Case Técnico - 4intelligence

Tiago Monteiro

Introdução

Este documento tem o intuito de estudar, modelar e predizer os valores do índice ABCR para veículos leves dos próximos anos, que mede o fluxo de carros nas rodovias e estradas com base na quantidade de veículos que passam por praças de pedágios no Brasil. Para tal, foram utilizadas 16 outras variáveis, além de técnicas e modelos de séries temporais, visando compará-los e decidir qual é o mais adequado para o problema.

Análise Exploratória

A Figura 1 mostra a tendência mensal do índice ABCR para veículos leves desde janeiro de 2010 até junho de 2023. Pelo gráfico, observa-se uma significativa sazonalidade anual, não se alterando muito ao longo dos anos. Observa-se também um crescimento em sua tendência, que pode ser melhor observada pela linha vermelha, representando as médias móveis anuais, até 2015, seguida de uma certa estabilidade, até 2020, quando houve uma rápida queda e quebra do padrão anterior, provavelmente causada pela pandemia de COVID-19, quando o fluxo de pessoas, e consequentemente de carros, diminuiu drasticamente.

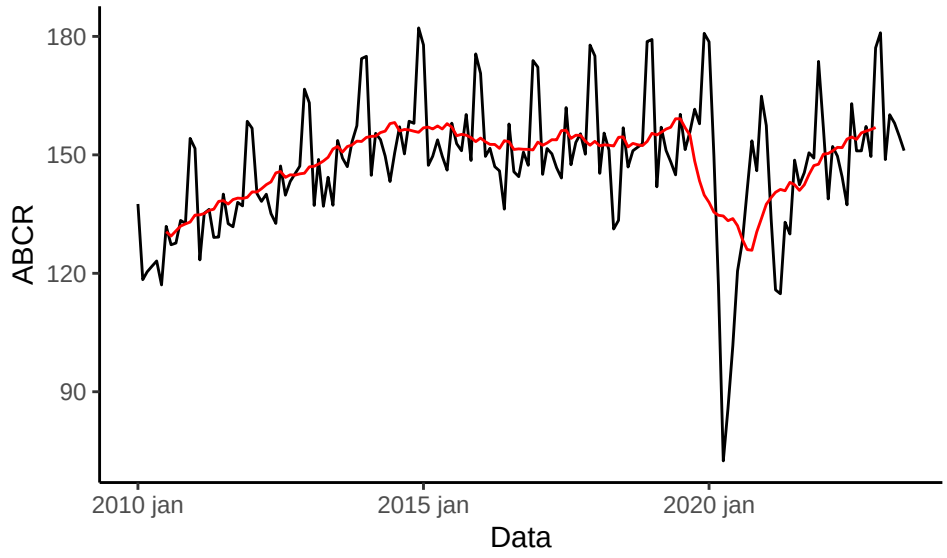

Figure 1: Tendência do índice ABCR (veículos leves) nos últimos anos

A Figura 2 mostra as funções de autocorrelação e autocorrelação parcial da série, deixando ainda mais claro os fortes padrões sazonais, mais precisamente, a cada 12 lags, ou seja, as observações tem o comportamento influenciado por suas observações referentes ao ano anterior.

