Relatório

Lucas Massaroppe

28/06/2019

Desafio B2W

Neste desafio foram consideradas duas bases de dados: (a) 'sales.csv' e (b) 'comp_prices.csv'. Em (a) consta informação transacional de nove produtos $P_i, i = 1, \cdots, 9$ para o ano de 2015 e em (b), contém dados de seis competidores $C_i, i = 1, \cdots, 6$, monitorados em horários diferentes, para as mesmas datas e produtos dos primeiros elementos.

Primeiramente analisaremos 'sales.csv' e, em uma segunda etapa, será dada atenção à 'comp_prices.csv', para que possamos assim fazer uma comparação entre as informações transacionais e os preços dos competidores.

Base 'sales.csv'

Como uma verificação rápida apresentamos as séries temporais dos nove produtos na Figura 1. Assim, podemos realizar o teste KPSS sobre cada uma delas a fim de saber se elas possuem raiz unitaria e tendência, ou seja, se são ou não estacionárias.

A tabela abaixo apresenta os resultados do teste KPSS.

Produto	Estatística	p-valor
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 que, de acordo com a hipótese nula do teste KPSS e utilizando um nível de significância $\alpha = 1\%$, as séries dos produtos que apresentam p-valor menor do α são estacionárias, ou seja, P_3, P_5 e P_7.

Previsão das séries dos produtos

Para todas as séries temporais utilizamos três modelos de previsões: o linear ARIMA(p, d, q), o não-linear de suavização exponencial ETS e o de rede neural autorregresivo NNETAR(p, k). Para avaliar qual o melhor modelo entre os três, utilizamos o desvio padrão dos resíduos como métrica de comparação e mostramos na tabela a seguir.

