Processo 4intelligence ADS

Luis Fernando Corrêa da Costa

20/06/2021

Questão 1

Análise Descritiva: Em anexo, você recebeu uma base de dados (Bases Final ADS Jun2021) com o consumo de energia residencial, comercial e industrial de cada região brasileira. Faça uma análise descritiva das variáveis e, eventualmente, da relação entre elas.

Questão 2

Modelagem: Utilizando-se das variáveis fornecidas na base de dados Bases Final ADS Jun2021.xlsx, forneça um modelo que projete, com a melhor acurácia possível, o consumo de energia industrial da região Sudeste para os próximos 24 meses.

1. Explique o método e a razão de utilizar a abordagem escolhida na sua projeção. Quais "insights" podem ser obtidos da modelagem?
2. Forneça medidas para avaliar a qualidade da projeção do modelo.
3. Justifique a escolha das variáveis explicativas e avalie o poder explicativo delas.

Questão 3

Levando em consideração a modelagem apresentada acima, escolha os 5 melhores modelos em termos de acurácia e argumente a razão de tê-los escolhido.

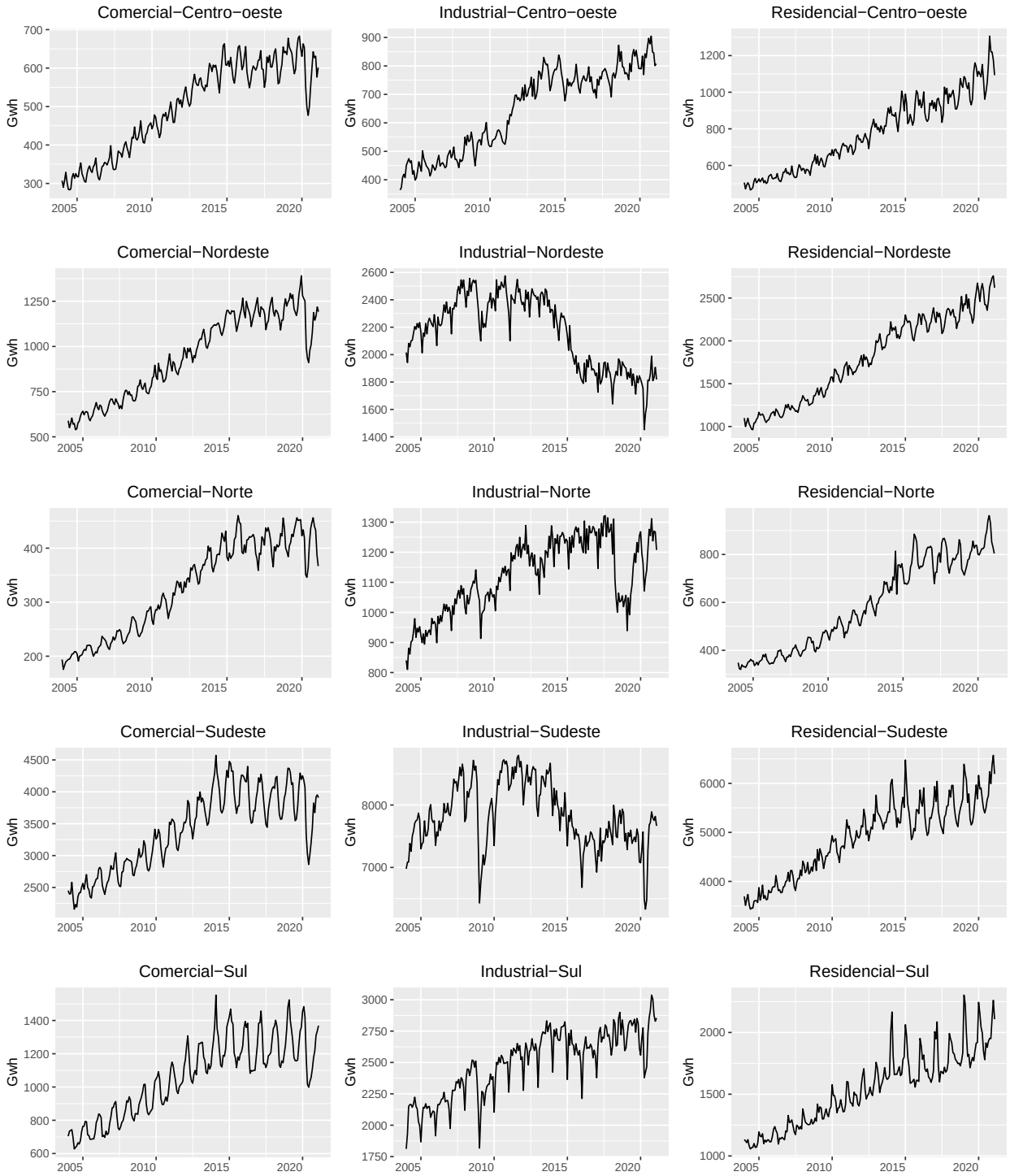
Questão 4

O que é possível tirar de conclusões a partir dos exercícios 1, 2, e 3?

Eu procuro responder a todas as questões acima ao longo da análise realizadas nas páginas seguintes

Análise descritiva do consumo de energia setorial nas regiões brasileiras

Plotagem das séries temporais relativas ao consumo de energia nos setores eregiões