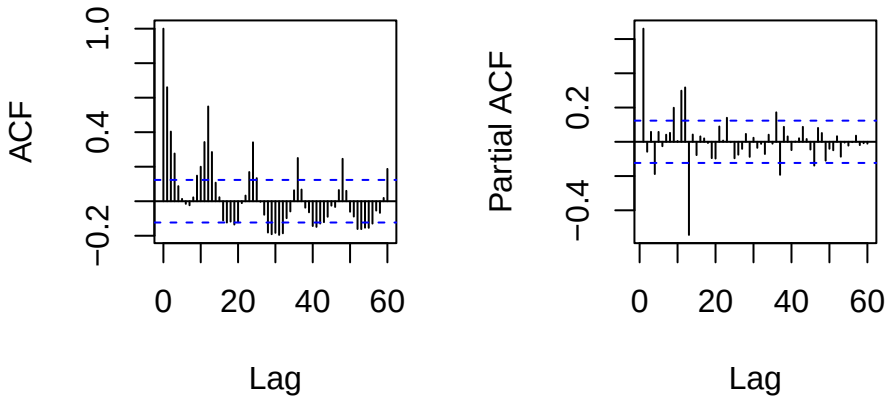

Figure 2: ACF e PACF de ABCR

A Figura 3 nos confirma o que foi visto nas análises anteriores, decompondo a série em suas componentes de tendência, sazonal e a parte aleatória, usando Loess, um método para estimar relações não lineares. Como dito anteriormente, vemos uma tendência de crescimento, com uma ruptura do padrão em 2020, e uma sazonalidade bem definida.

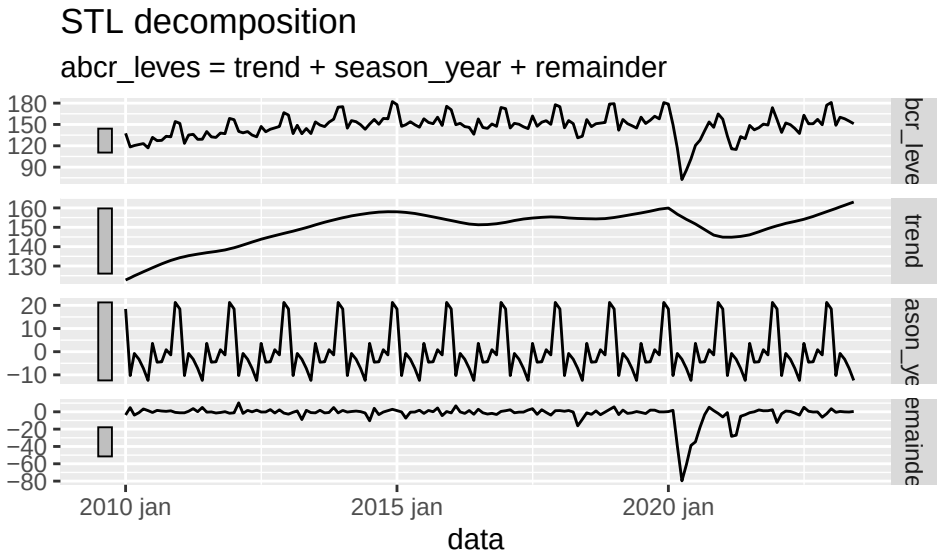

Figure 3: Decomposição STL de ABCR

Com o intuito de construir um modelo mais simples possível, a Figura 4, apesar de poluída pelo número de linhas, nos ajuda a entender como as variáveis interagem entre si pelas suas correlações, que são mais significativas a medida que a cor do quadrado de intersecção está mais próxima do vermelho (correlação positiva) ou do azul (correlação negativa). Observando a primeira linha de baixo, vemos que as variáveis que possuem maiores correlações com o índice ABCR são: `pmc_alimentos_bebidas`, `pmc-roupas_calcados`, `pmc_moveis_linha_branca`, `receita`, `reservas_internacionais` e `pmc_combustiveis_lubrificantes`, nesta ordem.

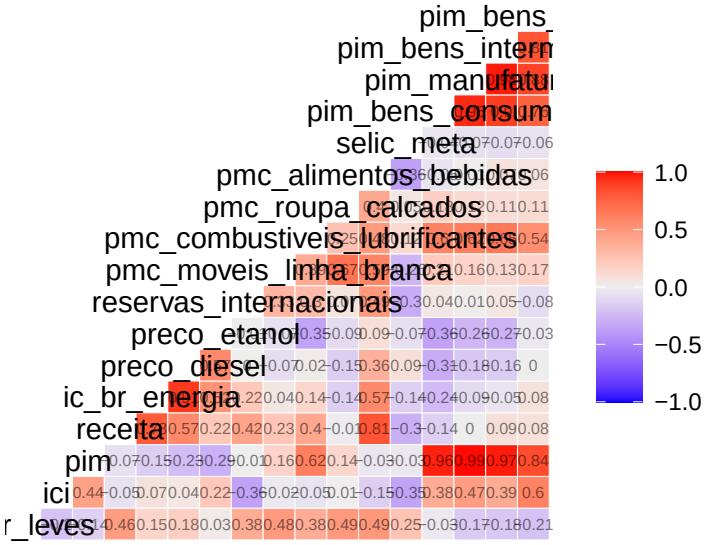

Figure 4: Matriz de correlação (todas as variáveis)