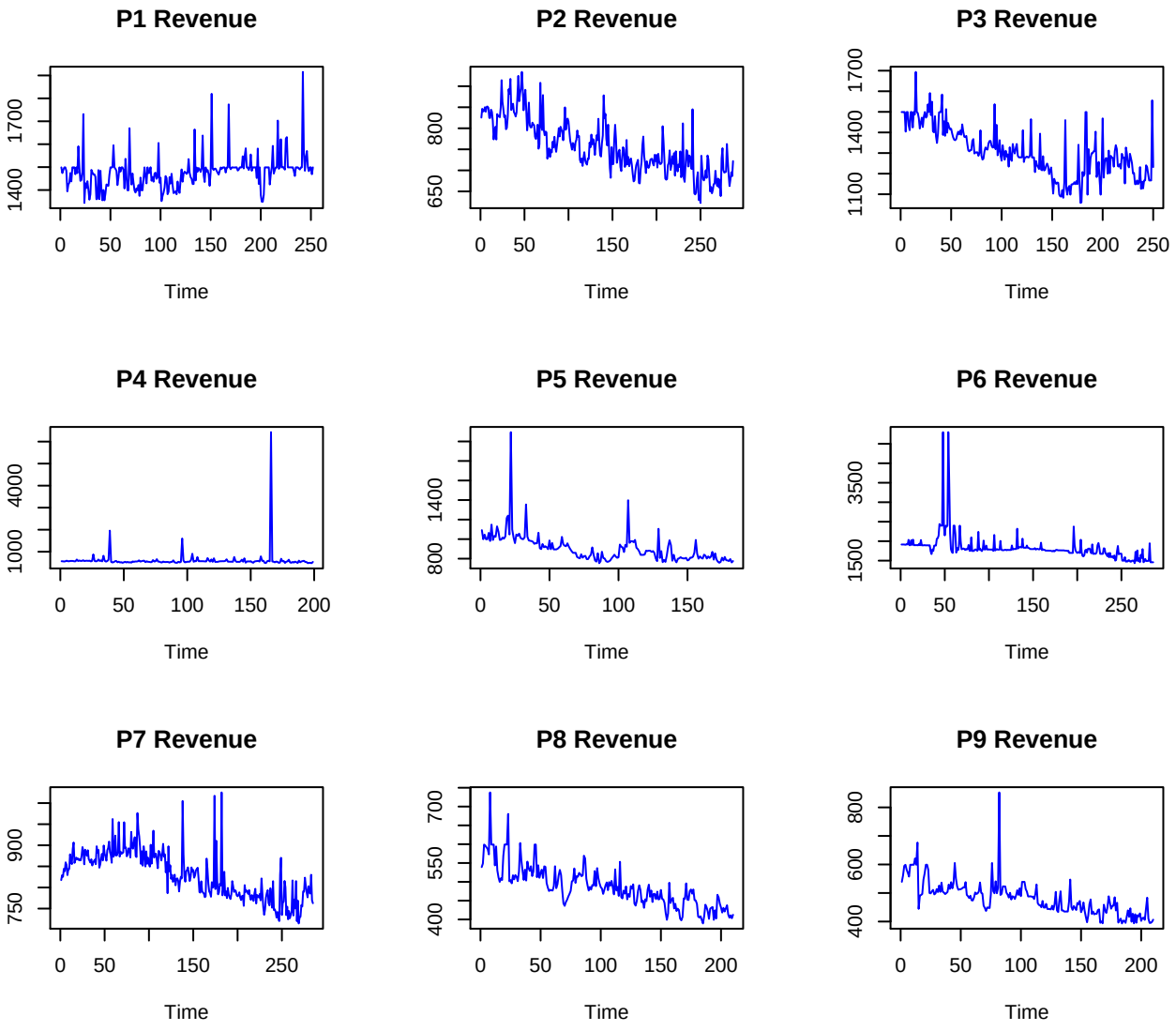

Figure 1: Preço médio dos produtos.

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$(6,4)$, $\widehat{\sigma_e} = 0{,}06$

Portanto, pela tabela anterior, é possível concluir que o melhor modelo em todos casos é o NNETAR(p,k), pois

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

em que se pode escrever o modelo NNETAR(p,k) da seguinte forma,

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

em que, da rede neural, tem-se que, $\mathbf{x}(n-r)$ é um vetor $k \times 1$ com as respectivas entradas, $\mathbf{\Phi}_r$ sd matrizes de parâmetros de dimensão $k \times k$, k o numero de camadas, $f(\cdot)$ a função de ativação (que, no caso, é a sigmóide) e $e(n)$ um processo independente, identicamente distributído, de média nula e variância σ_e^2 ($\{e(n)\}_{n\in\mathbb{Z}} \sim$ i.i.d.$(0,\sigma_e^2)$).

Porém, pela experiência prévia de manipulação de dados anuais do candidato e modelos NNETAR(p,k) é necessário se modificar os parâmetros p e k dos modelos dos produtos $P_i, i = 1, \cdots, 9$, para que os mesmos sejam capazes de capturar a dinâmica não-linear dos processos.

