Report

Lucas Massaroppe

06/28/2019

B2W Challenge

In this challenge two databases were considered: (a) 'sales.csv' and (b) 'comp_prices.csv'. In (a) there is transactional information of nine products $P_i, i = 1,\ cdots, 9$ for the year of 2015 and in (b), contains data from six competitors $C_i, i = 1, \cdots, 6$, monitored at different times, for the same dates and products of the data set.

At first, we will analyze 'sales.csv' and, in a second step, we will pay attention to 'comp_prices.csv', so that we can thus make a comparison between the transactional information and the prices of the competitors.

The 'sales.csv' data set

As a quick check the time series of the nine products are depicted in Figure 1, in such a way that we can perform the KPSS test on each to access the trend-stationarity null hypothesis of them.

The below table shows the KPSS test results.

Product ID	Test Statistic	p-value
P_1	0,14	0,07
P_2	0,17	0,02
P_3	0,53	$< 0{,}01$
P_4	0,04	$> 0{,}10$
P_5	0,26	$< 0{,}01$
P_6	0,08	$> 0{,}10$
P_7	0,42	$< 0{,}01$
P_8	0,09	$> 0{,}10$
P_9	0,06	$> 0{,}10$

Note that, according to the null hypothesis of the KPSS test and using a significance level of $\alpha = 1\%$, products with $p < \alpha$ is stationary, that is, P_3, P_5 and P_7.

Product revenue 30 day forecast

For all time series we use three prediction models: the linear autoregressive moving average (ARIMA(p, d, q)), the nonlinear exponential smoothing (ETS) and the autoregressive neural network (NNETAR(p, k)). To evaluate the best model among them, we used the standard deviation of the residuals as a mesure of comparison and we show in the following table.

Produto	ARIMA(p, d, q)	ETS	NNETAR(p, k)
P_1	ARIMA$(1, 1, 1)$, $\widehat{\sigma_e} = 66{,}87$	ETS(M, N, N), $\widehat{\sigma_e} = 0{,}04$	NNETAR$(4, 2)$, $\widehat{\sigma_e} = 0{,}04$
P_2	ARIMA$(0, 1, 2)$, $\widehat{\sigma_e} = 32{,}21$	ETS(A, N, N), $\widehat{\sigma_e} = 33{,}56$	NNETAR$(7, 4)$, $\widehat{\sigma_e} = 0{,}04$
P_3	ARIMA$(2, 1, 3)$, $\widehat{\sigma_e} = 71{,}28$	ETS(A, N, N), $\widehat{\sigma_e} = 72{,}13$	NNETAR$(7, 4)$, $\widehat{\sigma_e} = 0{,}05$
P_4	ARIMA$(0, 0, 0)$, $\widehat{\sigma_e} = 436{,}59$	ETS(A, N, N), $\widehat{\sigma_e} = 437{,}72$	NNETAR$(1, 1)$, $\widehat{\sigma_e} = 0{,}22$
P_5	ARIMA$(0, 1, 1)$, $\widehat{\sigma_e} = 109{,}04$	ETS(M, N, N), $\widehat{\sigma_e} = 0{,}11$	NNETAR$(4, 2)$, $\widehat{\sigma_e} = 0{,}08$
P_6	ARIMA$(5, 1, 0)$, $\widehat{\sigma_e} = 231{,}71$	ETS(M, A_d, N), $\widehat{\sigma_e} = 0{,}12$	NNETAR$(6, 4)$, $\widehat{\sigma_e} = 0{,}05$
P_7	ARIMA$(0, 1, 1)$, $\widehat{\sigma_e} = 34{,}35$	ETS(A, N, N), $\widehat{\sigma_e} = 34{,}37$	NNETAR$(8, 4)$, $\widehat{\sigma_e} = 0{,}03$
P_8	ARIMA$(2, 1, 1)$, $\widehat{\sigma_e} = 29{,}99$	ETS(M, N, N), $\widehat{\sigma_e} = 0{,}05$	NNETAR$(2, 2)$, $\widehat{\sigma_e} = 0{,}03$
P_9	ARIMA$(1, 1, 1)$, $\widehat{\sigma_e} = 39{,}47$	ETS(M, N, N), $\widehat{\sigma_e} = 0{,}08$	NNETAR$(6, 4)$, $\widehat{\sigma_e} = 0{,}06$