A Figura 5, assim como a Figura 4, mostra a matriz de correlação, porém restrita apenas às variáveis mais relevantes para o índice ABCR, além de mostrar gráficos de dispersão, densidades e os coeficiente de correlação de cada par de variável. Observa-se que as variáveis mais correlacionadas entre si são `receita` e `pmc_alimentos_bebidas`, com 0,81 de correlação, grande mas não a ponto de indicar uma possível multicolinearidade. Vemos também que algumas indicam uma possível relação não-linear em relação a ABCR, como `receita`, `pmc_combustiveis_lubrificantes`, `pmc_roupa_calcados` e `pmc_alimentos_bebidas`.

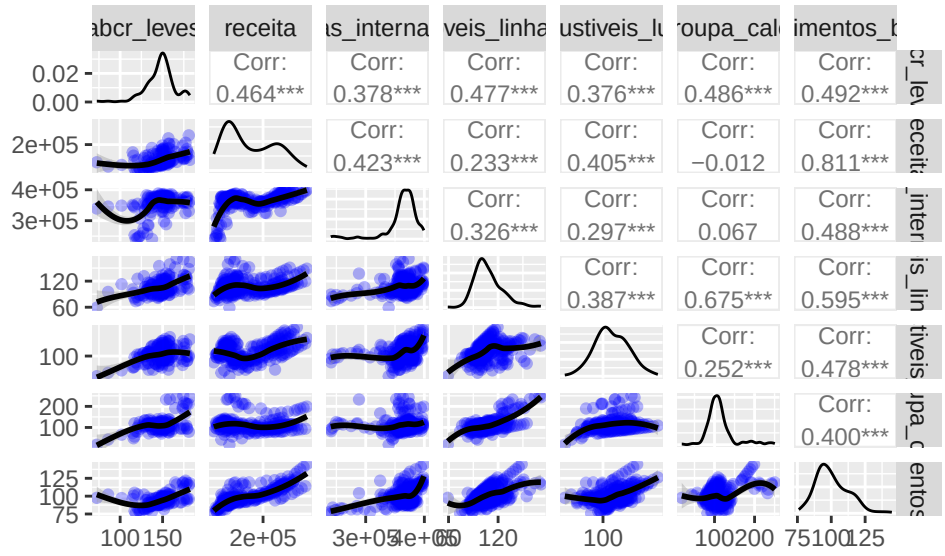

Figure 5: Matriz de correlação (variáveis mais relevantes)

A Figura 6 mostra os gráficos de tendência temporal das variáveis vistas no gráfico anterior, comparadas ao índice ABCR para veículos leves. Em todas, com exceção de `reservas_internacionais`, percebe-se uma sazonalidade anual, e com uma quebra na tendência por volta de 2020, com exceção de `pmc_alimentos_bebidas`, que se manteve estável.