Assim, como se trata de dados anuais e apesar de se possuir apenas um ano de dados, para ambos os parâmetros utilizam-se múltiplos de 12, ou seja $p = k = 24$ para que se possa ter a restrição de o moelo conseguir capturar anualidade.

De fato, observe que na próxima tabela os valores de $\widehat{\sigma_e}$NNETAR(p,k) diminuem em todos os casos.

Produto	NNETAR$(24,24)$
P_1	$\widehat{\sigma_e} = 2{,}83 \times 10^{-5}$
P_2	$\widehat{\sigma_e} = 1{,}08 \times 10^{-4}$
P_3	$\widehat{\sigma_e} = 2{,}32 \times 10^{-4}$
P_4	$\widehat{\sigma_e} = 1{,}81 \times 10^{-4}$
P_5	$\widehat{\sigma_e} = 9{,}82 \times 10^{-5}$
P_6	$\widehat{\sigma_e} = 1{,}17 \times 10^{-2}$
P_7	$\widehat{\sigma_e} = 1{,}38 \times 10^{-3}$
P_8	$\widehat{\sigma_e} = 7{,}36 \times 10^{-5}$
P_9	$\widehat{\sigma_e} = 8{,}32 \times 10^{-5}$

Na Figura 6 a seguir, mostra-se a previsão realizada por esses modelos.

