Hackathon JPA Agro 2021

Data Science Research Group - DSRG
Universidade Federal de Lavras - UFLA

Identificação da equipe

Nome da equipe: SufCy team (Sufficiency principle).
Integrante 1: Carlos Antônio Zarzar.

Descrição da solução

Buscando um equilíbrio entre as informações técnicas e a própria intuição da solução proposta para esse problema específico no desafiado no Hackaton JP Agro 2021, serei breve e simples nessa sessão.

Os desafios ao longo desse projeto foram surgindo à medida que fui avaliando os dados. Primeiramente observei que os dados não estavam completos, existiam dias com informações faltantes (395 dias) sobre o valor da polpa cítrica comercializada pela empresa. Portanto o primeiro desafio era saber se ajustava os dados em um modelo ignorando esses dias faltantes ou faria uma interpolação nos mesmo. Caso a última opção fosse o caminho definido, o segundo passo era saber qual interpolação seria melhor indicado para esse problema?

Seria mais prudente sugerir qualquer método de interpolação ou até mesmo qualquer modelo para predição dos dados futuros após conhecer as principais características dessa série temporal (valor de venda da polpa cítrica), afinal quanto mais informações você têm sobre o fenômeno mais fácil será as tomadas de decisões para o proposto objetivado. Ao menos é isso o que os cientistas de dados mais aconselham na sua profissão e considerando uma análise como essa, não foge a regra.

Devido as características da série, o padrão encontrado nos dados e meus recentes estudos optei por um modelo baseado em Séries Temporais Estruturais Bayesiana. Modelo derivado da especificação de Tendência Linear Local da família de modelos Bayesianos. Sua implementação foi utilizando linguagem Stan com interface ao R *software*. Stan é uma linguagem de programação com interface em R, Python e outros voltado para inferência Bayesiana escrito em C++ que se utiliza de rápida velocidade e eficiência de uma programação de baixo nível ao mesmo tempo que se intersecta com uma linguagem de programação orientada à objetos, bem avançada no processamento de dados e em análises estatísticas, o programa R.

Muitas análises e dúvidas surgiram ao longo do caminho, porém, considerando o tempo do desafio, a limitação das informações e a restrição do meu conhecimento sobre os métodos e as vastas ferramentas utilizadas em séries temporais, busquei o meu melhor na predição desses dados e na realização desse desafio.

1. Entendimento do negócio

Primeiramente, busquei entender a empresa JPA Agro e o grupo Grupo JPA, procurando no seu site profissional informações mais detalhadas sobre o negócio e consequentemente sobre a necessidade da empresa. Em seguida uma busca sobre o principal produto (polpa cítrica) da empresa para entender o mercado desse produto, região de produção, origem e destino desse nicho de mercado entre outras informações. Ao longo dos estudos, antes mesmo de conhecer os dados, compreendi que tanto o mercado do milho quanto a valorização do dólar americano podem estar diretamente ou indiretamente relacionados ao valor de venda do subproduto derivado das frutas cítricas. Dessa forma, séries temporais para o valor do dólar americano e para o valor do milho no mercado foram adquiridos através do site da CEPEA (Esalq/USP), banco de dados público disponível.

2. Pré-processamento dos dados

Todo processamento e análises foram realizadas em linguagem de programação R (R core team 2021). Após importação dos dados e declaração das variáveis de forma correta para o programa, foram feitas inspeções para detectar problemas nos mesmos. Encontrei dados não disponíveis ou pulos em alguns dias na coleta da informação (395 dias sem informação, dados faltantes) através do sumário (*summary()*) e principalmente coincidindo o vetor criado com as datas (dias/mês/ano) completas (sequência da data que eu esperaria se os dados fossem coletados todos os dias) com o vetor das datas reais nos dados fornecidos, série temporal do valor da polpa cítrica ("dataset_train.csv"). Depois foi correlacionada esses dias faltantes com os dias da semana para saber se havia algum padrão já que eram números significativos de informações não coletadas. Porém, não foi identificado o motivo da ausência de informação nesses específicos dias através dos dados, mostrando-se aleatório a ocorrência nos dias da semana desses dados que não estavam disponíveis.

