see in figure **??** that although the standard deviation has a small variation the mean increases over time. Also the results of the Dickey-Fuller test is that we can not reject H_0 (TS is non-stationary) because the test statistic value is not less than the critical value. Here is the data in detail:

Test Statistic	−0.2312
p-value	0.9347
Lags	22.0000
Observations	975.0000
Critical Value (1%)	−3.4371
Critical Value (5%)	−2.8645
Critical Value (10%)	−2.5684

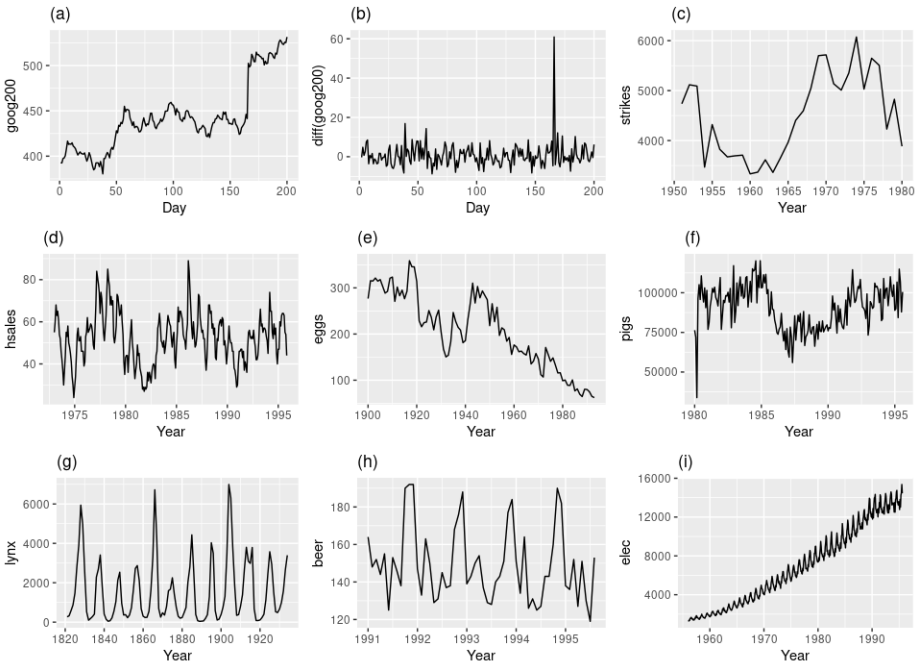

Figure 1: (a) Google stock price for 200 consecutive days; (b) Daily change in the Google stock price for 200 consecutive days; (c) Annual number of strikes in the US; (d) Monthly sales of new one-family houses sold in the US; (e) Annual price of a dozen eggs in the US (constant dollars); (f) Monthly total of pigs slaughtered in Victoria, Australia; (g) Annual total of lynx trapped in the McKenzie River district of north-west Canada; (h) Monthly Australian beer production; (i) Monthly Australian electricity production.

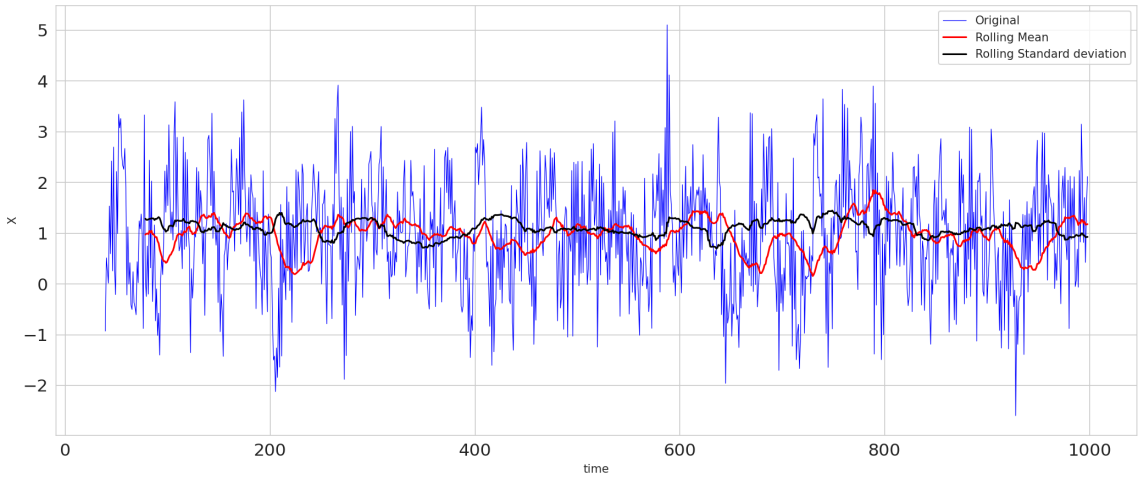

Figure 2: The result from smoothing.