A figura acima apresenta as séries temporais mensais do consumo de energia comercial, industrial e residencial nas cinco regiões brasileiras entre janeiro de 2004 e fevereiro de 2021.

A pricipal característica observada é a elevada sazonalidade presente nas séries, especialmente na setor comericial e no consumo residencial. Adicionalmente, observa-se o impacto durante os períodos de crise sobre o consumo de energia, sobretudo na indústria, a qual responde mais intesamente à atividade econômica.

As estatísticas descritivas do consumo de energia nos estados são apresentados a seguir em três tabelas classicadas para os setroes analisados.

Estatísticas descritivas do consumo de energia nas regiões

Consumo energia comercial

Estatisticas descritivas do consumo de energia no comercio

regiao	Min	Max	Media	Mediana	Desv_pad
Centro-oeste	283.1044	683.1114	499.4131	520.4980	117.08220
Nordeste	539.4518	1390.6223	949.4825	972.2660	230.05091
Norte	175.4953	460.2620	327.0770	340.6408	85.11528
Sudeste	2159.4795	4571.7170	3399.4874	3488.6665	622.48227
Sul	627.6517	1552.6660	1058.5056	1090.0800	229.85414

Consumo energia residencial

Estatisticas descritivas do consumo de energia residencial

regiao	Min	Max	Media	Mediana	Desv_pad
Centro-oeste	466.3588	1306.3738	776.6566	769.7115	200.2737
Nordeste	962.9343	2758.9849	1799.7010	1804.6390	503.5601
Norte	320.2374	962.1004	588.7700	566.8075	185.4983
Sudeste	3433.4425	6571.3116	4903.5937	5048.8835	765.9491
Sul	1056.8501	2302.7874	1540.2121	1565.9381	298.2800