Na seção seguinte mostra-se o caso para os dados dos competidores.

Base 'comp_prices.csv'

Para os dados da base 'comp_prices.csv' fez-se necessário realizar uma redução de dimensionalidade.

Primeiramente, agrupou-se os dados em relação à forma de pagamento, ou seja, à prazo ou à vista, obtendo-se os gráficos da Figura 3. Assim, como se observa nessa figura, percebe-se que não há diferença entre as formas de pagamentos e podemos classficar as variáveis independente desta que estamos analisando.

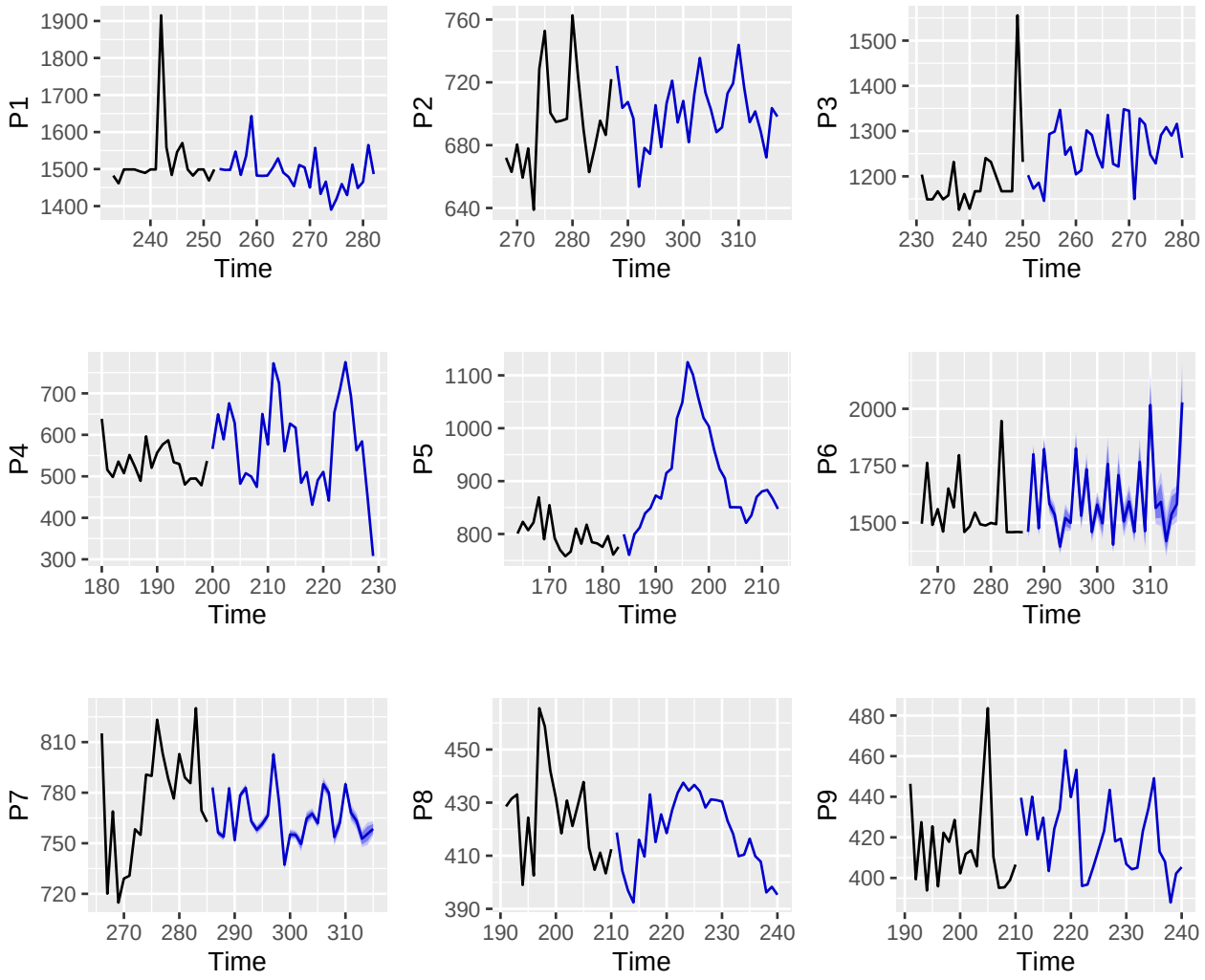

Figure 2: Previsão dos preços médios dos produtos para os próximos 30 dias.