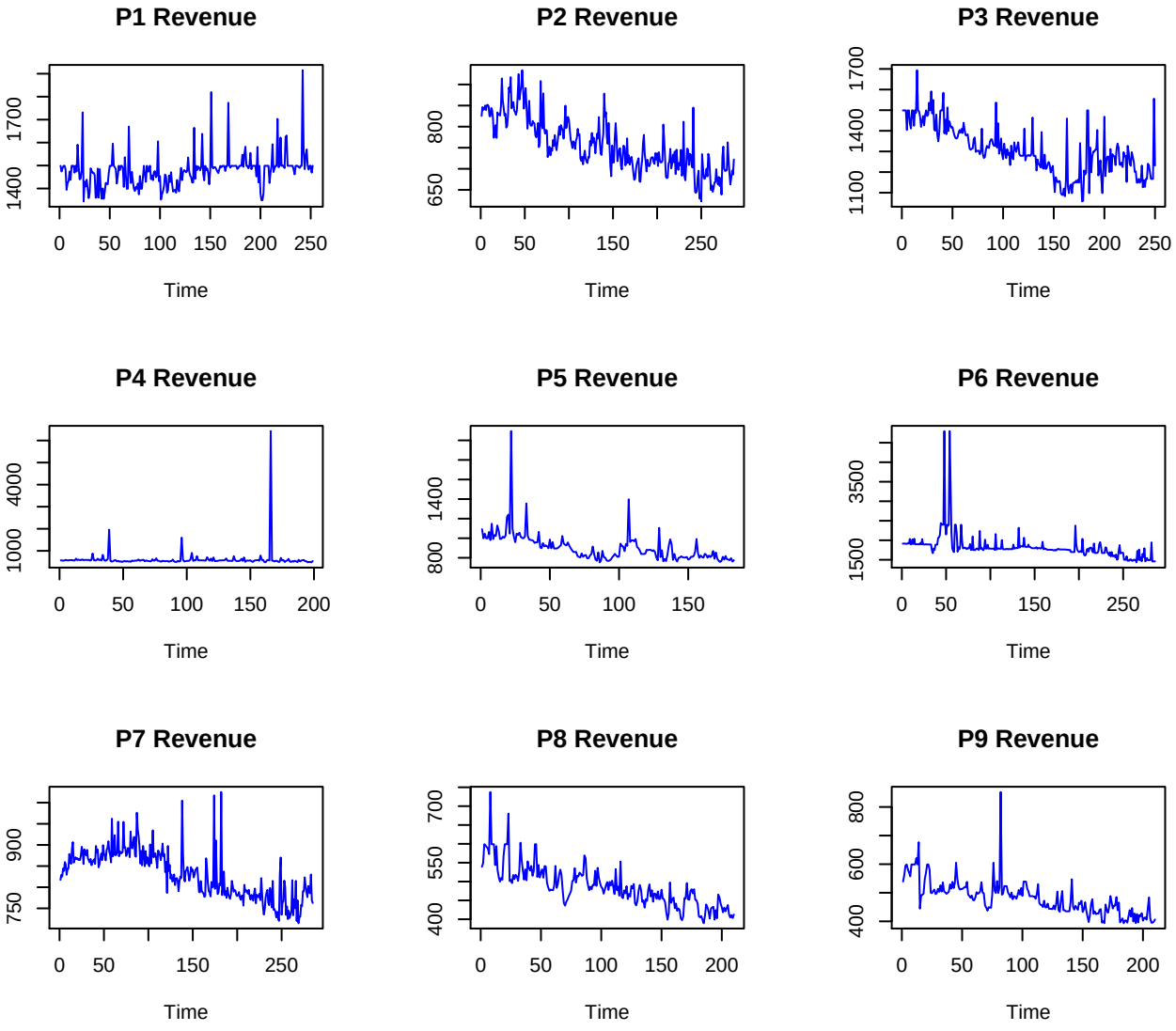

Figure 1: Products revenue

Therefore, from the previous table, it is possible to conclude that the best model in all cases is the NNETAR(p, k), since