Consumo energia industrial

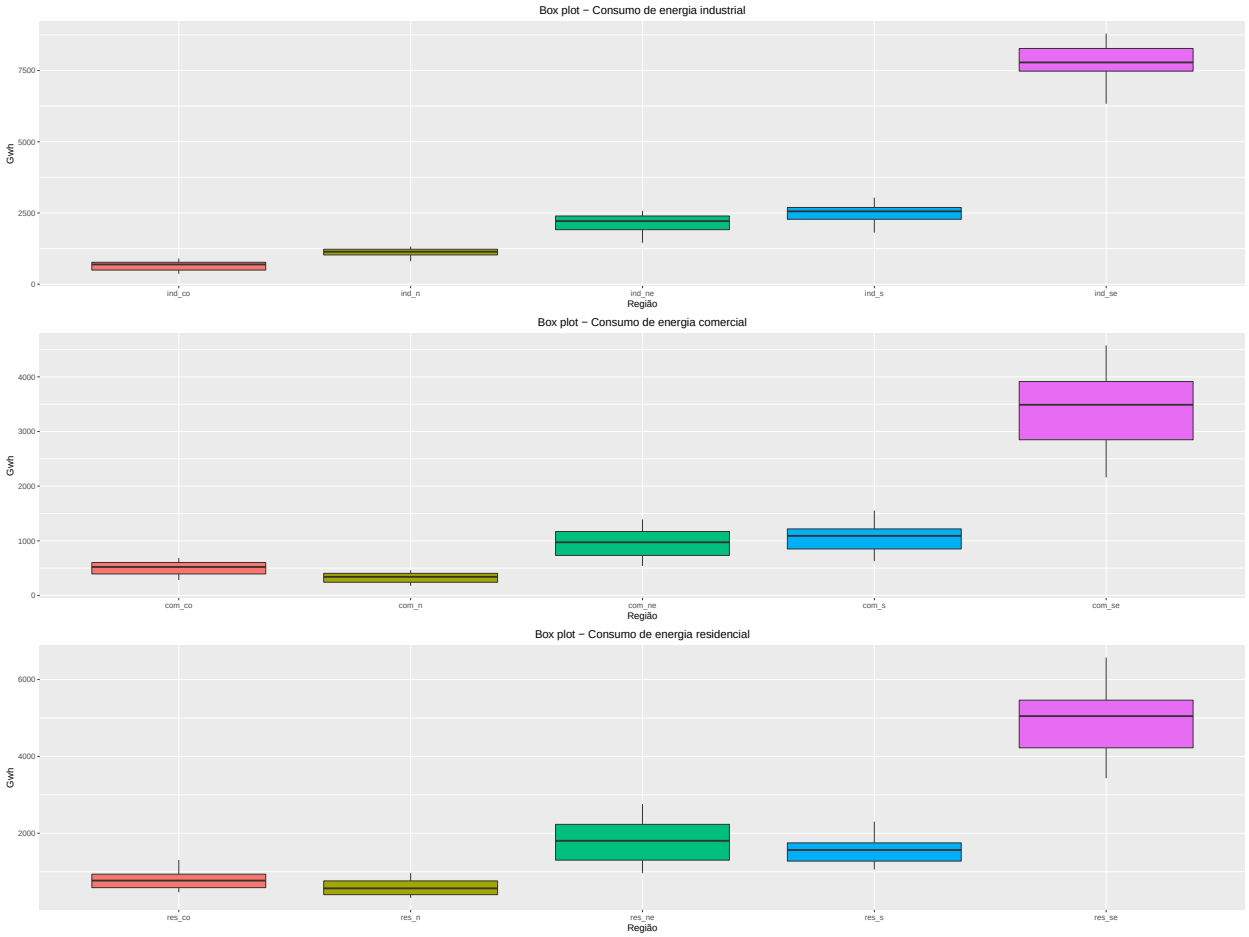
Estatisticas descritivas do consumo energia na industria

regiao	Min	Max	Media	Mediana	Desv_pad
Centro-oeste	364.3270	904.7817	645.9656	693.9215	146.8212
Nordeste	1452.1316	2574.7110	2164.5811	2215.2132	258.9640
Norte	810.2563	1321.9580	1121.5087	1140.0723	122.2530
Sudeste	6331.1189	8795.5540	7828.9091	7783.6148	519.7338
Sul	1810.9802	3037.0106	2495.4679	2557.7429	260.2037

As estaísticas descritivas bem como as distribuições das séries do consumo de energia pode ser visualizado na figura a seguir, a qual informa o boxplot das séries de consumo de energia dos estados e setores.

Tal como esperado, a região sudeste se destaca das demais, não apenas por ser a mais populosa, mas também por concentrar a maior parte da atividade econômica do país.

Boxplot da distribuição do consumo de energia dos setores e por região.



