Madrid weather prediction project

Project of forecasting future max./min. temperature of Madrid Aeropuerto weather station, based on historical data.

By: Juan Vicente Peluso

GitHub notebook

Table of contents

1. **Introduction**
   - The dataset, features and target value
   - Problem description
2. **Data load, formatting and inspection**
3. **Descriptive statistics**
   - General overview
   - Data visualization
4. **Exploratory data analysis**
5. **Model development**
   - Data split and SARIMA order lookup
   - Models fit, forecast and scoring
   - Future predictions
6. **Conclusions and final thoughts**

1. Introduction

The dataset, features and target value

The data is retrieved from the ***AEMET*** site, ***AEMET*** (Agencia Estatal de Metereología) is the Spanish official weather agency, which offers some of the data they gather for public use (OpenData).

AEMet
Agencia Estatal de Meteorología

We will try to forecast the **minimum** and **maximum** mean temperatures for the MADRID AEROPUERTO weather station in Madrid, Spain.

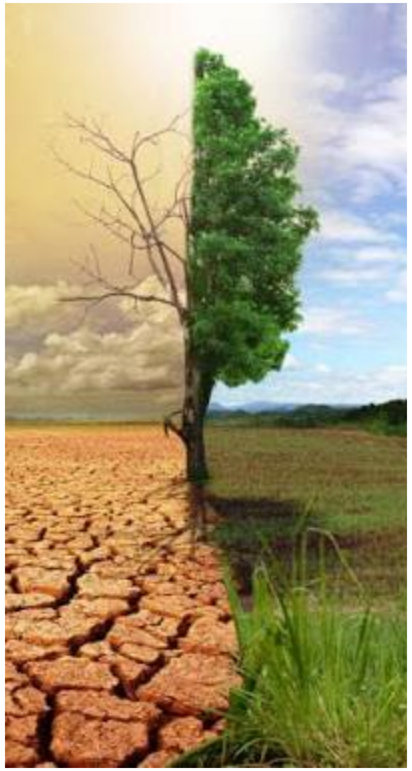
Encinar de los Reyes
LAS TABLAS
VALDEFUENTES
Aeropuerto de Madrid-Barajas Adolfo...
Paracuellos de Jarama
A-1
M-50
M-13
M-111
M-11
E-5
Los Berrocales del Jarama
HORTALEZA
M-11
BARAJAS
Agencia Estatal de Meteorología
M-12
M-14
A-2
ALAMEDA DE OSUNA

Problem description

As we all know, climate change is a huge problem we all have been suffering from since the last decade. The temperature is constantly rising, which creates abnormal changes that affect the ecosystem. Consequently, these changes impact the stations; longer and warmer periods and shorter but inconsistent cold periods.

To name a few of the consequences of this change, we can see how some species are on the verge of extinction, others, are forced to migrate elsewhere, breaking the balance of the local fauna. Plants experience a change in their life cycle, causing poor harvests, or even completely lost due to an extremely cold late winter, or an extraordinarily hot early summer. Not to mention the increase in weather phenomena that we've experienced over the last 20 years.

Being the Madrid the city where I currently live since 2009, I can tell the summers are warmer than previous, but mainly, the winters aren't colder as they used to be. The objective is to **forecast** he minimum and maximum temperatures from the MADRID AEROPUERTO station, to see if the increasing trend holds for the coming years.



2. Data load, formatting and inspection.

We access the *AEMET* data via **API**, and request the *daily temperature records* of the period 1990 - 2019 from the *MADRID AEROPUERTO* station. The information is retrieved in JSON format, loaded into a Pandas dataframe and filtered to have finally the columns we're interested in:

- ***fecha***: Date of the measure, index of the dataframe.
- ***tmin***: Minimum temperature registered that day.
- ***tmax***: Maximum temperature registered that day.

As is pointless to forecast daily temperatures, the dataframe was resampled to monthly periodicity using *average* as the aggregation method. Then, the data was split into 2 series (minimum and maximum) to work separately. No NaN or empty values were found running the integrity checks.

3. Descriptive statistics

General statistics

The datasets statistics tell us, the STD of the maximum temperatures is higher than the minimum one, which indicates the temperature fluctuations are more prominent.