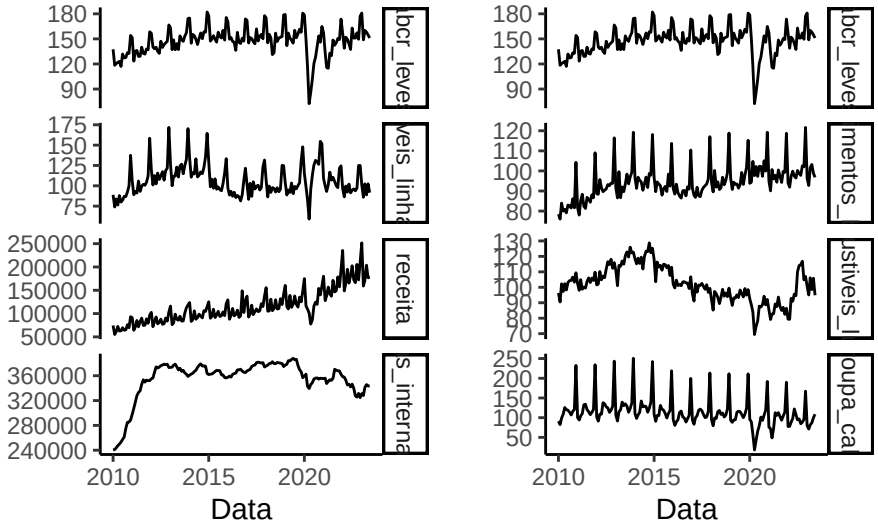

Figure 6: Tendências das variáveis comparadas ao índice ABCR

Resultados

Agora ajustaremos vários modelos com o intuito de escolher aquele com uma acurácia maior e, de preferência, que seja o mais simples possível, mantendo o padrão de qualidade. Como pudemos ver que o índice ABCR possui padrão sazonal, vamos ajustar um SARIMAX, que corresponde a um modelo de regressão com erros SARIMA (*Seasonal AutoRegressive Integrated Moving Average*).

Abaixo podemos ver os resultados de um ajuste com todas as 6 variáveis mais relevantes para a variável de interesse. Na Tabela 1 podemos ver os valores estimados das variáveis e dos parâmetros dos erros SARIMA(0,1,1)(0,1,1)[12], além de seus erros padrão, estatística t e o respectivo p-valor do Teste t, que verifica a significância do parâmetro a partir das hipóteses: H_0: o parâmetro é igual a 0 $vs.H_1$: o parâmetro é diferente de 0. Usando um nível de signficância de 5%, rejeitamos H_0 quando o p-valor for menor que 0,05. Na tabela podemos ver que, dado que todas as outras variáveis estão no modelo, não rejeitamos a hipótese de que `receita` e `pmc_moveis_linha_branca` são iguais a 0, ou seja, não são significativos para o modelo. Além disso, foi calculado o AIC (*Akaike Information Criterion*), que é uma medida de quão bom é o ajuste do modelo, levando em consideração sua complexidade, penalizando aqueles que possuem mais parâmetros, sendo preferível ajustes com menor AIC quando comparado com outros modelos.

Table 1: Coeficientes SARIMAX 1 (AIC = 935.72)

term	estimate	std.error	statistic	p.value
ma1	-0.6930	0.0129	-53.8635	0.000
sma1	-0.6497	0.0749	-8.6707	0.000
pmc_alimentos_bebidas	0.0485	NaN	NaN	NaN
pmc_roupa_calcados	0.5043	0.0434	11.6307	0.000
pmc_moveis_linha_branca	-0.0779	0.0687	-1.1331	0.259
receita	0.0001	0.0001	0.8369	0.404
reservas_internacionais	0.0001	NaN	NaN	NaN
pmc_combustiveis_lubrificantes	0.5655	0.1037	5.4510	0.000

Na Tabela 2 foi ajustado um modelo sem as 2 variáveis que não mostraram significância no ajuste anterior, e com os erros SARIMA(0,1,1)(1,1,3)[12]. Neste modelo vemos que `pmc_alimentos_bebidas` e `reservas_internacionais` também não se mostraram ter significância dado que as outras variáveis estão no modelo. O AIC do ajuste abaixo foi pouco maior do que o modelo anterior, indicando que este está levemente menos ajustado aos dados que o anterior.

Table 2: Coeficientes SARIMAX 2 (AIC = 935.86)

term	estimate	std.error	statistic	p.value
ma1	-0.6311	0.1286	-4.9095	0.0000
sar1	0.8094	0.1618	5.0032	0.0000
sma1	-1.5608	0.2578	-6.0531	0.0000
sma2	0.3975	0.2790	1.4248	0.1563
sma3	0.3201	0.1478	2.1659	0.0319
pmc_alimentos_bebidas	0.0331	0.3925	0.0842	0.9330
pmc_roupa_calcados	0.5200	0.0841	6.1819	0.0000
reservas_internacionais	0.0002	0.0003	0.5353	0.5933
pmc_combustiveis_lubrificantes	0.5343	0.1858	2.8752	0.0046

Na Tabela 3 foi ajustado um SARIMAX novamente sem as 2 variáveis que não se mostraram significativas no modelo anterior, com erros SARIMA(0,1,1)(1,1,3)[12], e como podemos ver abaixo, todas as variáveis

desde modelo se mostraram significativas, porém, com um ajuste pior do que os modelos anteriores, pois o AIC foi superior.