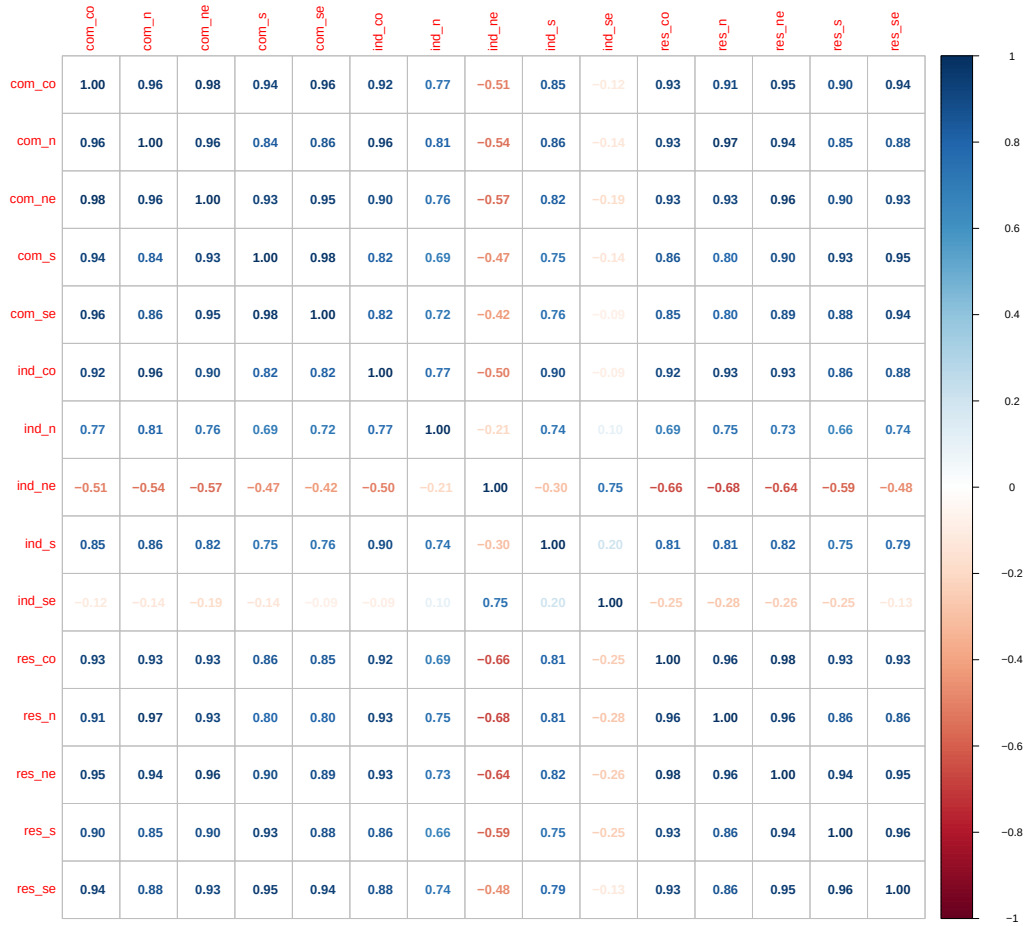
A correlação entre as variáveis pode ser vista na figura a seguir.

Ressalta-se que uma forte correlação positiva entre as váriáveis em geral. Contudo, dois pontos principais podem ser destacados:

1. Diferentemente da tendência geral, o consumo de energia industrial do nordeste possui uma correlação negativa como as demais variáveis, exceto o consumo industrial do sudeste.

2. Tal como ressaltado, o consumo industrial do nordeste e de do sudeste são positivamente correlacionados. No caso dessa última região, a correlação da energia industrial guarda baixa correlação com as demais variáveis.

Gráfico de correlação entre o consumo de energia dos setores e regiões do país.