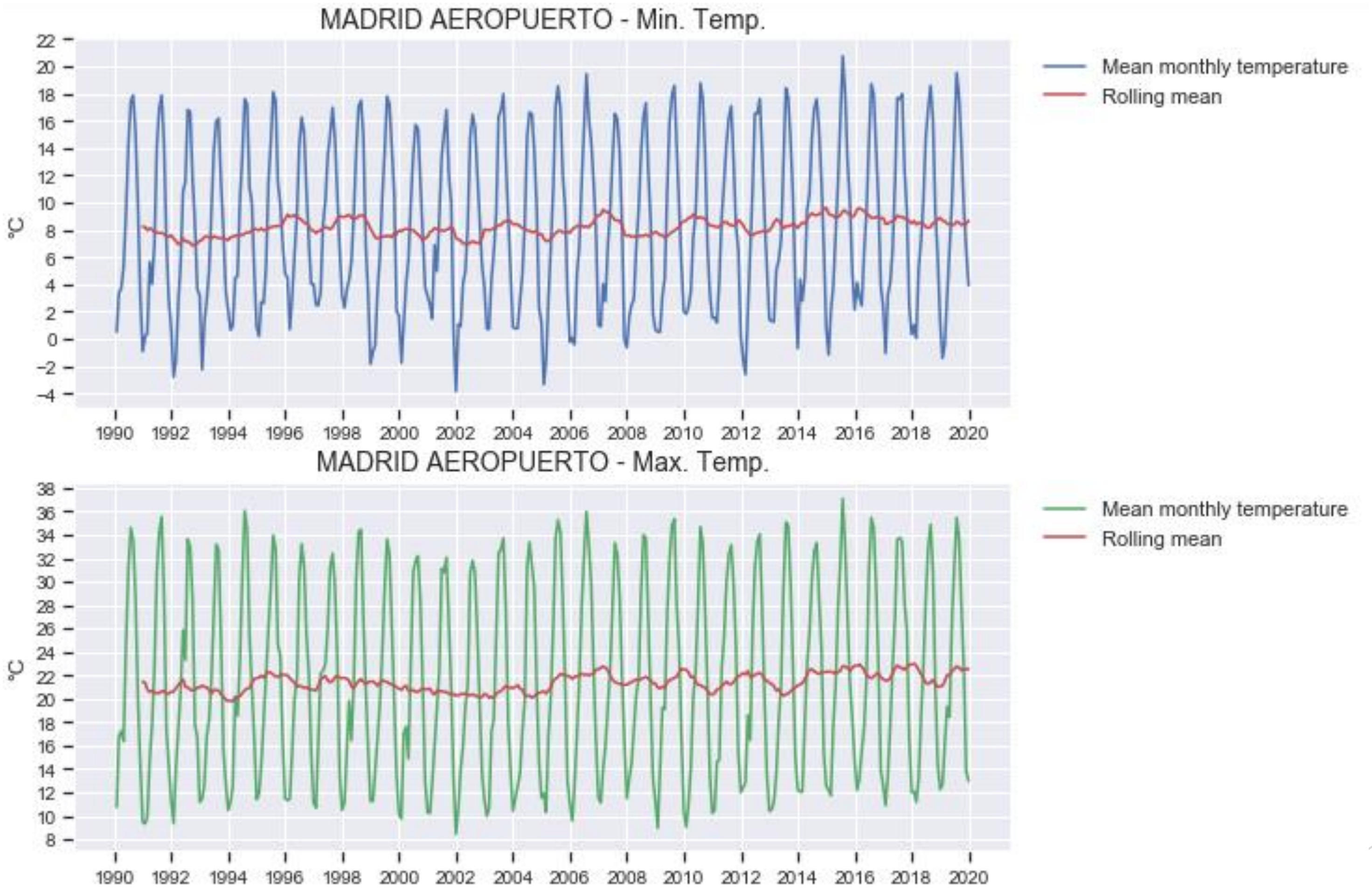
When we see the higher and lower record of daily, monthly and, yearly temperatures we note the following:

- All coldest temperatures happened all before 2001.
- The warmest month and year occur in 2015, except the hottest day ever recorded, which happened in July 1995.
- The warmest minimum temperatures all happened in the 2010s decade.
- The coldest maximum temperatures for month and year occur before 2001, except the coldest maximum temperature ever recorded, which happened in January 2010.

All of this statistics can be seen in the notebook.