- 1° Primeiro desafio

O primeiro desafio que me deparei nessa análise era saber o que fazer com esses dados faltantes (dias não coletados) para alcançar o principal objetivo fornecido pela empresa (predição de 30 dias à frente dos dados fornecidos) (Figura 1). Poderia ignorar praticamente 1 ano de informação não coletada, ou fazer uma interpolação entre eles. Se seguisse esse último caminho, qual interpolação seria mais indicado? Então conclui que se separasse 30 dias do banco de dados de treinamento fornecido, poderia quantificar métricas de predições e assim, em função de melhores predições poderia definir qual desses dois caminhos deveria seguir. Métodos gráficos e análises realizadas procurando algum padrão sobre esses dados ausentes podem ser encontrados no *script* fornecido (arquivo: Script_SufCy.R)

- Características da série

Antes de sugerir qualquer método de interpolação foi importante conhecer as características da série temporal fornecida (valor da venda da polpa cítrica). Análise para identificar a tendência da série indicaram que se tratava de uma série estacionária, que não possui sazonalidade e o ruído branco caracterizado por ser heterocedástico (ou seja variância não constante) e não independentes. Assim sendo, devido uma memória forte

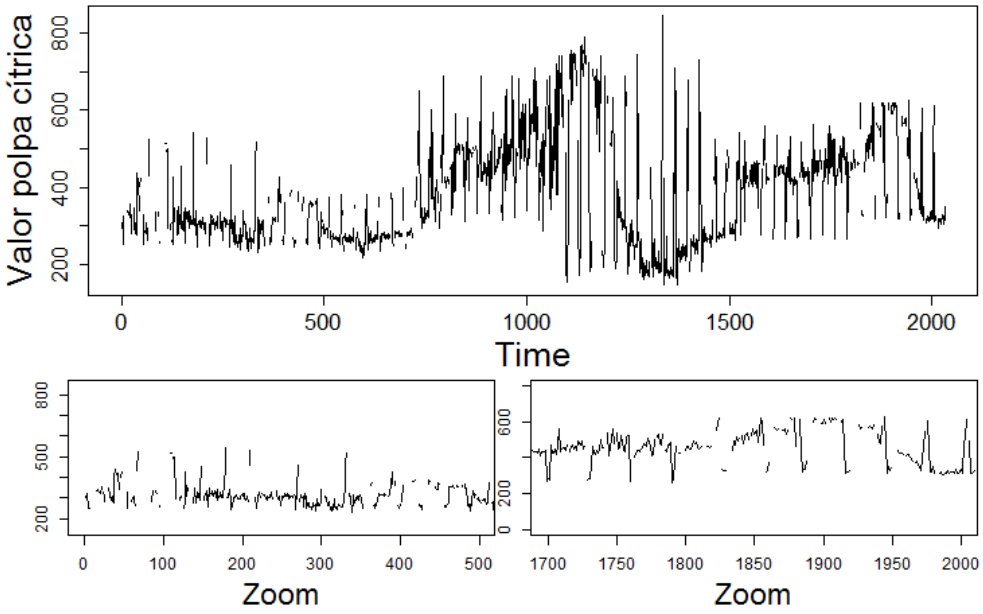

Figura 1: Marcantes dados faltantes impressos no gráfico da Série temporal do valor da polpa cítrica e em seu respectivo zoom para facilitar a visualização

da série e as demais características descritas, me direcionaram para uma interpolação *random walk* sobre os dados faltantes.