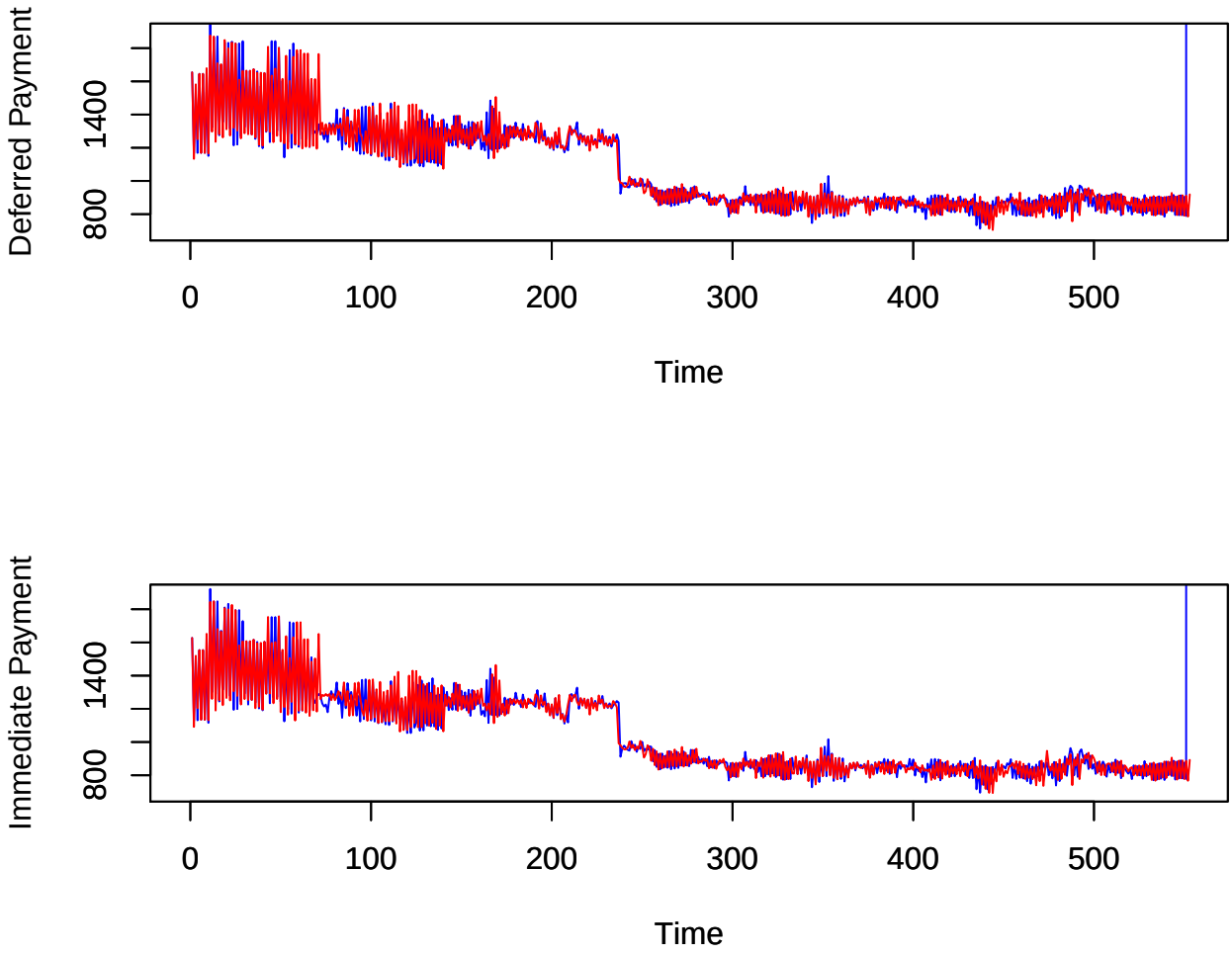

Figure 3: Tipo de pagamento dos competidores.