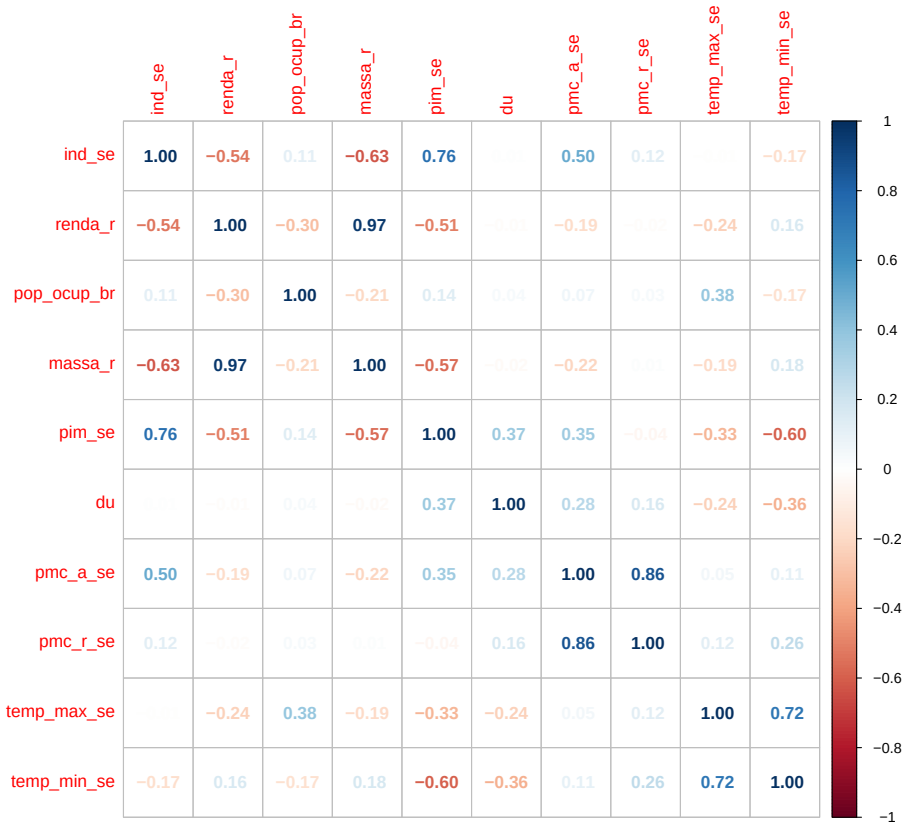
Modelagem

Essa seção abrda a modelagem do consumo de energia da região industrial do SE e projeção para 22 meses à frente.

Preparação da base de dados para a modelagem

Os dados foram selecionados após 2012-03-01, uma vez que todas as variáveis possuem observações a partir dessa data.

Como exercício inical, analisa-se a correlação entre a variável dependente, o consumo de energia industrial do sudeste, e os regressores. A figura a seguir ilustra tais correlações.



	ind_se	renda_r	pop_ocup_br	massa_r	pim_se	du	pmc_a_se	pmc_r_se	temp_max_se	temp_min_se
ind_se	1.00	−0.54	0.11	−0.63	0.76		0.50	0.12		−0.17
renda_r	−0.54	1.00	−0.30	0.97	−0.51		−0.19		−0.24	0.16
pop_ocup_br	0.11	−0.30	1.00	−0.21	0.14		0.07		0.38	−0.17
massa_r	−0.63	0.97	−0.21	1.00	−0.57		−0.22		−0.19	0.18
pim_se	0.76	−0.51	0.14	−0.57	1.00	0.37	0.35		−0.33	−0.60
du					0.37	1.00	0.28	0.16	−0.24	−0.36
pmc_a_se	0.50	−0.19	0.07	−0.22	0.35	0.28	1.00	0.86		0.11
pmc_r_se	0.12					0.16	0.86	1.00	0.12	0.26
temp_max_se		−0.24	0.38	−0.19	−0.33	−0.24		0.12	1.00	0.72
temp_min_se	−0.17	0.16	−0.17	0.18	−0.60	−0.36	0.11	0.26	0.72	1.00

A despeito das correlações acima vistas, algumas hipóteses podem ser consiradas no tocante dos efeitos das variáveis explicativas sobre o consumo de energia industrial.

1. Espera-se relações positivas entre o consumo de energia e a produção industrial, bem como com variáveis relacionadas à renda, a qual não é afirmada pelas correlações vistas.
2. Pode-se pensar também em relações positivas com o dados de comércio.

Estimação dos modelos

Para medir a acurácia dos modelos, as projeções são realizadas inicialmente dentro da amostra, ou seja, um sample de treino para a estimação dos modelos e um sample de treino para projeção e avaliações. O sample

de treino vai até fevereiro de 2019, ao passo que o sample de teste entre março de 2019 e fevereiro de 2021, ou seja, 24 meses foram reservados para comparar os valores previstos com o valor realizado.

A abordagem para se medir o efeito pretendido é através de estimações de modelos de regressões linear. Dezesseis diferentes especificações foram testadas, em que classifica-se o modelo de acordo com seu poder preditivo para dentro da amostra. Para tal, as medidas de comparação são RMSE, MAE, MPE e MAPE, os quais medem o desvio entre projetado e realizado.

O procedimento é feito da de acordo com os seguintes passos:

- Seleção das variáveis no training sample.
- Estimação da regressão no training sample.
- Criação da matriz para projeção dentro da amostra.
- Realizar a projeção dentro da amosta.
- Cálculo das medidas de acurácia do modelo e comparação.