Nos modelos de regressão dinâmica temporal, a variável dependente é explicada por seus valores defasados (lag), pelos valores atuais, sazonalidade caso ocorra, tendência, periodicidade e variáveis causais ou exógenas. Um cuidado redobrado com as variáveis exógenas em séries temporais pois séries com tendências regredida contra outras séries temporais, frequentemente, revelam um relacionamento forte, mas espúrio. Principalmente se for utilizado a correlação de Pearson, pois é um tipo de análise de correlação entre duas variáveis de forma linear uma da outra. Enfatizando que nem sempre a correlação é uma causalização, o que significa que só porque duas coisas parecem estar relacionadas entre si não significa que uma causa a outra. No caso de séries temporais duas séries sempre estão relacionadas pelo tempo (t) e sua tendência induz uma correlação que muitas vezes não é verdadeira. Para realmente saber se as variações em uma série estão correlacionadas com as variações em outra é indicado modelar a tendência em cada série temporal e usar esse modelo para removê-la. Em outros casos não paramétricos não requerem modelagem, apenas tomar sucessivas diferenças da série original até encontrar uma série estacionária, removendo por completo qualquer tendência existente. A função de correlação cruzada ajuda a determinar quais *lags* da série temporal X preveem o valor da série temporal Y. No entanto, se uma das séries tiver autocorrelação, ou as duas séries compartilharem tendências comuns, é difícil identificar relações significativas entre a duas séries temporais. Nesse projeto as duas variáveis que se acreditava ter grandes relações apresentaram pouca correlação cruzada entre as series temporais sem tendência do valor do milho e do valor do dólar americano contra série do valor da polpa cítrica (Figura 2 e 3).

- Iterpolação *random walk*

A interpolação *random walk* consiste em uma amostragem a partir de uma distribuição aleatória, aqui definida como uma normal devido o padrão do ruído branco. Então

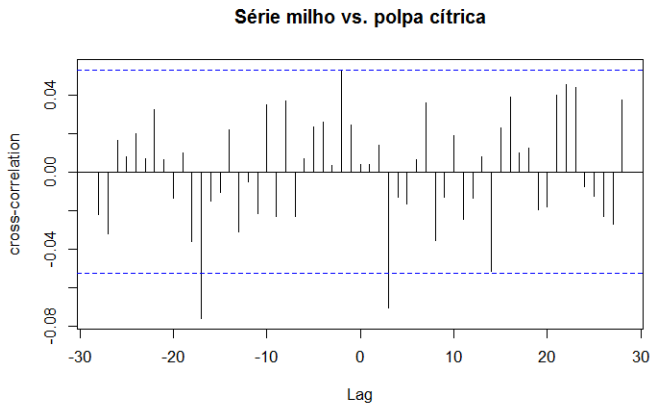

Figura 2: Baixa correlação cruzada entre a série temporal sem tendência do valor de milho versus a série do valor da polpa cítrica

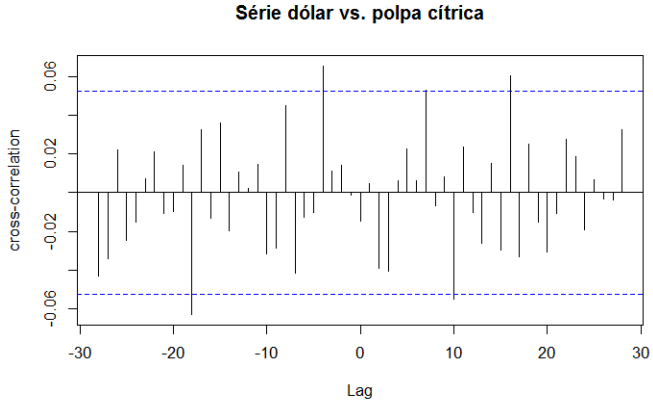

Figura 3: Baixa correlação cruzada entre a série temporal sem tendência do valor do dólar americano versus a série do valor da polpa cítrica

a cada dado faltante sobre a série defasada de um *lag* ($y_{\Delta t} = y_t - y_{t-1}$) se obtém as estatísticas (média e desvio padrão) da série do início até o momento anterior do dado faltante, amostra-se um valor aleatório oriunda de uma distribuição de probabilidade $Normal(\mu_{y_{\Delta t}}, \sigma_{y_{\Delta t}})$ e interpola no específico dado faltante do referente momento (t) e em seguida continua o algoritmo até acabar todos dados faltantes da série. Após a interpolação de todas as defasagens da série, adiciona-se os dados faltante na série original (não defasada) interpolando os dados não existente (Figura 5).

3. Modelos

O modelo ajustado foi baseado em Séries Temporais Estruturais Bayesianas. Este modelo deriva da especificação de Tendência Linear Local da família de modelos Bayesian Structural (*Time Series*), implementado em linguagem Stan. Posteriormente, foram feitas predições de 30 dias como descrito por Scott e Varian (2014) para previsões de modelo de série temporal estrutural Bayesiana de tendência linear local.