Data visualization

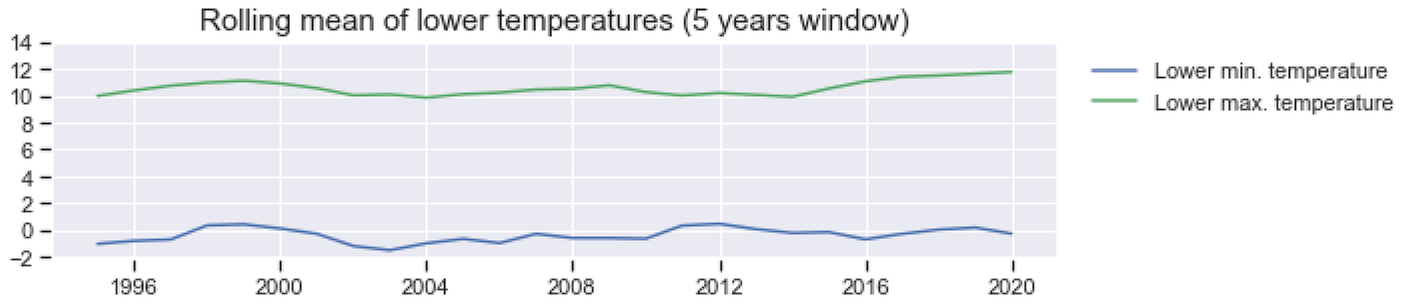
The plots of the series, we see a *seasonal time series* in both graphs, with no clear trend, neither the actual data or the rolling mean, calculated on a 12 months window.

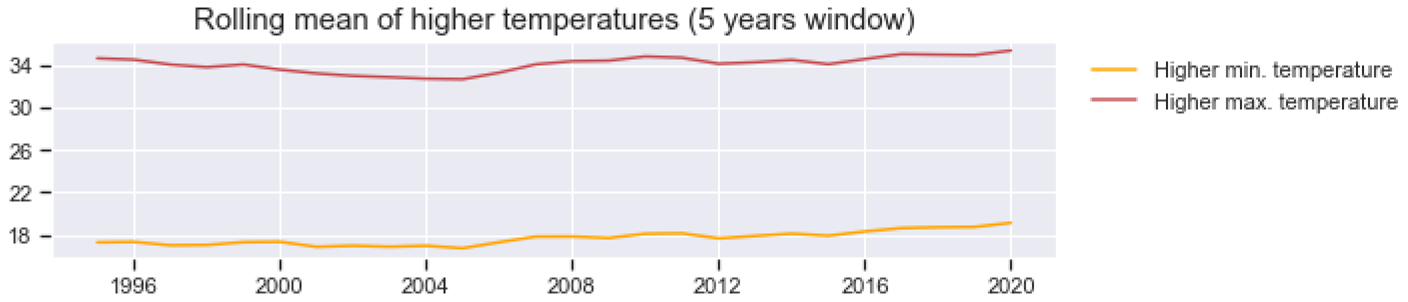


MADRID AEROPUERTO - Min. Temp.

MADRID AEROPUERTO - Max. Temp.

But when we calculate the year rolling mean for the minimum and maximum data, with a window of 5 years, we see that both higher temperatures and the lowest maximum average temperature have a clear rising trend, except the lower maximum temperature, which confirms the fact that temperatures are rising.