I will use the moving average smoothing method which we used in the exercise class to eliminate the trend. First i calculated the rolling mean with the pandas function and then subtracted this from the original series (code: timeseries/1.2.py). Then i dropped the nan values that resulted from averaging over the time window (i tried out different sizes of that window and ended up with a value of 60). In figure 2 we see the result and that we eliminated the trend. Let's take a look at the test statistics:

Test Statistic	−10.7860
p-value	0.0000
Lags	1.0000
Observations	937.0000
Critical Value (1%)	−3.4373
Critical Value (5%)	−2.8646
Critical Value (10%)	−2.5684

The value is lower than all the critical values. Therefore we can say at least with 99% confidence this is now a stationary time-series.

- For answering this question we need to find the p-value. For this value to find we look at the plot of the partial autocorrelation function (PACF). From figure 3 we find a p-value of 3. That means that 3 values of at prior times are expected to directly affet a given current value of the time series.
-

1.3

-
-
-

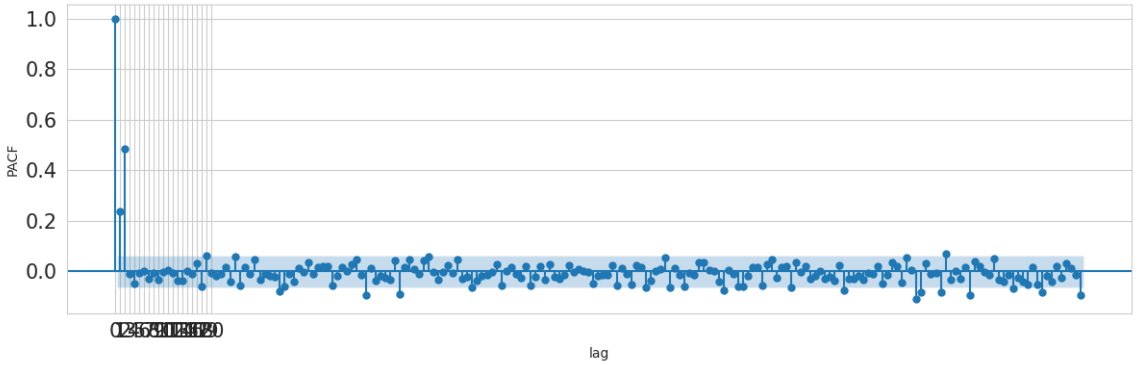

Figure 3: Plot of the partial autocorrelation function. The x-axis are the lags. One can see that the first time the PACF crosses the confidence interval (blue) is at lag 3.

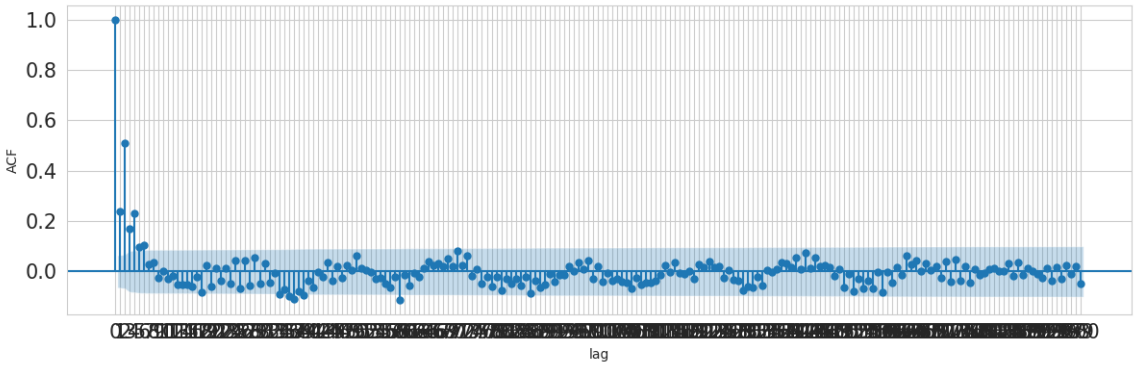

Figure 4: The autocorrelation function. The function crosses the confidence interval at lag 7 $\implies$ $q = 7$.