Avaliação dos modelos estimados

A seguinte tabela apresenta os cinco melhores modelos no tocante a acurácia da previsão do consumo de energia industrial do sudeste. Estes são ordenados de acordo com as méticas, em que valores menores são preferíveis.

Metricas de avaliacao dos modelos estimados

.model	RMSE	MAE	MPE	MAPE
model15	215.6926	172.5929	-0.1336165	2.405873
model16	232.6798	174.4354	-0.2630219	2.438893
model7	245.8711	185.6872	-0.9243784	2.592447
model13	247.9913	195.3635	-0.7174970	2.731600
model3	257.5887	191.1845	-1.7827584	2.684581

Melhor modelo dentro da amostra

Dentre todos os modelos estimados, o modelo 15 apresentou ser o mais acurado para realizar a projeções para dentro da amostra, dadas as métricas vistas na tabela acima.

Pode-se observar através do output da regressão abaixo que a produção industrial do SE, o número de dias úteis e o próprio consumo de energia defasado são significativos para explicar o consumo de energia industrial no SE. Embora não significativos, a massa de rendimento real e PMC amploda mensal foram deixadas no modelo baseada puramente nas métricas para as previsões.

Observa-se assim, que fatores ligados à prórpia indústria são mais signifativos para explicar o consumo de energia do que fatores ligados à renda e ao comércio. No entanto, analisar estas última variáveis defasadas é relevante, uma vez que a indústria tende a responder posteriormente ao aumento da da demanda desencadeado pelo aumento da renda e da atividade comercial. Neste exercício, contudo, nota-se o comportamento autorregrssivo do processo, em que os efeitos consumo de energia tende a se propagar entre dois períodos.

Ademais, o R2 em torno de 0.80, mostra o quanto variações nas variáveis explicativas afetam variações no consumo de energia, o qual é relativamente significativo.

```
##
## Call:
## lm(formula = ind_se ~ massa_r + pim_se + du + pmc_a_se + lag_ind_se,
##     data = na.omit(train_in15))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -445.84 -121.44   -9.14  133.45  606.50
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.979e+03  7.713e+02   5.159 1.88e-06 ***
## massa_r     -1.440e-03  2.046e-03  -0.704    0.484
## pim_se       2.018e+01  3.633e+00   5.553 3.84e-07 ***
## du          -8.514e+01  2.065e+01  -4.123 9.34e-05 ***
## pmc_a_se     3.576e+00  3.770e+00   0.949    0.346
## lag_ind_se   4.616e-01  1.021e-01   4.522 2.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.5 on 77 degrees of freedom
## Multiple R-squared:  0.8145, Adjusted R-squared:  0.8024
## F-statistic:  67.6 on 5 and 77 DF,  p-value: < 2.2e-16
```

No tocante às análise quanto à possíveis problemas econométricos, os seguintes testes foram realizados:

Teste de heterocedasticidade

Não podemos rejeitar a hipótese nula de resíduos homocedásticos, tal como visto pelo teste Breusch-Pagan abaixo

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_15
## BP = 1.6907, df = 5, p-value = 0.8901
```

Teste de especificação do modelo.

A hipótese para a correta especificação do modelo não pode ser rejeitada pelo teste RESET abaixo.

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_15
## BP = 1.6907, df = 5, p-value = 0.8901
```

Teste de multicolinearidade

Tal como observado pelas estatística VIF, não há indicação de que os regressores sejam colineares.

```
##    massa_r     pim_se         du   pmc_a_se lag_ind_se
##   2.104449   2.858672   1.486460   2.659811   4.364706
```

Teste de autocorrelação residual

No entanto, observa-se a presença de autorrelação residual, o qual deve afetar a variabilidade dos intervalos de confiança dos parâmetros. A estimação da matriz de confiança robusta é indicada nesse sentido. Como estamos interessados em realizar projeções, relaxamos essa hipótese.

```
##
##  Durbin-Watson test
##
## data:  modelo_15
## DW = 2.2121, p-value = 0.7655
## alternative hypothesis: true autocorrelation is greater than 0
```

dwtest(modelo_15)