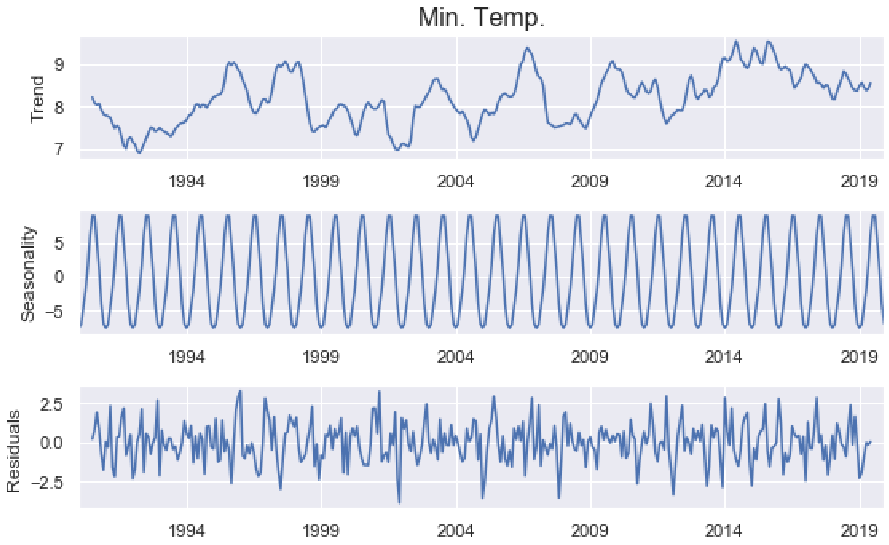


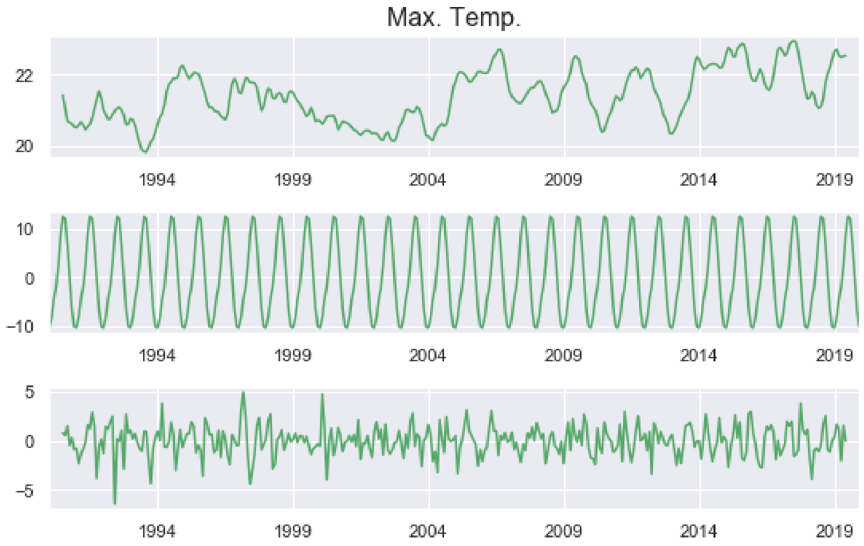


4. Exploratory data analysis

To begin with, we must verify if the time series are stationary, we've applied the Augmented Dickey-Fuller test in both series, and we can affirm with ***99% confidence*** that they ARE STATIONARY, both scored less than 0.01 p-value. The results of the test can be seen in the notebook.

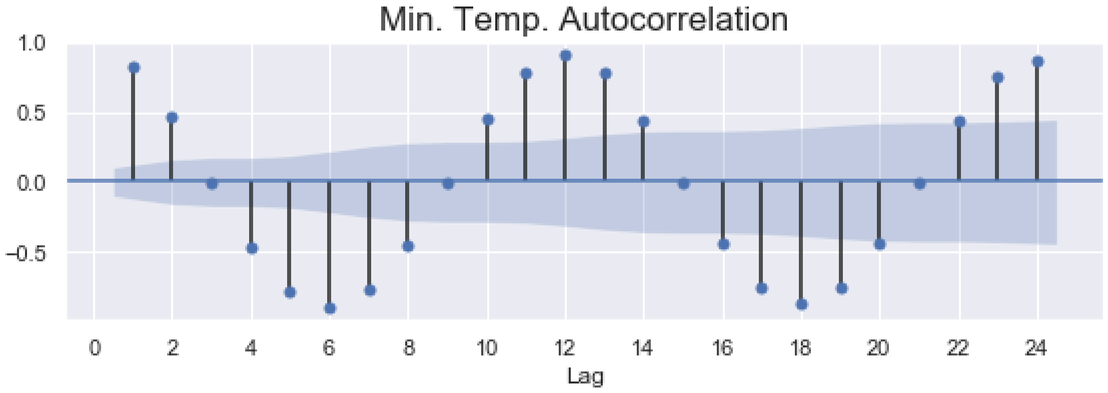
Being a seasonal time series, in the *statsmodel* package we can decompose the series in it's 3 main parts: *Trend*, *Seasonality*, and *Residuals*, to ensure our assumptions are correct.