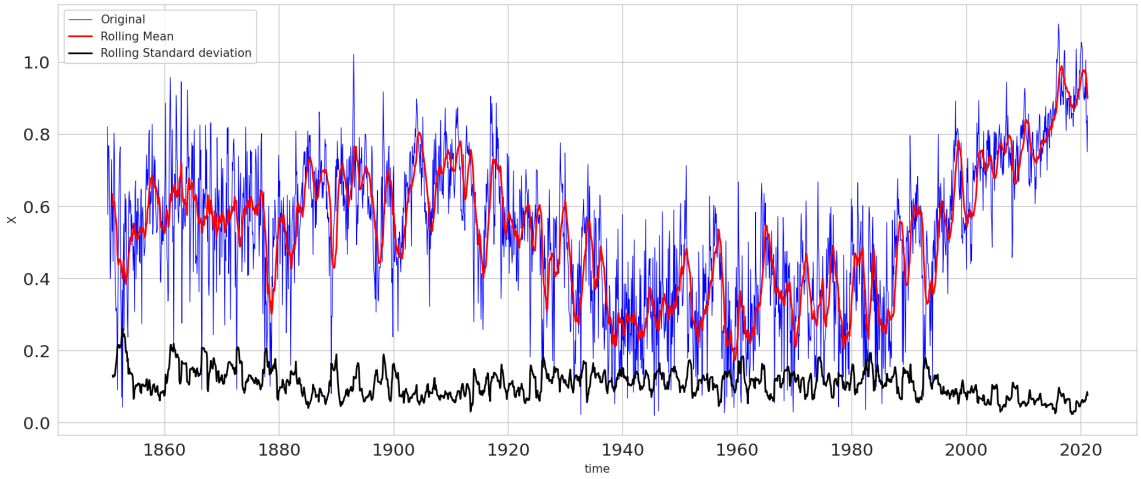

Figure 5: Average temperature anomaly raw data.

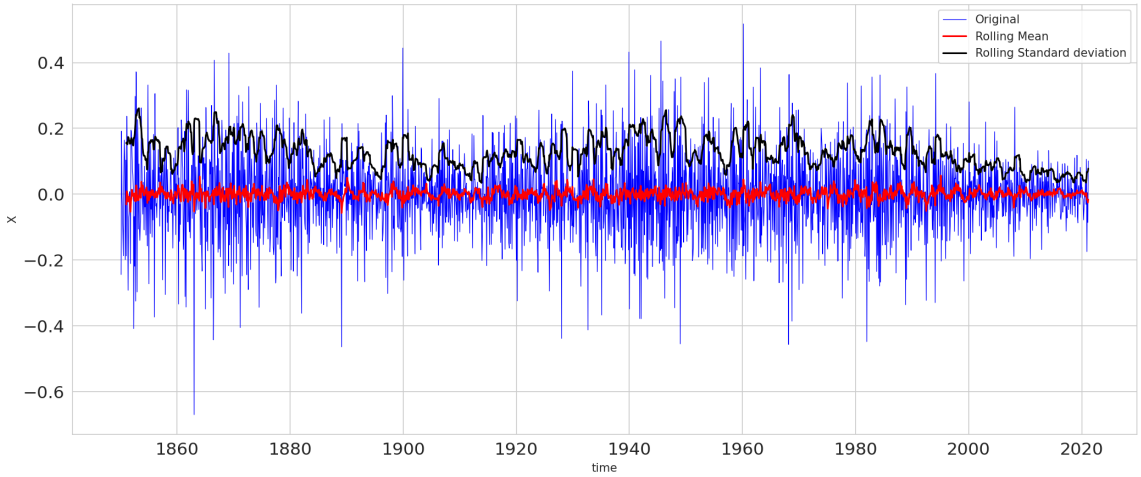

Figure 6: Average temperature anomaly after differencing.

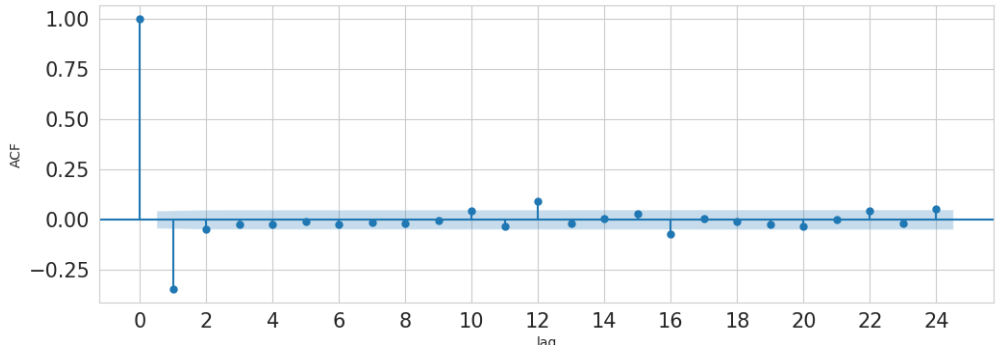

Figure 7: The autocorrelation function.