Melhores modelos

Em termos de acurácia, o modelo discutido anteriormente é preferível devido às métricas utilizadas para previsão do consumo de energia indutrial do SE dentro da amostra, tal como o RMSE. Baseado nessas mesmas estatísticas, os quatro seguintes modelos são apresentados a seguir:

2. Modelo 16

```
##
## Call:
## lm(formula = ind_se ~ pim_se + du + lag_ind_se, data = na.omit(train_in16))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -458.02 -143.26   -1.49  149.03  614.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3350.92452  542.95016   6.172 2.72e-08 ***
## pim_se        19.71494    3.19378   6.173 2.71e-08 ***
## du           -80.61050   18.40098  -4.381 3.60e-05 ***
## lag_ind_se     0.55572    0.06839   8.126 4.95e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.1 on 79 degrees of freedom
## Multiple R-squared:  0.8104, Adjusted R-squared:  0.8032
## F-statistic: 112.5 on 3 and 79 DF,  p-value: < 2.2e-16
```

3. Modelo 7

```
##
## Call:
## lm(formula = ind_se ~ massa_r + pim_se + du + pmc_a_se, data = train_in)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -506.85 -153.01    9.86  135.98  686.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.504e+03  6.040e+02  10.768  < 2e-16 ***
## massa_r     -3.667e-03  2.253e-03  -1.628    0.108
## pim_se       2.958e+01  3.332e+00   8.876 1.70e-13 ***
## du          -1.121e+02  2.196e+01  -5.104 2.24e-06 ***
## pmc_a_se     1.487e+01  3.172e+00   4.688 1.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.9 on 79 degrees of freedom
## Multiple R-squared:  0.7628, Adjusted R-squared:  0.7508
## F-statistic: 63.51 on 4 and 79 DF,  p-value: < 2.2e-16
```

4. Modelo 13

```
##
## Call:
## lm(formula = ind_se ~ massa_r + pim_se + du + pmc_a_se, data = train_in13)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -506.85 -153.01    9.86  135.98  686.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.504e+03  6.040e+02  10.768  < 2e-16 ***
## massa_r     -3.667e-03  2.253e-03  -1.628    0.108
## pim_se       2.958e+01  3.332e+00   8.876 1.70e-13 ***
## du          -1.121e+02  2.196e+01  -5.104 2.24e-06 ***
## pmc_a_se     1.487e+01  3.172e+00   4.688 1.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.9 on 79 degrees of freedom
## Multiple R-squared:  0.7628, Adjusted R-squared:  0.7508
## F-statistic: 63.51 on 4 and 79 DF,  p-value: < 2.2e-16
```

5. Modelo 3

```
##
## Call:
## lm(formula = ind_se ~ renda_r + massa_r + pim_se + du + pmc_a_se,
##     data = train_in)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -509.95 -141.99   12.52  131.25  516.93
##
## Coefficients:
```