$$\widehat{\sigma_e}\text{NNETAR}(p,k) < \widehat{\sigma_e}\text{ETS} < \widehat{\sigma_e}\text{ARIMA}(p,d,q),$$

where the model NNETAR(p, k) can be written as follows,

$$x(n) = f\left(\sum_{r=1}^{p} \mathbf{\Phi}_r \mathbf{x}(n-r)\right) + e(n),$$

in which, $\mathbf{x}(n-r)$ is a $k \times 1$ input vector, $\mathbf{\Phi}_r$ is a $k \times k$ of synaptic weights (parameter) matrix, k is the number of network hidden layers, $f(\cdot)$ is the activation function (sigmoid) and $e(n)$ independent, identically, distributed process (i.i.d.) with null mean and variance σ_e^2 ($\{e(n)\}_{n\in\mathbb{Z}} \sim$ i.i.d.$(0, \sigma_e^2)$).

However, previous candidate experience of manipulating annual data and NNETAR(p, k) models indicates the necessity of a fine tuning procedure of the parameters p and k for the products $P_i, i = 1,\ cdots, 9$, such that the new models are capable of capturing the nonlinear dynamics of the processes and possible other effects treated below.

As we are dealing with daily data and despite the apparent lack of seasonality in the graphs of Figure 1, the parameters $p = k = 30$ are imposed on the neural network, so that it is able to capture the monthly variation and the nature of the nonlinear dynamics of the data.

As a matter of fact, in the next table, note the decrease in value of $\widehat{\sigma_e}$NNETAR(p, k) in all cases.

Produto	NNETAR$(30, 30)$
P_1	$\widehat{\sigma_e} = 2{,}67 \times 10^{-5}$
P_2	$\widehat{\sigma_e} = 4{,}74 \times 10^{-5}$
P_3	$\widehat{\sigma_e} = 7{,}59 \times 10^{-5}$
P_4	$\widehat{\sigma_e} = 1{,}54 \times 10^{-4}$
P_5	$\widehat{\sigma_e} = 1{,}00 \times 10^{-4}$
P_6	$\widehat{\sigma_e} = 9{,}82 \times 10^{-3}$
P_7	$\widehat{\sigma_e} = 2{,}95 \times 10^{-4}$
P_8	$\widehat{\sigma_e} = 7{,}28 \times 10^{-5}$
P_9	$\widehat{\sigma_e} = 8{,}26 \times 10^{-5}$

Figure 2 illustrate the forecast produced by the NNETAR$(24, 24)$ models for the products revenues.

In the following section the case for competitors' data is shown.

The 'comp_prices.cs' data set

For the database 'comp_prices.cs' it was necessary to realize a reduction of dimensionality.

First, we groupped the data by payment type, product id, date and time and then take the mean value of the prices, regardless the competitors and products. Graphs in Figure 3 shows that there are no differences between neither dianurnal (12am — 12pm, blue line) nor nocturnal (12pm — 12am, red line) period of the day, and payment type and, thus, can be classified as independent variables.

Now to investigate the importance of each competitor, we clustered the information by competitor, date and time and then take the mean value of the prices, regardless the competitors and product id, obtaining Figure 5.

Noticed that Figure 5 shows no differentiation between the diurnal and nocturnal, for all competitors.Therefore, henceforth no distinction will be made between these variables (that is, $C_i, i = 1,\ cdots, 6$).

Next, Figure 5 shows the time series of clustered data by product id, date and time and then take the mean value of the prices, regardless the competitors.