Essa proposta de modelo esboça uma maneira flexível de descrever a tendência de uma variável de interesse μ_t de uma série temporal. Se supõe que as observações da série temporal são a soma de vários componentes como a tendência, sazonalidade, quais quer

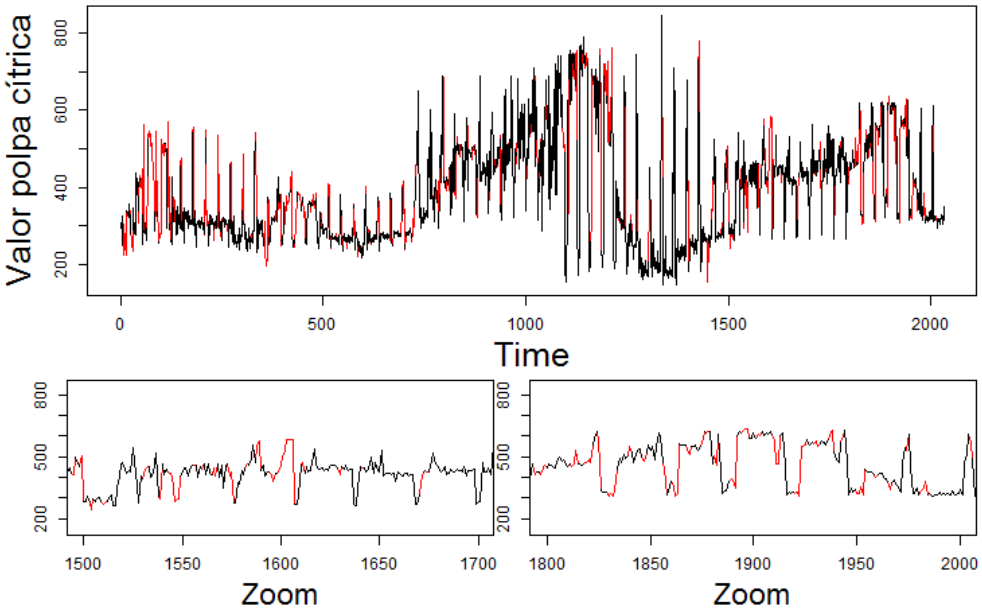

Figura 4: Método de *random walk* na interpolação dos dados faltates e seus respectivos zooms para facilitar a visualização

regressores e ruído:

$$Y_t \sim Normal(\lambda_t, \sigma_Y); \qquad \lambda_t = \mu_t + \tau_t + \beta^T x_t$$

Na série temporal do valor de venda da polpa cítrica não há indícios da componente sazonal, portanto foram retirados do modelo. E os regressores (no nosso caso relação do valor da polpa cítrica com o dólar americano e o valor do milho no Brasil) são certamente interessantes, mas como na análise não apresentaram fortes correlações cruzadas com a série de interesse acabou não sendo implementado nesse desafio proposto. Espera-se projetar a tendência de μ_t que é uma função linear do tempo, o que se equivale dizer $\mu_t = \mu_{t-1} + \delta_t$ considerando janelas *lag* iguais a um. Portanto, espera-se que a tendência na próxima etapa (t) seja a tendência anterior mais uma diferença que muda suavemente δ_t. Adicionando algum ruído estocástico a esta relação, obtemos:

$$\mu_t = \mu_{t-1} + \delta_t + \epsilon_{\mu,t}; \qquad \epsilon_{\mu,t} \sim Normal(0, \sigma_\mu)$$

Se a mudança δ_t a cada tempo t da seria é constante em δ, então a tendência seria igual a uma regressão linear em relação ao tempo. Caso contrário, se δ_t varia ao longo do tempo, a tendência pode se tornar uma função bastante flexível do tempo, que é desejado no nosso caso, embora seja localmente linear no tempo. Assim, o modelo se resumo em:

$$\delta_t = \delta_{t-1} + \epsilon_{\delta,t}; \qquad \epsilon_{\delta,t} \sim Normal(0, \sigma_\delta)$$

Definindo-se distribuições a priori de probabilidade para σ_Y, σ_μ e σ_δ como seminormal (Half-Normal), porém, escala diferentes. Diferente de Taylor e Letham (2018) que em sua especificação de modelo para inferência Bayesiana, eles escolheram $\delta_j \sim Laplace(0, \tau)$ como distribuição a priori para os valores da taxa de mudança a cada ponto no tempo t. Isso significa que as alterações são geralmente mais próximas de zero, mas podem ser grandes de tempos em tempos.