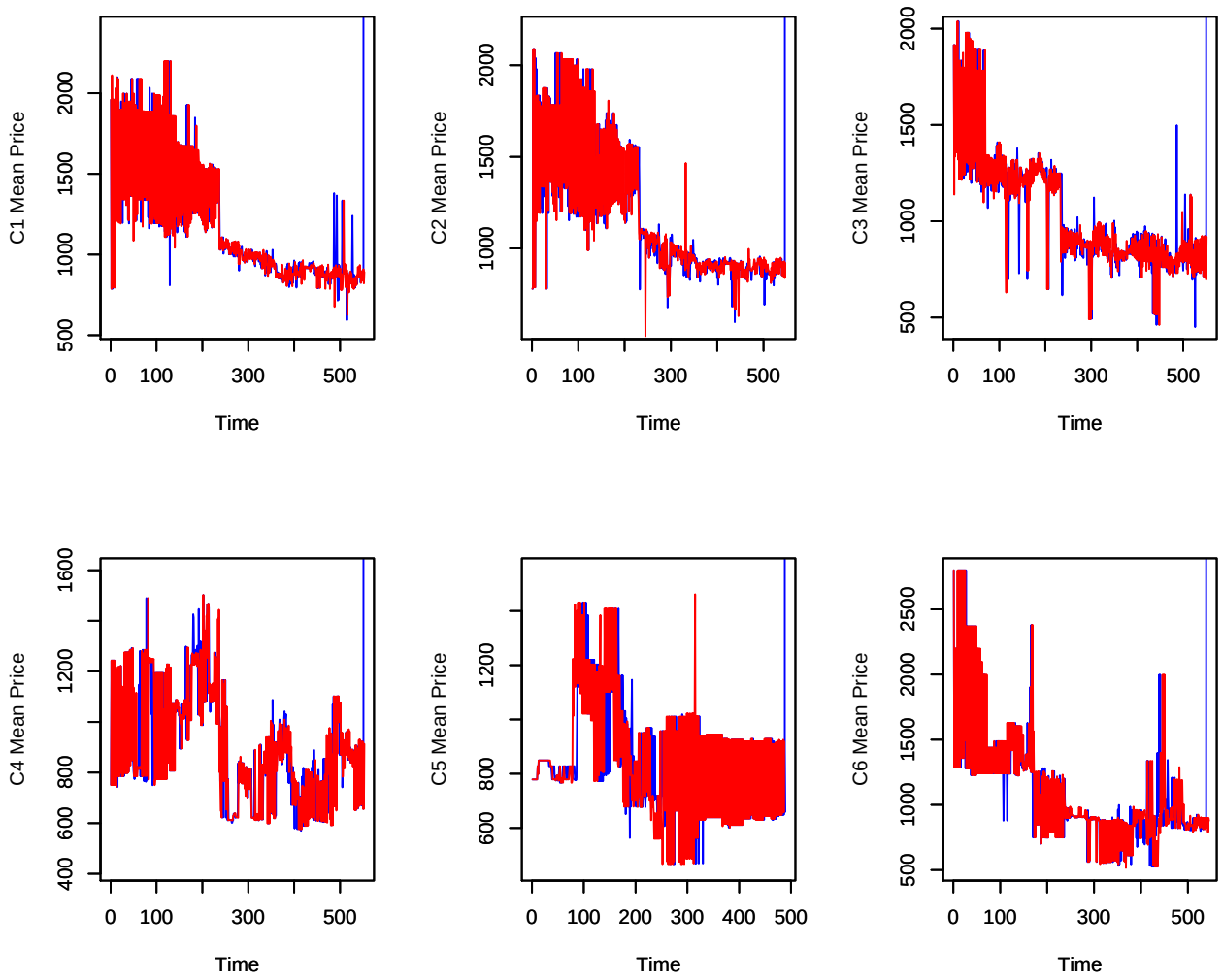

Figure 4: Pagamentos segundo os competidores e períodos diúrno (linha azul) e noturno (linha vermelha).

Já para investigar a importância de cada competidor, é importante aglomerar as variáveis acima e organizá-las indiferentemente dos produtos, porém mantendo ordem temporal, obtendo-se a Figura 5.

Pela Figura 5, percebe-se que não há diferenciação entre os períodos diurnos (linha azul) e noturno (linha vermelha) para todos os competidores. Logo, daqui para frente não será feita diferenciação entre essas variáveis (ou seja, $C_i, i = 1, \cdots, 6$).

A seguir, mostra-se as séries temporais dos produtos aglomerados independentemente dos competidores.

Como percebe-se, aqui não diferenciação entre os períodos do dia e, portanto, para se obter apenas uma série por produto, utilizou-se a média entre os diúrnos e noturno, para se poder fazer a previsão dos preços como é mostrado na Figura **??**, a seguir.

Assim como na seção anterior, aqui foram utilizados modelos NNETAR(p, k), com $p = k = 24$ para que esse seja capaz de capturar dinâmicas tais como possíveis sazonalidades anuais dos preços e não-linearidades e por fim, na tabela seguinte mostra-se uma avalição dos modelos a partir dos desvios padrões de seus resíduos.