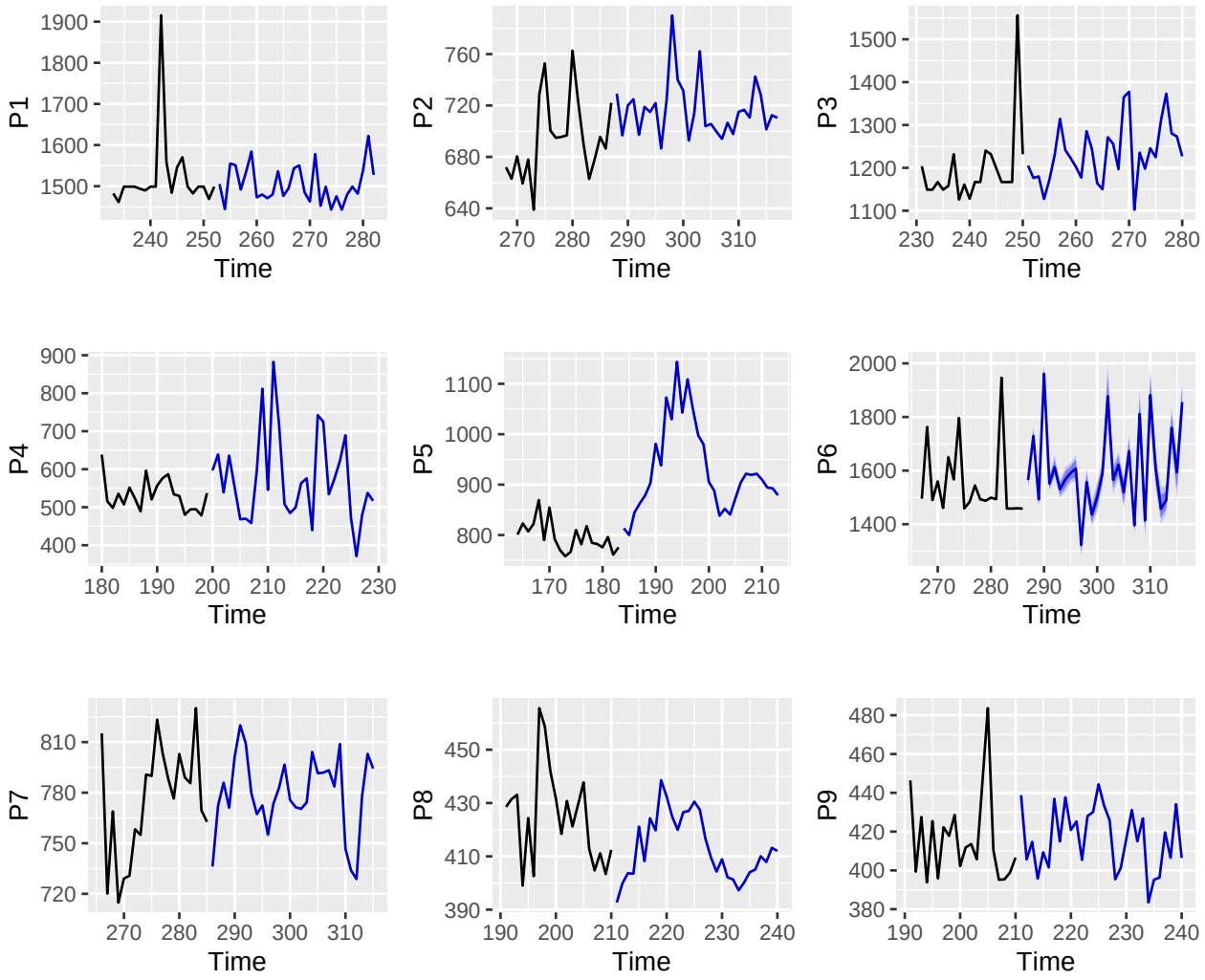

Figure 2: Product revenue forecast for the next 30 days.