4. Avaliação da solução

O modelo foi avaliado retirando 30 dias do banco de dados de treinamento fornecido e fazendo a predição para confrontar com os dados reais e calcular o índice RMSE (Root Mean Square Error). Além disso, foi dividido a série na metade e ajustados os dados, em seguida realizada a predição de 30 dias após a metade dos dados e novamente utilizado os dados reais para calcular o RMSE. Com esse índice percebi que a interpolação melhorou levemente a acurácia do modelo e então pude escolher qual o método e os caminhos a seguir a partir de uma metodologia definida. E utilizando diferentes momento da série na modelagem e predição, definimos assim através do índice, um método mais robusto e com menor tendência a um viés regional da série temporal estudada.

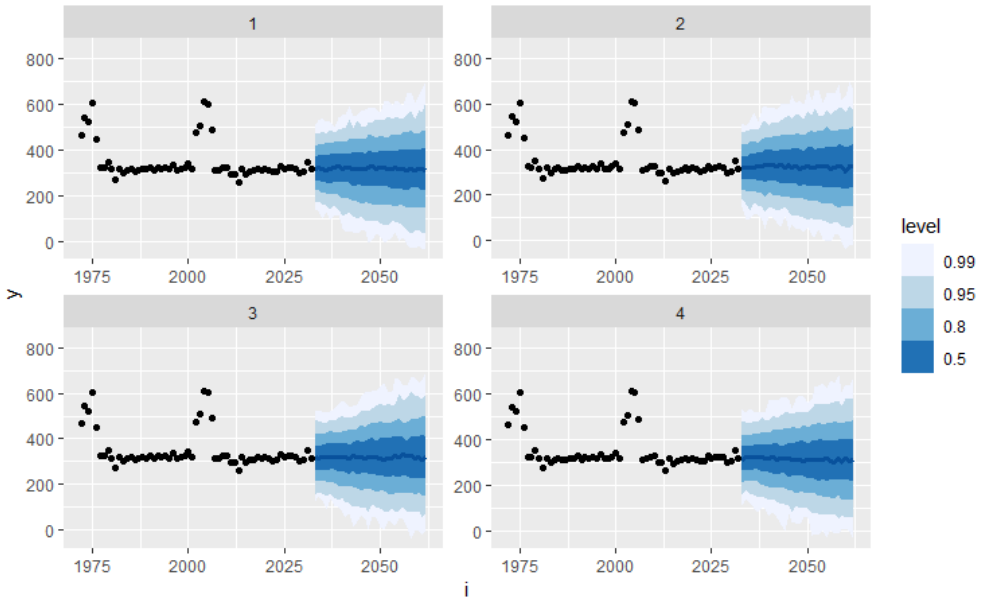

Figura 5: Predição da distribuição a posteriori com níveis de credibilidade do valor da polpa cítrica de 30 dias após a data final da série temporal fornecida para 4 cadeias MCMC e com 1.000 iterações cada para o modelo Estrutural Bayesiano

Todas as análise mais detalhas e gráficos informativos se encontram no script fornecido assim como os dados de predição solicitado em arquivo csv em formato de txt. Embora sempre aja limitações de tempo, informações, restrição do meu conhecimento sobre os métodos e a ignorância das vastas ferramentas utilizadas em séries temporais, busquei o meu melhor na predição desses dados e na realização desse desafio e espero ter alcançado um nível satisfatório de predição desejado.

Referências

[1] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[2] Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.

[3] Scott, S. L., and Varian, H. R. (2014). Predicting the present with Bayesian structural time series. International Journal of Mathematical Modelling and Numerical Optimisation, 5(1–2), 4–23.

[4] Taylor, S. J., Park, M., and Letham, B. (2017). Forecasting at scale at Facebook. Menlo Park, California, United States.

[5] Savage, J. (2019). Balanced budgets and American politics. Cornell University Press.