Table 3: Coeficientes SARIMAX 3 (AIC = 940.96)

term	estimate	std.error	statistic	p.value
ma1	-0.6580	0.0663	-9.9225	0.0000
sar1	0.7670	0.1637	4.6860	0.0000
sma1	-1.5609	0.2676	-5.8327	0.0000
sma2	0.4479	0.2695	1.6619	0.0986
sma3	0.2971	0.1396	2.1274	0.0350
pmc_roupa_calcados	0.5829	0.0472	12.3579	0.0000
pmc_combustiveis_lubrificantes	0.4658	0.1086	4.2912	0.0000

Como visto na Figura 5, algumas variáveis apresentam uma possível relação não linear em relação ao índice ABCR. Aplicando o logarítimo em `pmc_roupa_calcados`, geramos o modelo com coeficientes observados na Tabela 4, com erros SARIMA(0,1,1)(0,1,1)[12]. Vemos que o modelo possui todas as variáveis significativas, e ainda reduziu o AIC em relação aos modelos anteriores.

Table 4: Coeficientes SARIMAX 4 (AIC = 905.29)

term	estimate	std.error	statistic	p.value
ma1	-0.6578	0.0682	-9.6484	0e+00
sma1	-0.8499	0.0787	-10.7940	0e+00
log(pmc_roupa_calcados)	40.1207	2.7474	14.6033	0e+00
pmc_combustiveis_lubrificantes	0.3947	0.1050	3.7568	2e-04

A Tabela 5 acresce ao último ajuste, com mesmos erros SARIMA, a variável `reservas_internacionais`, e diminui o AIC em relação ao modelo sem a variável.

Table 5: Coeficientes SARIMAX 5 (AIC = 902.34)

term	estimate	std.error	statistic	p.value
ma1	-0.6611	0.0670	-9.8700	0.0000
sma1	-0.8340	0.0773	-10.7917	0.0000
log(pmc_roupa_calcados)	38.0995	2.8865	13.1994	0.0000
pmc_combustiveis_lubrificantes	0.4359	0.1048	4.1609	0.0001
log(reservas_internacionais)	34.1950	16.0563	2.1297	0.0348

A Tabela 6 remove a variável `log(reservas_internacionais)` e aplica log em `pmc_combustiveis_lubrificantes`, com mesmos erros SARIMA, resultando em um AIC pouco maior que o anterior, mas com todas as variáveis sendo significativas.

Table 6: Coeficientes SARIMAX 6 (AIC = 903.25)

term	estimate	std.error	statistic	p.value
ma1	-0.6623	0.0695	-9.5224	0e+00
sma1	-0.8588	0.0815	-10.5355	0e+00
log(pmc_roupa_calcados)	38.5192	2.9102	13.2360	0e+00

term	estimate	std.error	statistic	p.value
log(pmc_combustiveis_lubrificantes)	42.6988	10.5203	4.0587	1e-04

A Tabela 7 resume tudo que foi visto acima, mostrando o número de parâmetros e o AIC de cada um dos modelos ajustados. O modelo com menor AIC é o modelo 5, que possui 5 parâmetros, mas o modelo 6 conseguiu se ajustar quase tão bem, mas com 1 parâmetro a menos.

Table 7: AIC e número de parâmetros por modelo

Modelo	Par.	AIC
SARIMAX 1	8	935.72
SARIMAX 2	9	935.86
SARIMAX 3	7	940.96
SARIMAX 4	4	905.29
SARIMAX 5	5	902.34
SARIMAX 6	4	903.25

Analisando os resíduos dos 3 modelos com menores AIC's, vemos na Figura 7 a tendência, autocorrelação e distribuição dos resíduos do SARIMAX 4, que indica que os resíduos são estacionários, sem possuir autocorrelação significativa entre seus lags, com distribuição centrada em 0 mas não muito próxima de uma normal, pois indica uma certa assimetria.