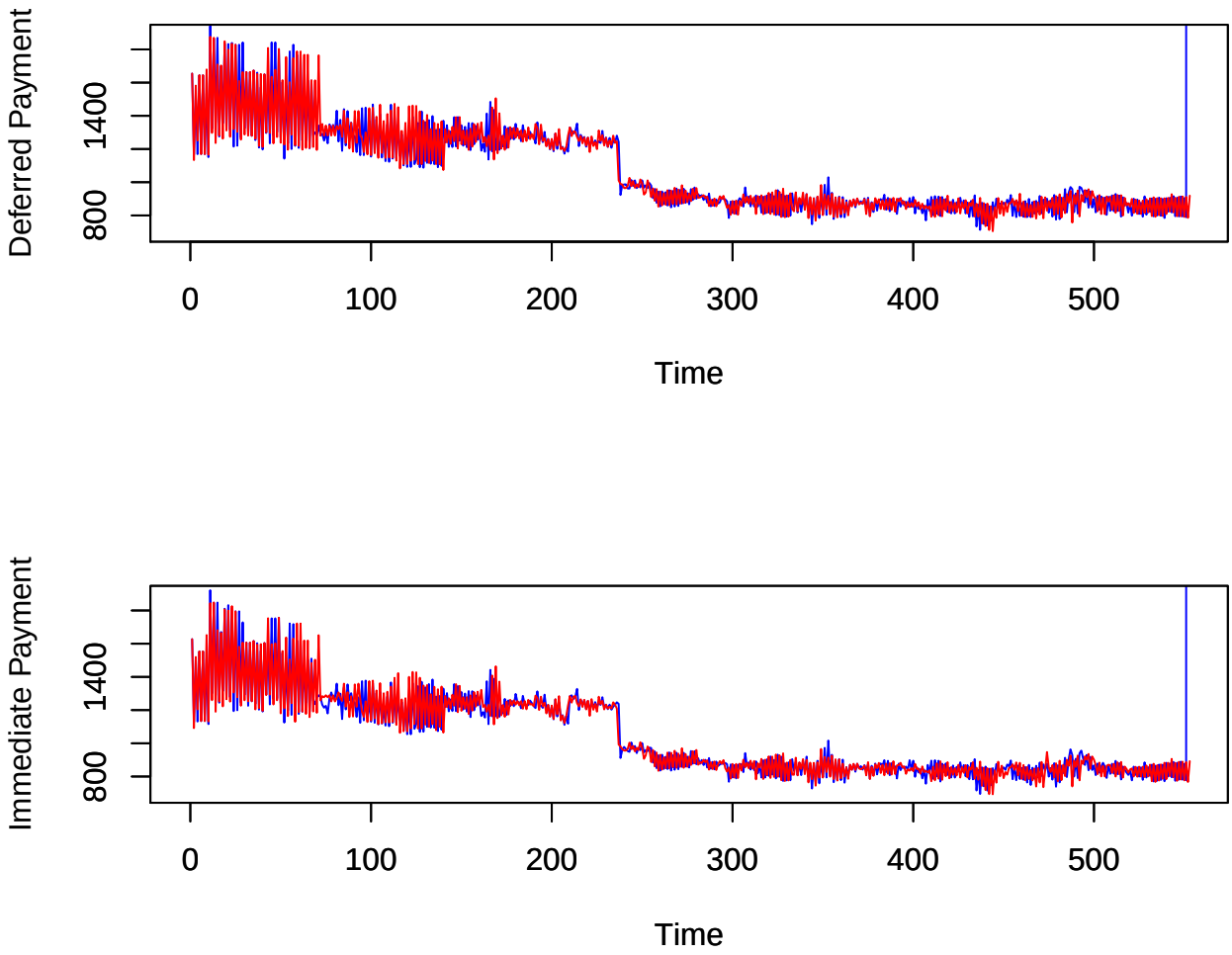

Figure 3: Payment type divide by period of the day: dianurnal (blue line) and nocturnal (red line).