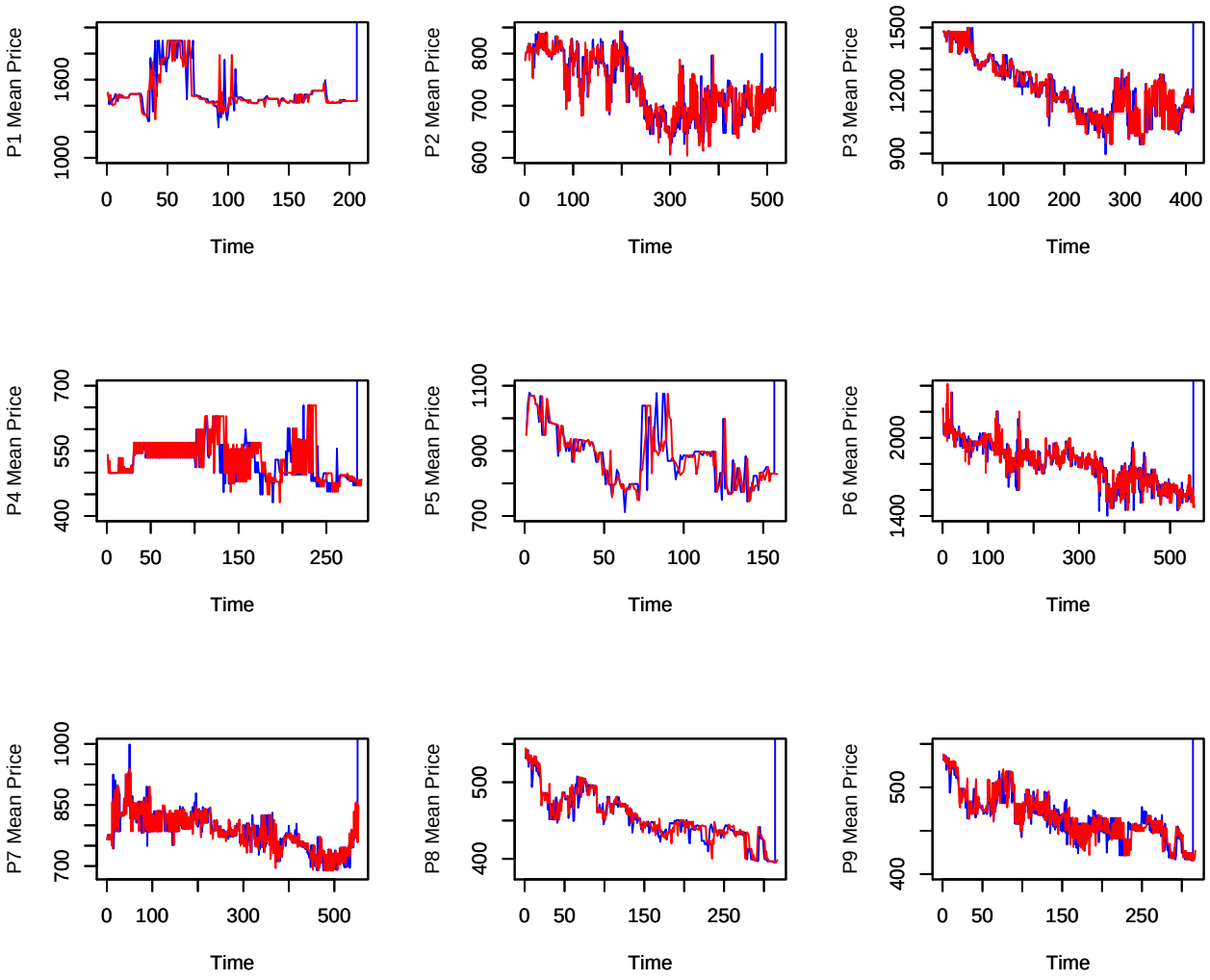

Figure 5: Séries dos produtos, mostrando os períodos diúrno (linha azul) e noturno (linha vermelha).