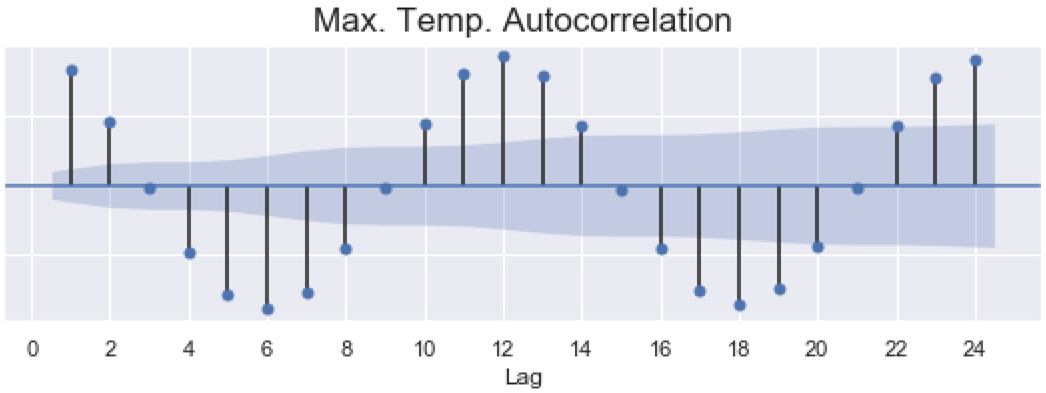




- In both series, there's no clear trend.
- In both series, as we've seen before, there's a clear seasonality.
- In both series, the residuals follow the *white Gaussian noise* pattern, as we cannot see any obvious structure.

Although we've already seen that the seasonality is annual, with the **AFC** plot we confirm that both time series follow a twelve-lag pattern.





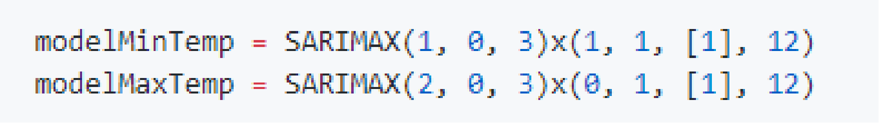
5. Model development

Data split and SARIMA order lookup

First, the data for both datasets were split in train and test, using the data from the 1990-2014 period for training, leaving the remaining data (2015-2019) to validate the model.

For this time series task, we've used a **SARIMA** (Seasonal Autoregressive Integrated Moving Average) model. To find the order for seasonal and non-seasonal autoregressive, integrating difference, and moving average algorithms, which compose the ARIMA model, we've used the *auto_arima* function of the python package *pmdarima*, which loops through different combinations of model orders, and returns the best-scored model. In this case, the **AIC** (Akaike information criterion) score was selected as the metric.

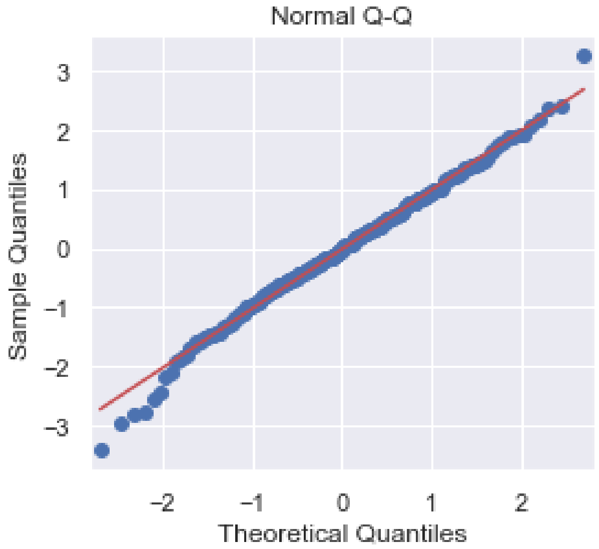
Indicating that the seasonal cycle is 12, we've run the lookup process and the best set of orders was the following:



```
modelMinTemp = SARIMAX(1, 0, 3)x(1, 1, [1], 12)
modelMaxTemp = SARIMAX(2, 0, 3)x(0, 1, [1], 12)
```