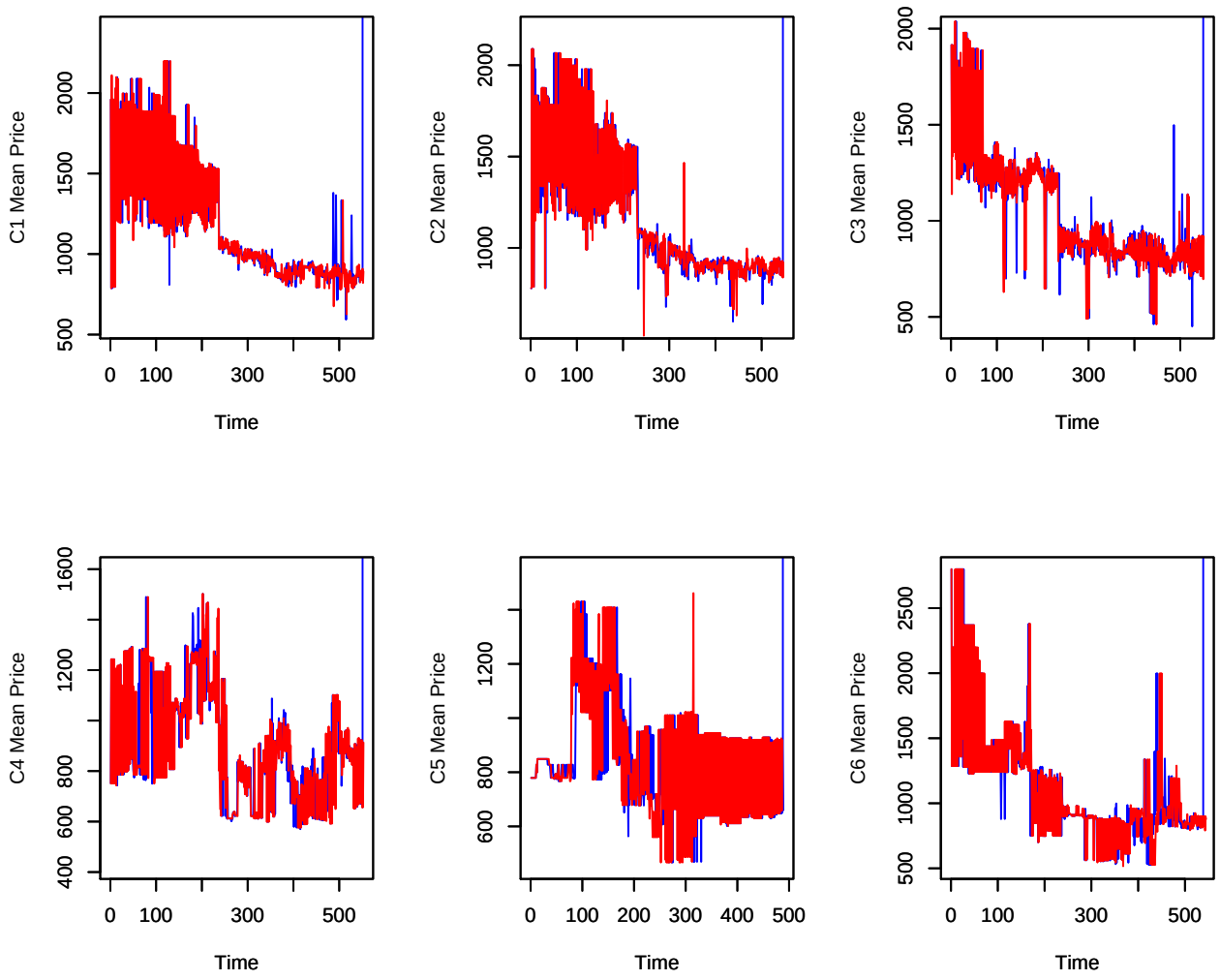

Figure 4: Prices due to competitors and diurnal (blue line) and nocturnal (red line) periods.