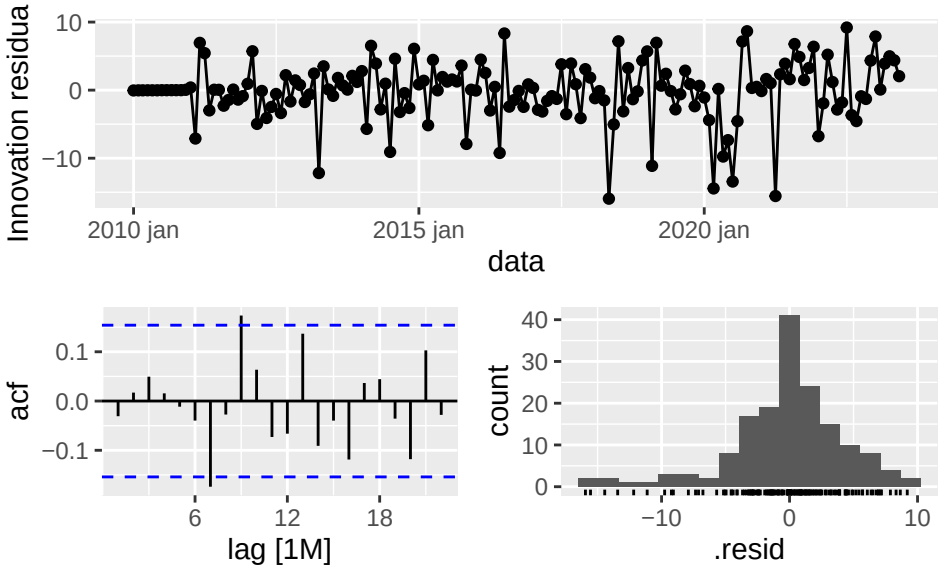

Figure 7: Resíduos do SARIMAX 4

A Figura 8 mostra os mesmos gráficos para o SARIMAX 5, indicando estacionaridade dos resíduos, sem autocorrelação significativa aparente e com distribuição similar a anterior.

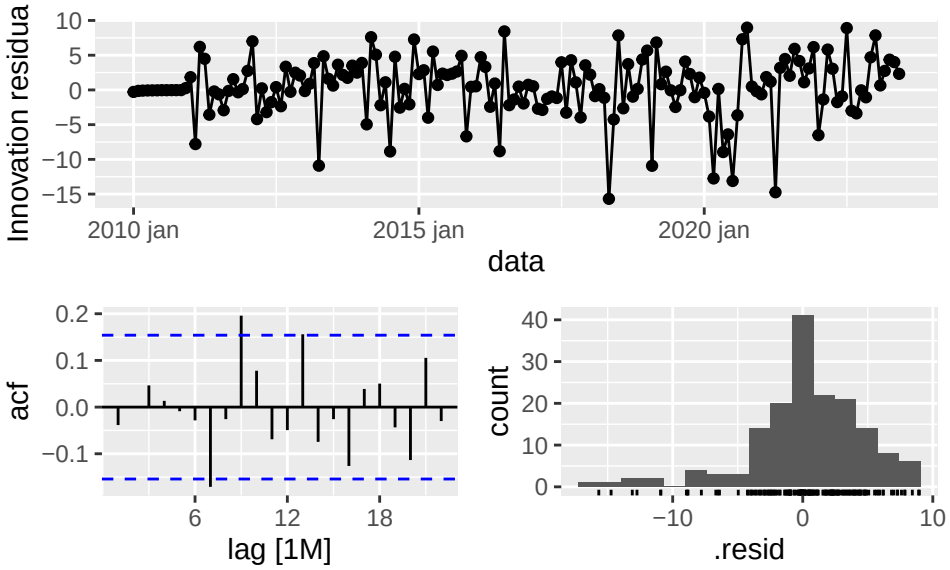

Figure 8: Resíduos do SARIMAX 5

A Figura 9 também mostra os resíduos do ajuste SARIMAX 6, com características similares aos modelos anteriores, com um distribuição mais distante de uma normal.