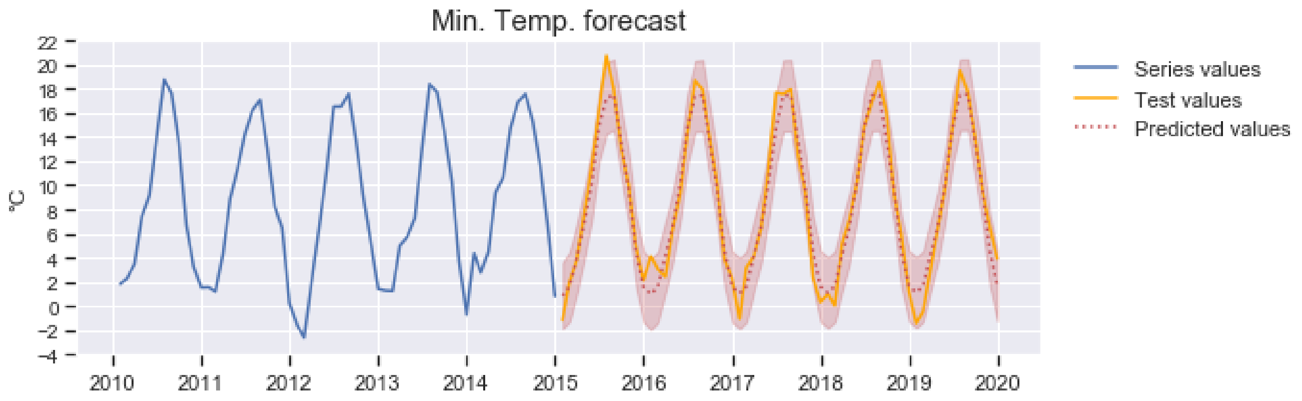
Both models residuals are NOT normally distributed (Jarque-Vera p-value 0.07/0.37 respectively) and NOT auto-correlated (Ljung-Box p-value 1.00/0.21 respectively).

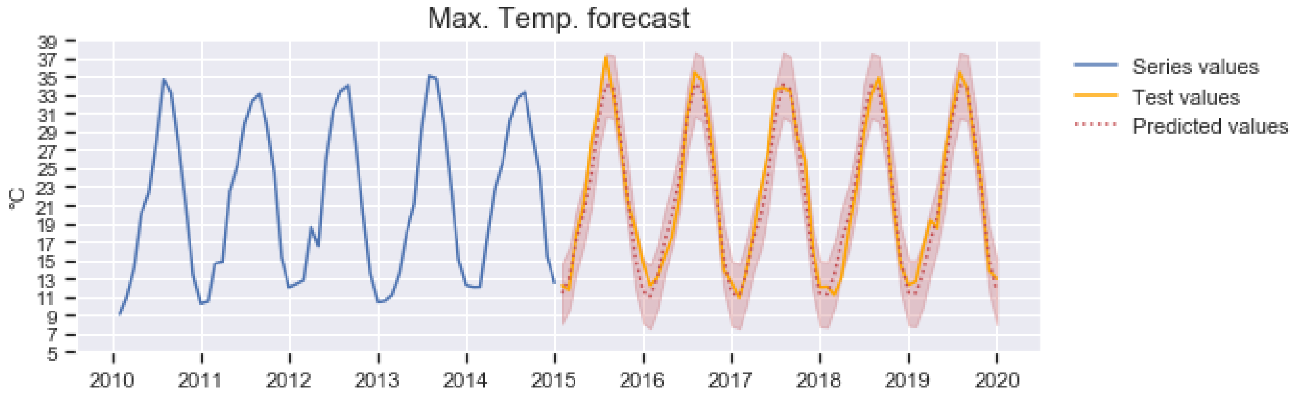
It is worth mention, that in the diagnostic plots of the model for the minimum temperatures, the residuals, KDE histogram, and correlogram look correct, except the **Q-Q plot**, which at the lower end, has some points that fall outside the line. Here we can see the plot, other plots can be seen in the notebook.