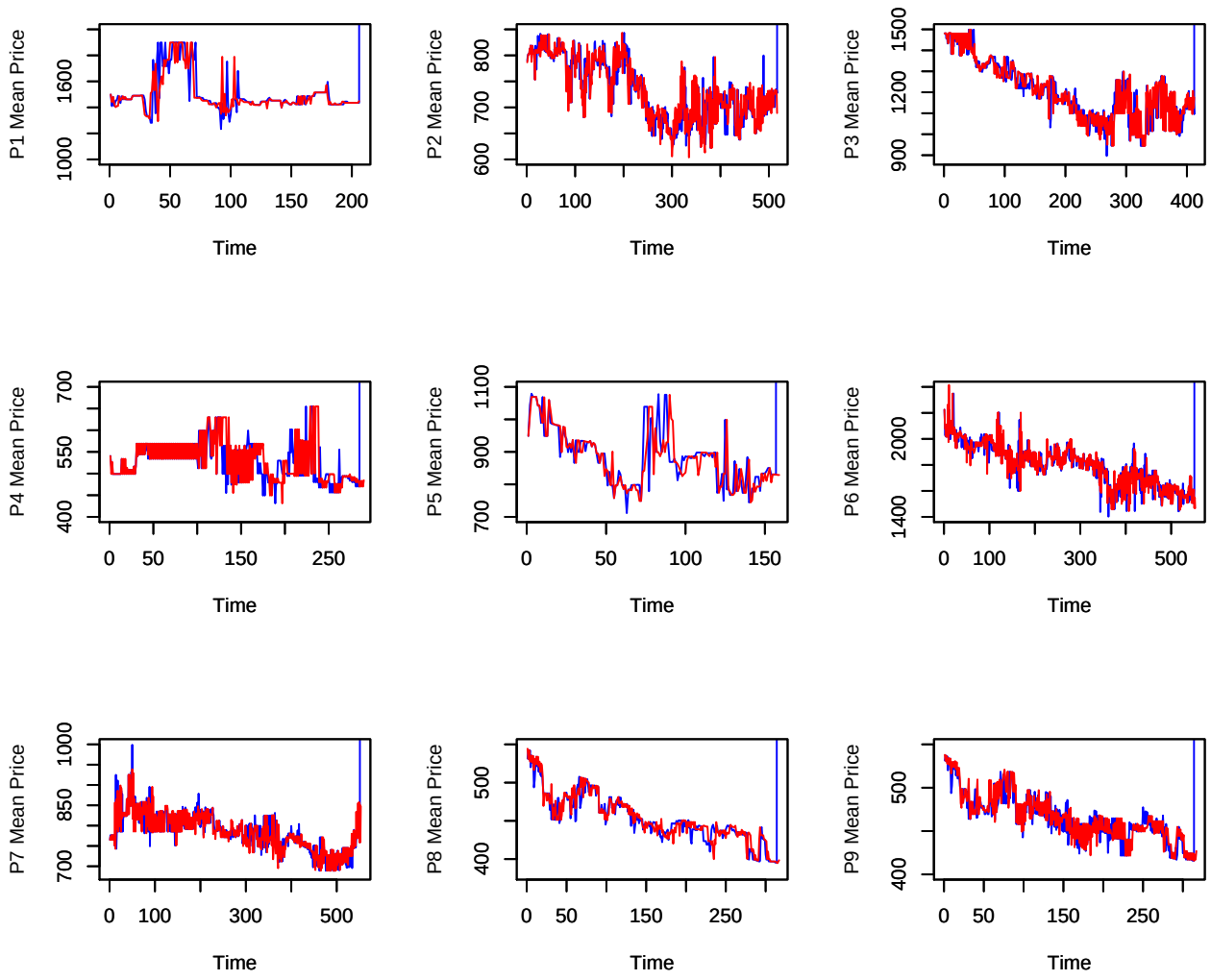

Figure 5: Competitors products time series, showing the diurnal (blue line) and nocturnal (red line) periods.