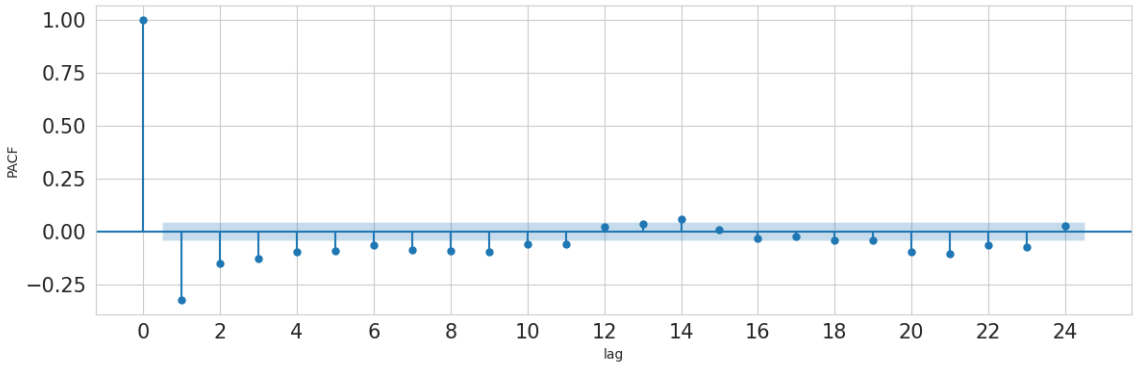

Figure 8: The partial autocorrelation function.

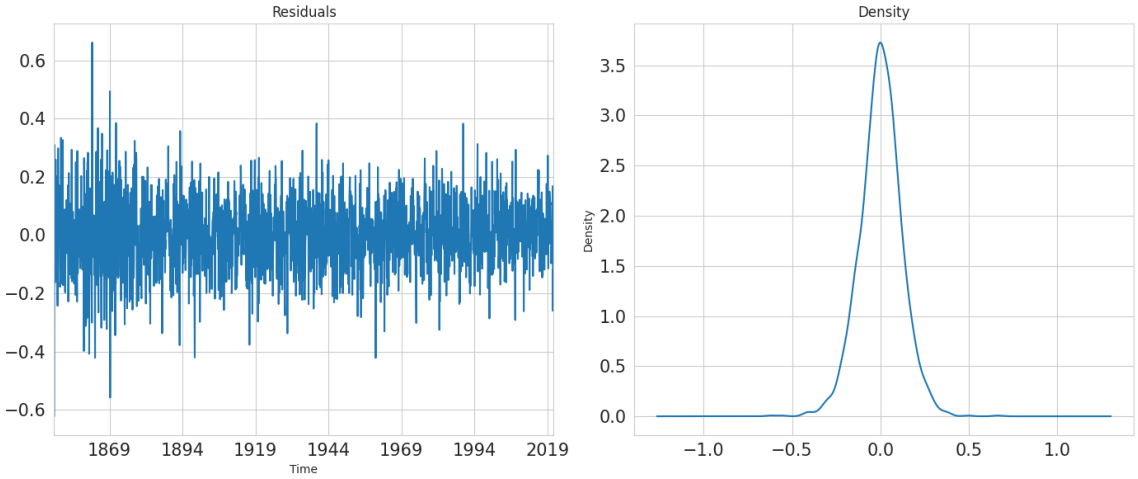

Figure 9: Residuals and density.