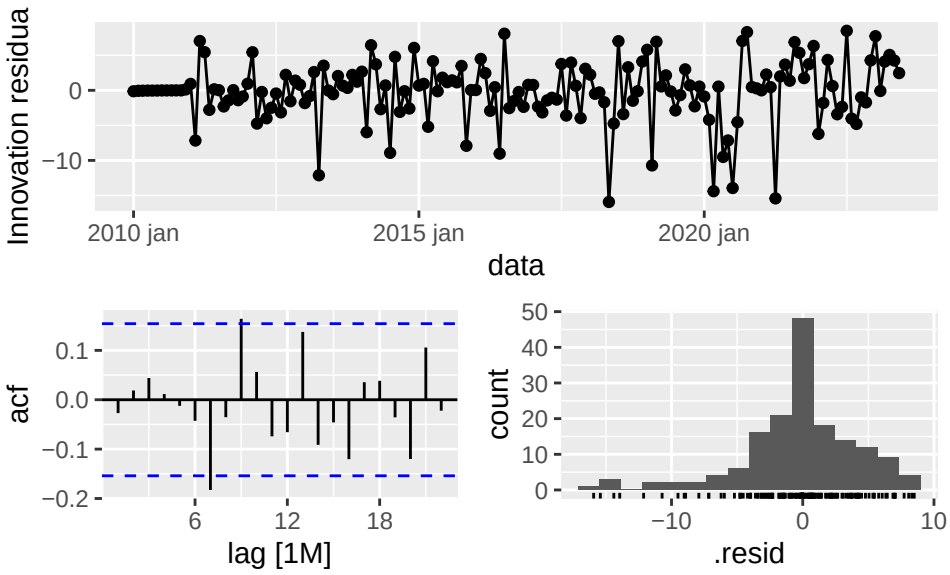

Figure 9: Resíduos do SARIMAX 6

Na Tabela 8 vemos os p-valores dos testes KPSS (que testa H_0: os resíduos são estacionários $vs. H_1$: os resíduos não são estacionários), Shapiro-Wilk (que testa H_0: os resíduos têm distribuição normal $vs. H_1$: os resíduos não têm distribuição normal), e Box-Pierce (que testa H_0: os resíduos são não-autocorrelacionados $vs. H_1$: os resíduos são autocorrelacionados). Com nível de significância de 5%, rejeitamos H_0 quando obtivermos um p-valor $< 0{,}05$. Como podemos ver, todos os modelos rejeitam normalidade dos resíduos, indicando que não capturaram bem a dinâmica dos dados.

Table 8: Testes estatísticos dos modelos

Modelo	KPSS	Shapiro.Wilk	Box.Pierce
SARIMAX 4	0.1	0	0.11031
SARIMAX 5	0.1	0	0.05196
SARIMAX 6	0.1	0	0.11553

Realizando validação cruzada para os 3 modelos, ou seja, separando so dados em treino (80%) e teste (20%) de maneira que é feita a previsão 1 passo a frente e depois retreinando o modelo repetidamente, até completar as observações, posteriormente é calculado o erro de predição comparado com valores observados. A Tabela 9 mostra a raíz do erro quadrático médio de cada um dos modelos, mostrando que o SARIMAX 5 foi o que teve a maior acurácia nas predições, se mostrando o mais adequado para ser colocado em produção.

Table 9: RMSE das previsoes

Modelos	RMSE
sarimax04	5.292482
sarimax05	5.057089
sarimax06	5.278949

Conclusões

Dos modelos propostos anteriormente, o modelo SARIMAX 5, com as variáveis `log(pmc_roupa_calcados)`, `pmc_combustiveis_lubrificantes` e `reservas_internacionais`, com erros SARIMA(0,1,1)(0,1,1)[12] se mostrou o com maior acurácia dentro dos modelos mais simples, sendo o mais adequado para realizar predições. A Figura 10 mostra os valores preditos para ABCR dos próximos meses até 2030, junto com um intervalo de predição de 95%, para o cenário proposto das covariáveis do modelo.