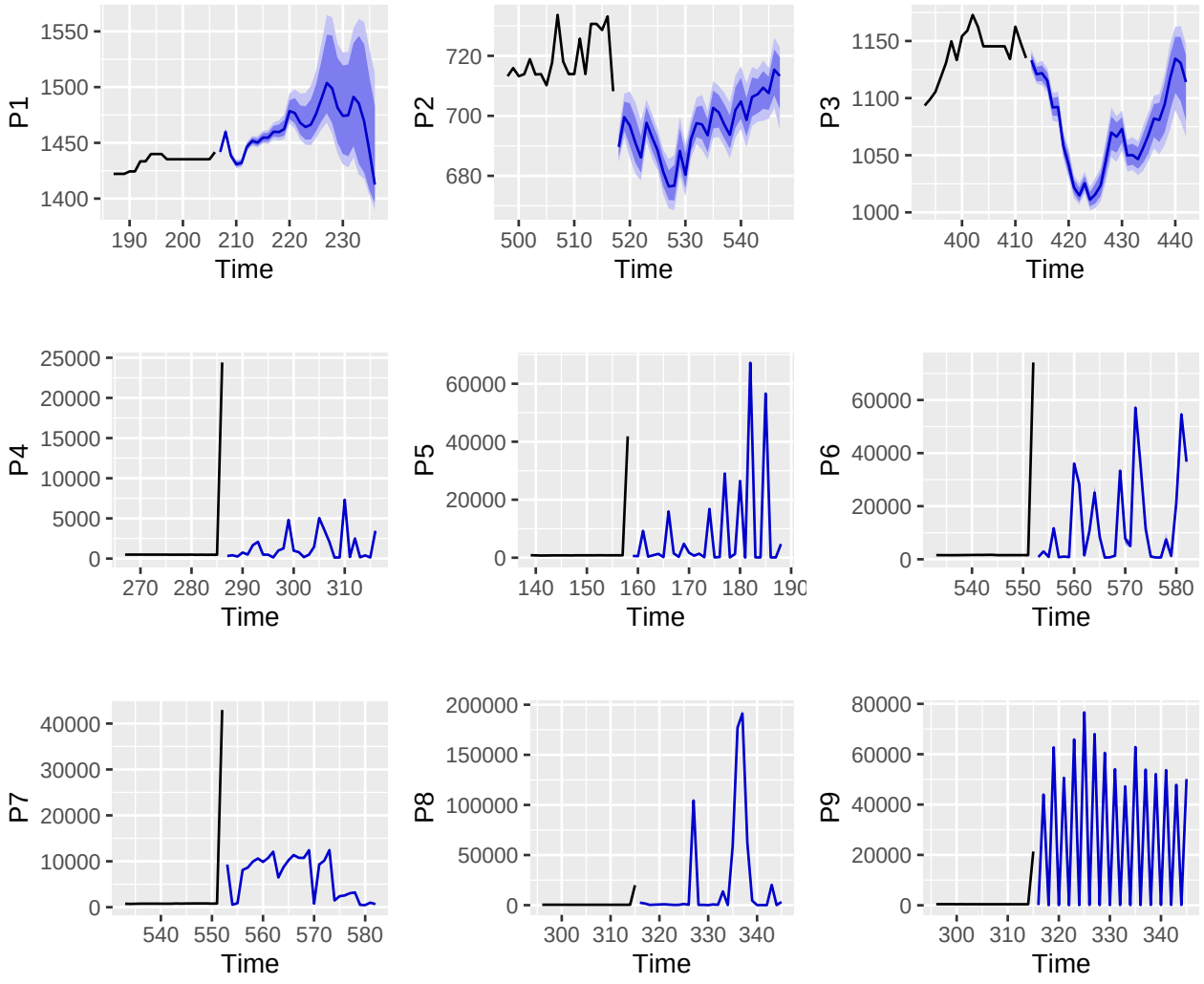

Figure 6: Forecast of the average prices of the competitors' products for the next 30 days.

As can be seen, here there is no differentiation between the periods of the day and, therefore, to obtain only one series per product, the mean between the diurnal and nocturnal periods were used, in order to be able to forecast the prices as shown in Figure **??**.

As in the previous section, we used NNETAR$models(p, k)$, with $p = k = 30$ so that it is able to capture dynamics such as possible mensal seazonality and nonlinear dynamics. Finally, in the following table it is shown an evaluation of the models given the standard deviations of its residues.

Produto	NNETAR$(30, 30)$
P_1	$\widehat{\sigma_e} = 1{,}01 \times 10^{-3}$
P_2	$\widehat{\sigma_e} = 4{,}05 \times 10^{-3}$
P_3	$\widehat{\sigma_e} = 4{,}03 \times 10^{-3}$
P_4	$\widehat{\sigma_e} = 5{,}61 \times 10^{-3}$
P_5	$\widehat{\sigma_e} = 5{,}77 \times 10^{-4}$
P_6	$\widehat{\sigma_e} = 1{,}29 \times 10^{-2}$
P_7	$\widehat{\sigma_e} = 8{,}92 \times 10^{-3}$
P_8	$\widehat{\sigma_e} = 5{,}19 \times 10^{-3}$
P_9	$\widehat{\sigma_e} = 5{,}37 \times 10^{-3}$

Therefore, by comparing the above table with the last table of the previous section it is possible to infer

that all the models for the products of the database 'sales.csv' produced a better predictive model, since the standard deviations of the residuals were less than those of competitors.

Therefore, the predictive power of the data provided for B2W are higher than for the competition.

References

1. Hyndman, Rob J. and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2018.