Models fit, forecast and scoring

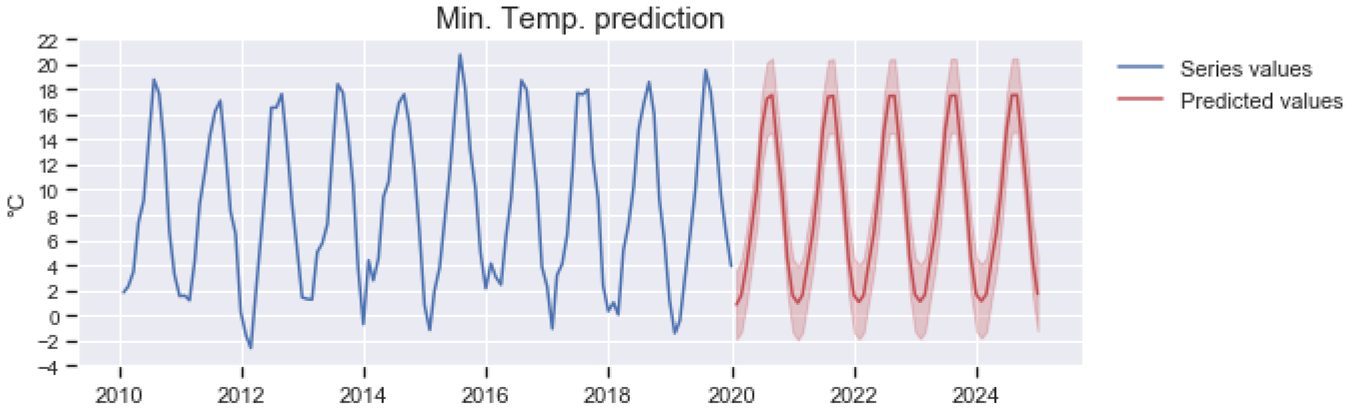
We've fit both models to the training data and forecasted the values for the 2015-2019 period. The metric to measure the quality of the models was **MAE** (Mean average error), and the results for both models were **0,9972** for the minimum model, and **1,4293** for the maximum model. When we've plotted the forecast against the real data, we see that the predictions are reasonably accurate and that almost all real values fall inside the 95% confidence interval range.