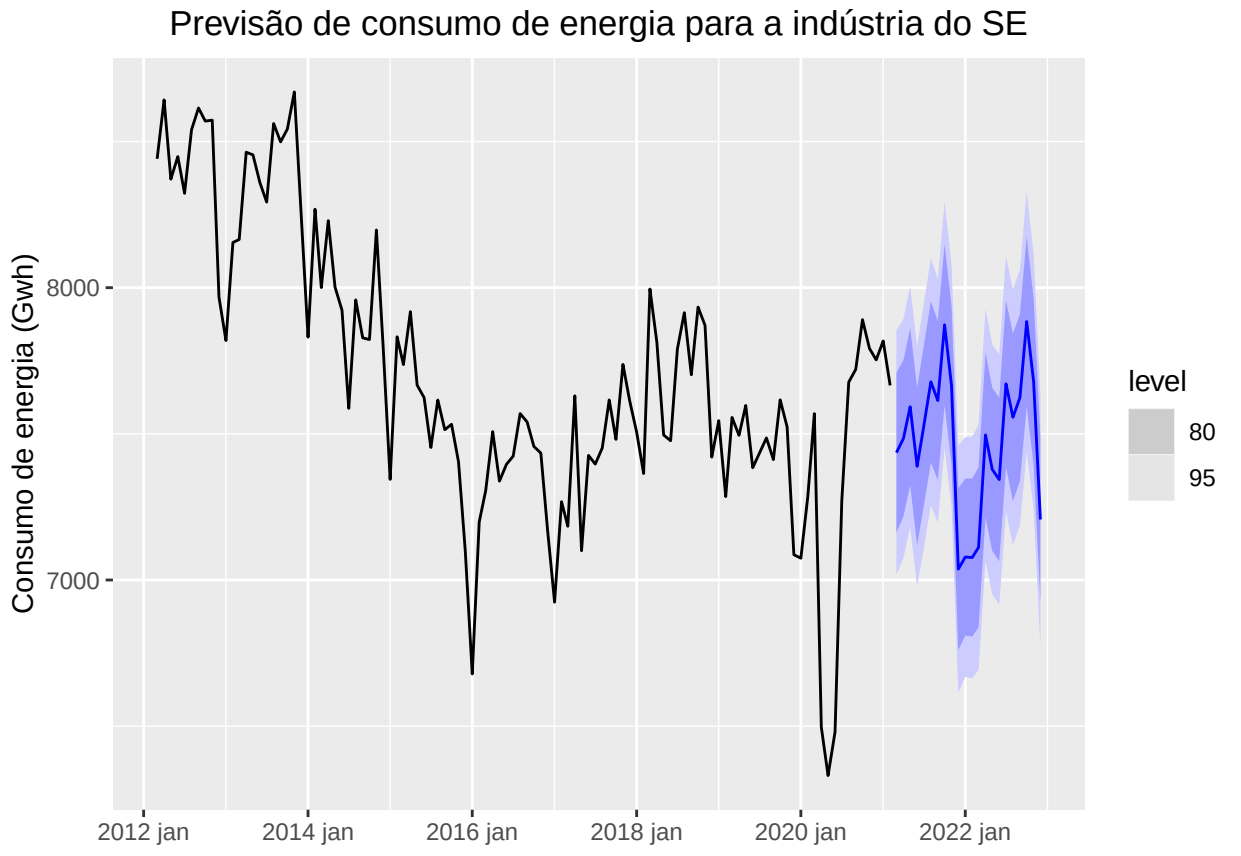
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.421e+03  6.576e+02   8.245 3.16e-12 ***
## renda_r      1.722e+00  5.231e-01   3.293 0.001492 **
## massa_r     -1.665e-02  4.478e-03  -3.718 0.000377 ***
## pim_se       2.784e+01  3.186e+00   8.738 3.48e-13 ***
## du          -1.107e+02  2.072e+01  -5.345 8.72e-07 ***
## pmc_a_se     1.516e+01  2.993e+00   5.066 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 216.8 on 78 degrees of freedom
## Multiple R-squared:  0.7918, Adjusted R-squared:  0.7784
## F-statistic: 59.31 on 5 and 78 DF,  p-value: < 2.2e-16
```

Previsão fora da amostra com o melhor modelo (M15)

A previsão fora da amostra para os 22 meses à frente fora feita da seguinte forma:

1. Reestima-se o modelo com a série completa.
2. Realiza-se a projeção com um modelo sem variáveis defasadas (modelo 7). A projeção para o consumo deste servirá de proxy na matriz de projeção fora da amostra.
3. Realiza-se o forecast do modelo 15 fora da amostra.

O seguinte gráfico apresenta a predição do modelo 15 para o consumo de energia elétric apara os próximos meses.



Considerações finais

Através dos exercícios 1 a 3 consegue-se mapear o fenômeno de interesse. Embora chegou-se a especificações para a projeção do consumo de energia industrial no SE, a análise não se encerra nesse ponto. Ganhos de eficiência no contexto preditivo podem ser alcançados ao testar outras especificações para o modelo linear, assim como da utilização de outros procedimentos tal como modelos random forest, apenas para citar um.

Essa sequência de análise define brevemente a forma de abordagem da questão. É necessário explorar as variáveis em contexto, bem como as relações entre as mesmas. Hipóteses iniciais acerca do fenômeno são feitas e em seguida testadas através de estimações. Por fim, projeções para variável de intersses são realizadas e comparações entre as estimações são feitas. Após este primeiro round, o ciclo se reinicia, uma vez que entendemos mais do fenômeno e somos mais ábeis para aperfeiçoar as análises.