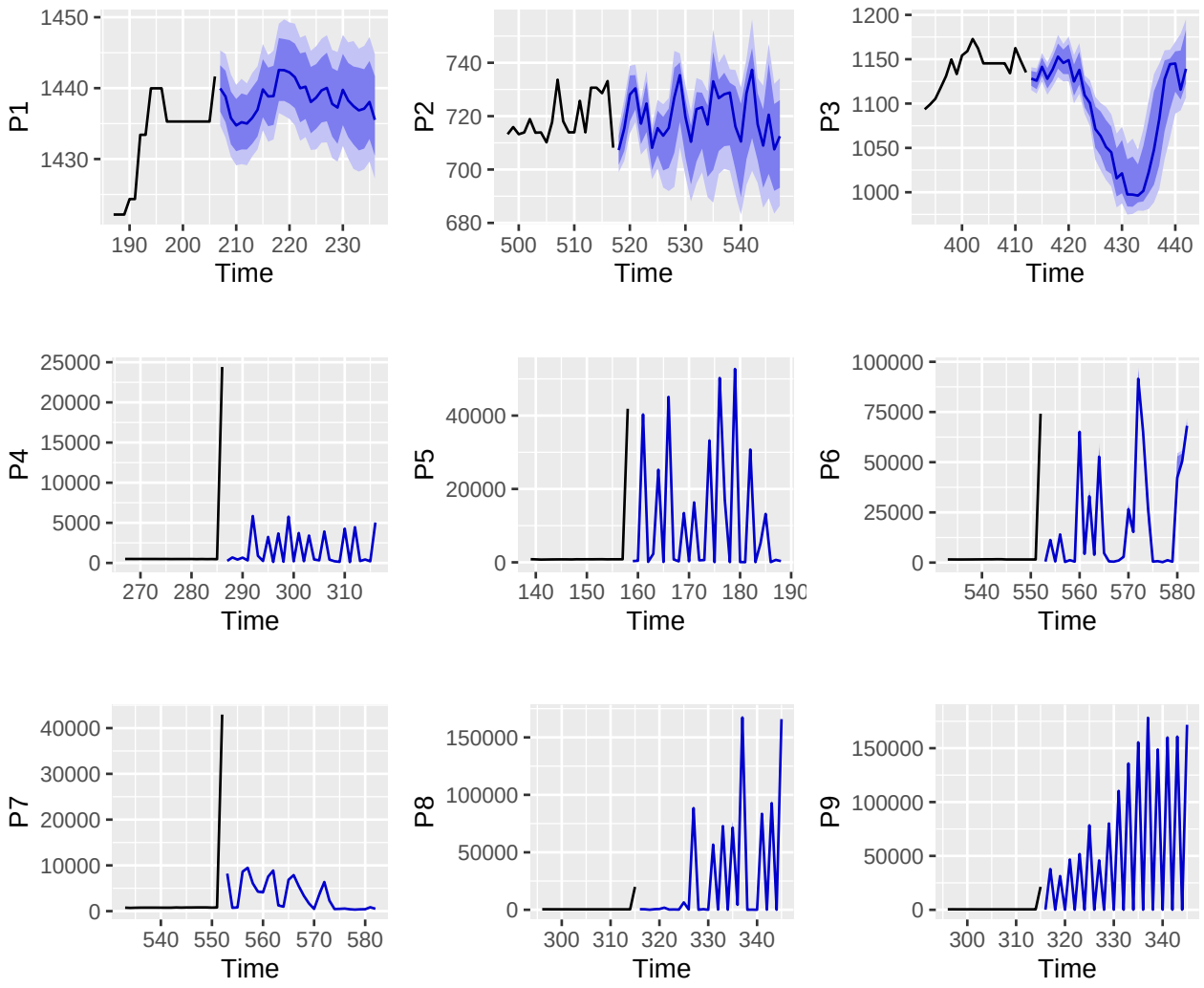

Figure 6: Previsão dos preços médios dos produtos para os próximos 30 dias.

Produto	NNETAR$(24, 24)$
P_1	$\widehat{\sigma_e} = 1{,}72 \times 10^{-3}$
P_2	$\widehat{\sigma_e} = 5{,}74 \times 10^{-3}$
P_3	$\widehat{\sigma_e} = 5{,}79 \times 10^{-3}$
P_4	$\widehat{\sigma_e} = 9{,}10 \times 10^{-3}$
P_5	$\widehat{\sigma_e} = 2{,}97 \times 10^{-3}$
P_6	$\widehat{\sigma_e} = 1{,}63 \times 10^{-2}$
P_7	$\widehat{\sigma_e} = 1{,}05 \times 10^{-2}$
P_8	$\widehat{\sigma_e} = 7{,}47 \times 10^{-3}$
P_9	$\widehat{\sigma_e} = 8{,}28 \times 10^{-3}$

Portanto, comparando a última tabela com a última tabela da seção anterior é possível inferir que todos os modelos para os produtos da base de dados 'sales.csv' foram superiores, pois os desvios padrões dos resíduos dos modelos das previsões para os produtos $(P_i, i = 1, \cdots, 9)$ foram menores para os da B2W do que os dos competidores.

Assim, o poder preditivo dos dados fornecidos para a B2W são superiores do que para os da competição.