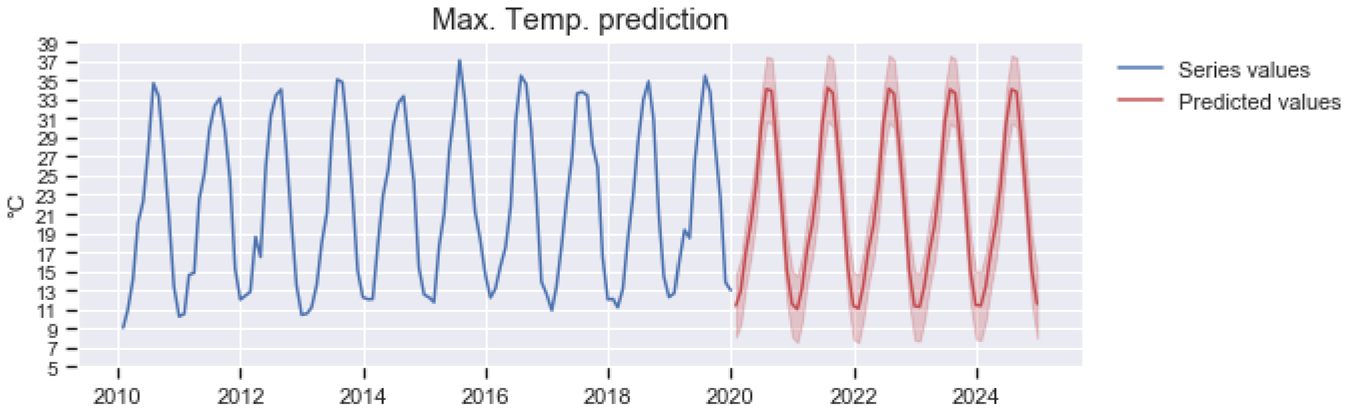




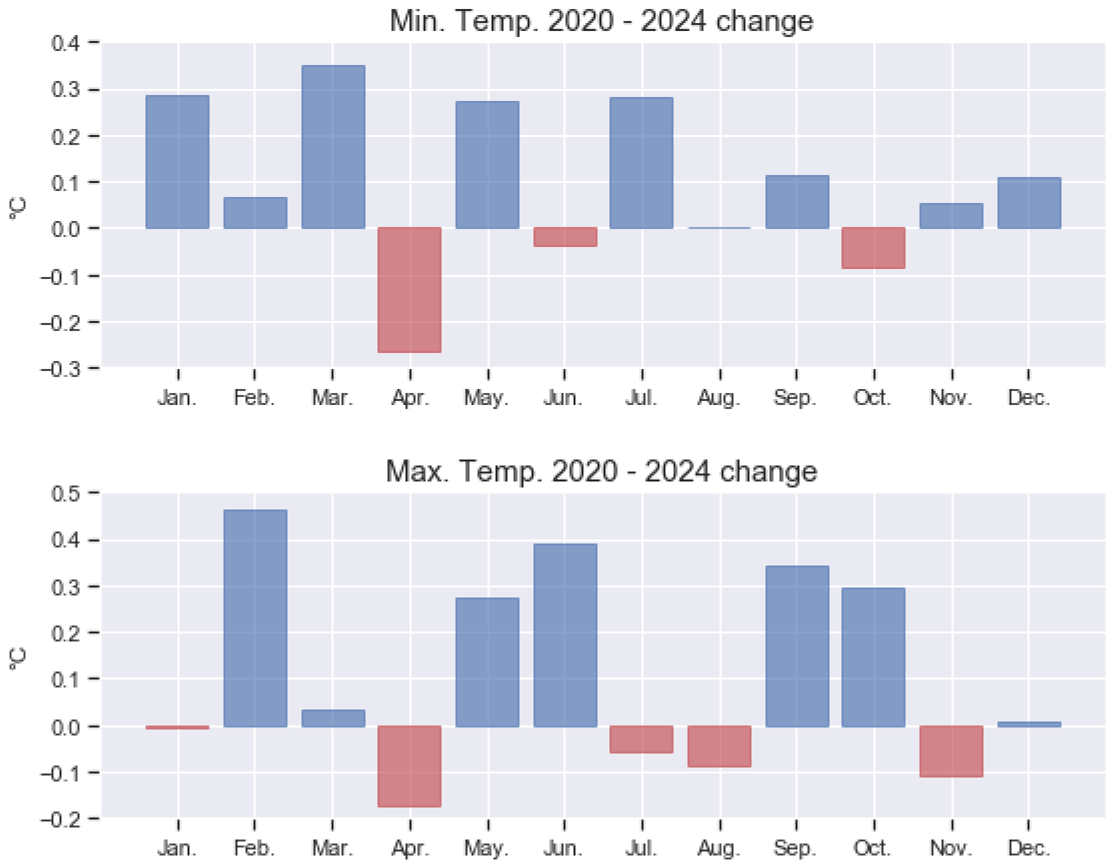
Final predictions

Finally, we've made predictions for the 2020-2024 period (60 months), and although there are no clear indications that the trend holds when we see the plot.


Min. Temp. prediction


Max. Temp. prediction

But if we calculate the difference per month, of the last-year prediction minus the first one, we see that the trend holds.



Having seen the bar plot, we draw the following conclusions:

- 9 out of 12 months minimum temperature rise, only April, June and October descend.
- January, March and July show an increase of 0.3°C for the minimum temperature over five years.
- April shows a drastic fall of almost 0.3°C which, along with the March rise, indicates a warm winter end and a cold spring beginning.
- 7 out of 12 months maximum temperature rise. Oddly enough, the warmest months of the year (July and August) decreases.
- February and June show an increase of 0.4°C of the maximum temperature over five years, a concerning fact since February (0.46°C rise) is one of the coldest months of the year.
- In addition to the above mentioned July and August, April and November minimum temperatures also decrease being most noticeable again April, with almost 0.2°C loss.
- If we put together that, May and June rise, July and August decrease and September and October rise again, we will have summer high temperatures for almost 6 months!

6. Conclusions and final thoughts

Data has spoken, is a fact that temperatures are not only increasing at a blinding pace, but the four seasons as we know them will practically disappear. Having warmer winters is an ecological catastrophe, not to mention hottest and longer summers will progressively trigger energy consumption, creating a vicious circle, since energy production (non-renewable, predominant in the world) generates more and more heat, thus increasing the total computation over and over again.

There's a fact that's disturbing, February and March raises and April decreases sharply, this means that the period where trees have the processes of flowering and pollination will start earlier, and by the time the fruits start to grow (April approximately), lower temperatures will interrupt the successful growth of most the harvest.

We could continue enumerating consequences; economic, social, demographic, etc. The point is, the temperature is rising and we, as a society are not doing all we could to reverse the trend.
Think, that with this pace, Madrid in 30 years would have average temperatures like **Marrakesh**! If the trend is the same in the poles, cities like Amsterdam or Venice will practically be drowned, as the poles' ice will melt, and the sea level will rise over a meter!

I as an individual can only ask you to do your part, may our actions allow our children to have a brighter future.