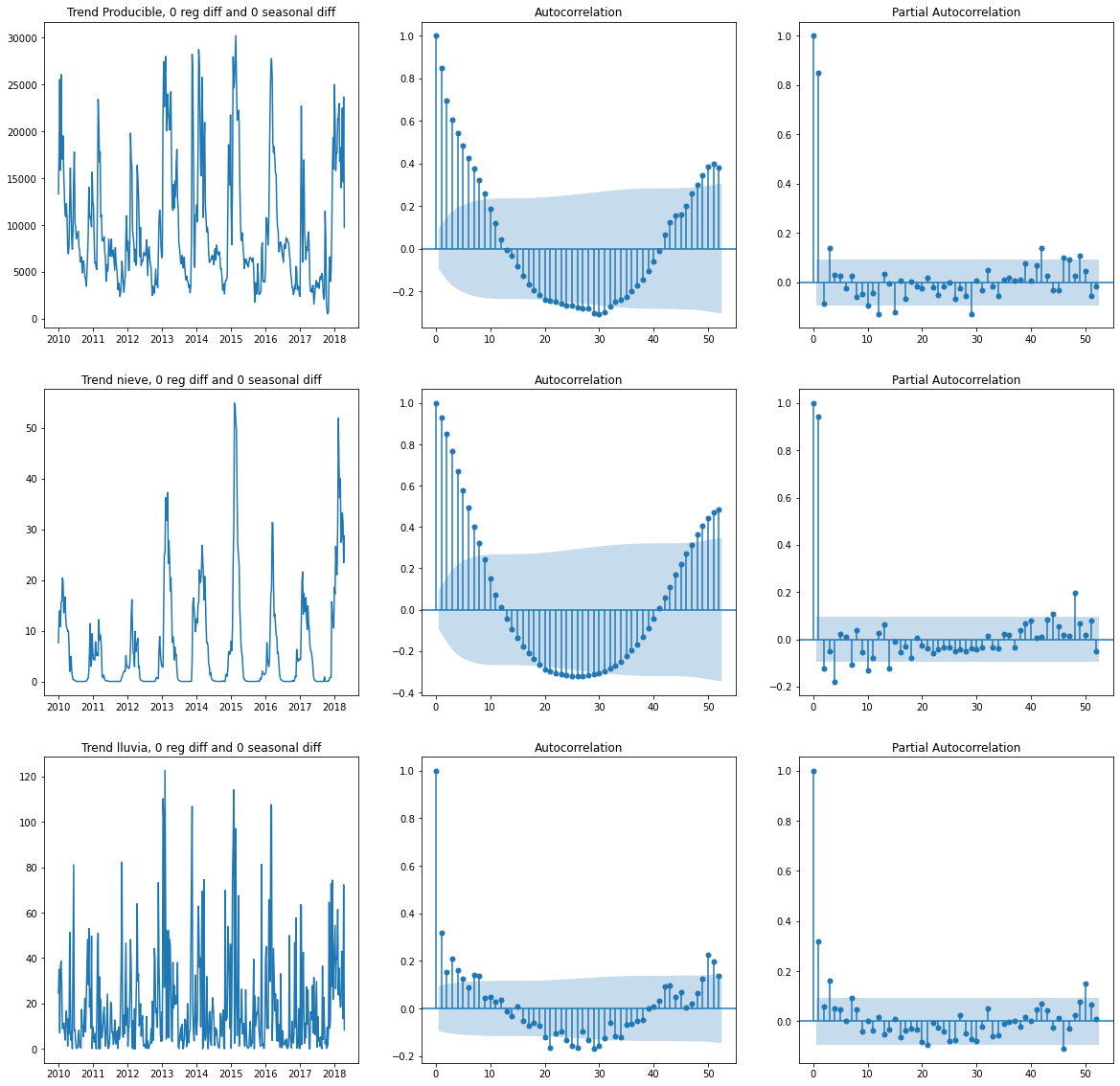
1. Find the best time series model, SARIMA type, for Y (streamflow).

The 3 variables in the dataset had an ADF test p-value lower than 0,05 meaning they are stationary. As you can see in the graph below, the 3 columns have a stationary mean and a high variance. Although in the charts the features appear to have seasonality, the seasonal trend is not constant, meaning not every 52 weeks or any specific interval of time the trend appears.

1 Trend, ACF and PACF of variables



Looking at the PACF chart of our target variable (Producible) has 2 lags out of bounds (1 and 3). Although it has lags out of bound around lag 45, they are not taken into account.

The first model to tried an AR(3).

2 Residual of AR(3) (Box Test: 0.251, Squared Box Test:0.002, Shapiro:0.000)

Interfaz de usuario gráfica

Descripción generada automáticamente

Looking at the chart and the respective test p-values the residuals of the model are white noise, so the model can not be improved in a linear way. The box test for the squared residuals shows autocorrelation, but looking at the ACF and PACF the explainability is low, so no model for the variance will be implemented.

1. Find the best transfer function model (regression model with time series errors) to explain Y as a linear function of X1 and X2, and the past of Y.

The second model to try was a regression model with formula y = 5175 + Nieve\*442 + Lluvia\*90, (all variables were relevant).

3. Residuals of Regression

Gráfico, Histograma

Descripción generada automáticamente

The residuals of the model were stationary but had autocorrelation, so an Arima model to consider this information was applied, The AR(1) was the best fit and after applied the residual were white noise.

4. Residuals of AR(1)

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Finally, both models were putted together and all the variables were found relevant.

1. Find the best random forest model to explain Y as a function of X1, X2 and the past of Y.

The first random model was fitted without taking in to account the past of Producible. For the second model the past of y was included by adding 3 columns with the last 3 lags of Producible. A Random Forest model was trained, leaving the last 100 periods for testing

1. Perform a back-testing out of sample with the best models found in the previous 3 points. To do this, leave the last 100 weekly real values to perform this out of sample exercise. Perform a one-step-ahead forecasting exercise.

To perform the back-testing the last 100 weeks were left for the test set. The 2 random forest were tested for 100 steps and the models with Arima were tested for 1, 4 and 100 steps in expand way to train with maximum amount of information. 2 metrics were used for evaluation MAPE and RMSE

5. Evaluation of Models

Tabla

Descripción generada automáticamente

The best models that included Arima were in the 1 step forecasting, because the uncertainty was lesser. The best model in RMSE was the random forest with lags and for the MAPE the best model was the Arima. In this case, because the values of the target variable have a high volatility (sometimes producible is too low that MAPE can go above 1000%) RMSE seems like a better metric.

6 Back Testing (one step ahead)

Gráfico, Histograma

Descripción generada automáticamente