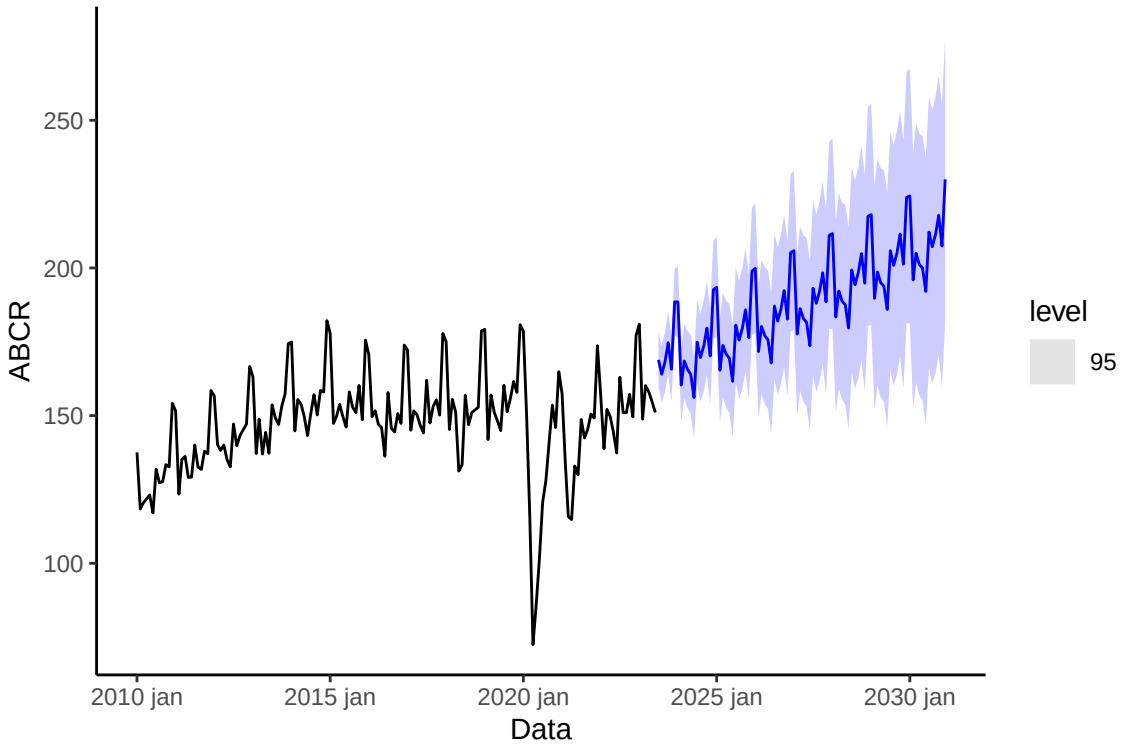

Figure 10: Predição do SARIMAX 5 até dez 2030

A interpretação do modelo seria mais apropriada se os resíduos tivessem distribuição normal, mas podemos entender a relação de `abcr_leves` com as demais variáveis da seguinte maneira:

$$ABCR_t = 37,8 \times log(X_{1t}) + 0,44 \times X_{2t} + 34,2 \times log(X_{3t}) + \eta_t$$

em que X_{1t} representa `pmc_roupa_calcados`, X_{2t} representa `pmc_combustiveis_lubrificantes`, X_{3t} representa `reservas_internacionais`, e η_t são os erros SARIMA.

Dessa forma, se aumentarmos o indicador referente ao comportamento do comércio de roupas e calçados em 1 unidade, é esperado que ABCR aumente em $0,378\%$, quando as outras variáveis são mantidas fixas. Se aumentarmos o indicador referente ao comportamento das vendas de combustíveis e lubrificantes em 1 unidade, é esperado que ABCR aumente em 0,44 unidades, quando as outras variáveis estão constantes. Finalmente, quando os ativos externos disponíveis do país aumentam em 1 unidade, é esperado um aumento de $0,342\%$ no índice ABCR para veículos leves.