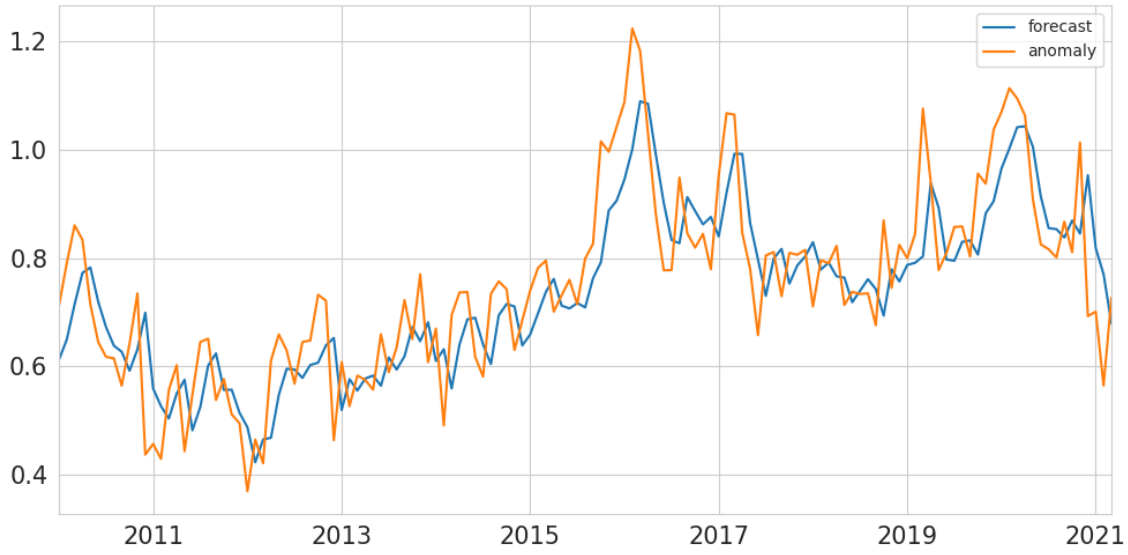

Figure 10: Forecast 2010 - 2021

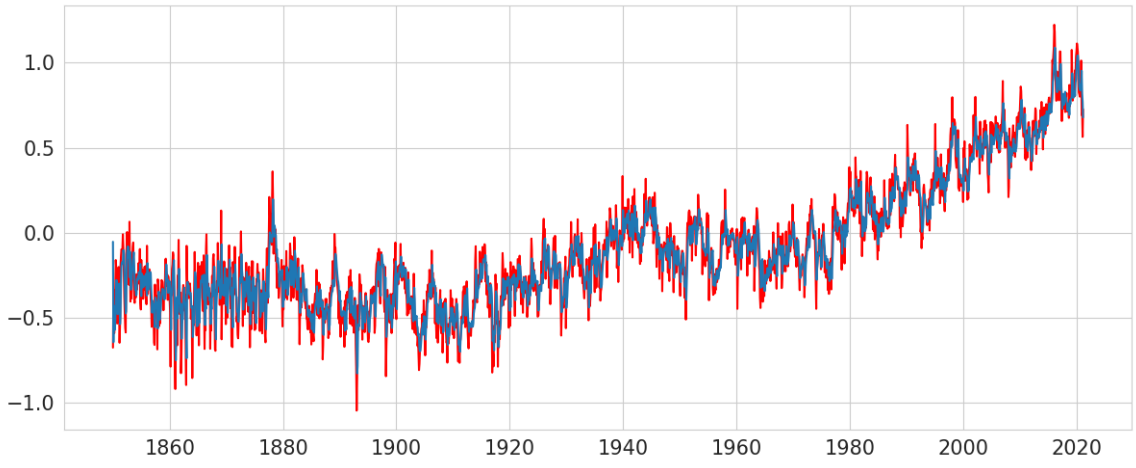

Figure 11: Forecast 2010 - 2021 RMSE

2 Image classification

2.1 Image classification

If better means higher accuracy then i would choose a CNN for image classification. RF performs well on categorical data while a CNN handles numerical input very well. Here we have pixels with numerical values between 0 and 255. We can also tune more parameters in a CNN like kind/number of layers, epochs and learning rate. Also different activation functions for the neurons can be chosen. Finally a CNN learns how to apply a filter to an input during training such that certain features in the image can be recognized. A RF can not take advantage of such structures in images.

On the other hand a CNN needs a lot of data to perform well. I don't know exactly if our dataset is large enough but if i compare with other models they used hundred thousands or even millions of images to train a CNN.

2.2 Classification performance

- Accuracy CNN: 99.00%, that means the error rate was 1%. It misclassified 10/1000 images from my test set.
- Accuracy RF: 90.5%, that means the error rate was 0.5%. It misclassified 5/1000 images from my test set.
- Both models performed surprisingly good. I checked many times if i excluded any test data from the training. No model performed significantly better.
- Both models agreed in all images only the CNN misclassified image 001212.jpg where it predicted the class 'headCT' instead of 'Hand'. As far as i can tell the RF made no prediction error on the unlabeled dataset. I included the corresponding files in the folder `report/data` in the repository (`predictions_cnn.csv`, `predictions_rf.csv`).

2.3 Hyper parameter tuning

2.4 Interpretation of model performance

2.5 Bonus Question

Figure 12: The prediction from the convolutional neural network. It misclassified one image. Can you find it?