Estudo de correlação e causalidade entre a cotação internacional da soja e a taxa de câmbio entre dólar e yuan *

Matheus Augusto de Bortoli de Campos† Prof. Me. Jorge Luis Boeira Bavaresco‡

2022

Resumo

A cotação da soja é regulada pela oferta e demanda do produto no mercado internacional, o que torna difícil a previsão de seu valor. Com os avanços recentes no campo da ciência de dados, um ramo em crescimento é o da análise quantitativa, um tipo de análise que se beneficia de estudos de correlação e causalidade, fazendo uso dos mesmos no auxílio à construção de modelos matemáticos que visam prever comportamentos de séries temporais. A China é o maior importador de soja do mundo e costumava adotar medidas que fixavam a variação da cotação de sua moeda, o yuan, em relação ao dólar. Contudo, o país vem flexibilizando suas políticas monetárias nos últimos anos e permitindo a flutuação dessa taxa de câmbio. Dessa forma, é interessante que sejam estudados os impactos dessas flutuações no preço da soja. O presente estudo busca estudar relações de correlação e causalidade entre a cotação internacional da soja, em dólar, e a taxa de câmbio entre dólar e yuan. Os conjuntos de dados foram obtidos, tratados e estudados por meio de métodos de análise estatística que apresentaram indícios de uma correlação negativa entre as séries temporais, mas uma ausência de causalidade entre elas.

Palavras-chaves: linguagem R, coeficiente de spearman, causalidade de granger

*Trabalho de Conclusão de Curso (TCC) apresentado ao Curso de Bacharelado em Ciência da Com-putação do Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense, Campus Passo Fundo, como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação, na cidade de Passo Fundo, em 2022.

†Bacharelando no curso de Ciência da Computação no Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense, Campus Passo Fundo.

‡Orientador do trabalho. Mestre em Computação Aplicada. Professor do Curso de Bacharelado em Ciência da Computação no Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense, Campus Passo Fundo

1 Introdução

A produção de soja está entre as atividades econômicas que mais apresentaram crescimento nas últimas décadas. Isso se deve à diversos fatores, tais como o desenvolvimento e estruturação de um sólido mercado internacional relacionado ao comércio de produtos do complexo agroindustrial da soja, a consolidação da oleaginosa como importante fonte de proteína vegetal, especialmente para atender demandas crescentes dos setores ligados à produção de produtos de origem animal, bem como a criação e oferta de novas tecnologias, que viabilizaram a expansão da exploração da commodity nas mais diversas regiões do mundo (HIRAKURI; LAZZAROTTO, 2014).

Entre os diversos usos da soja, se destacam a produção de ração, a produção de óleo vegetal, e a produção de biodiesel, que é realizado após a extração do óleo vegetal. Atualmente, três países lideram a produção de soja no mundo, sendo eles o Brasil, os Estados Unidos e a Argentina, nessa ordem. Segundo dados coletados na Statista (2022b), o Brasil produziu um total de 139 milhões de toneladas de soja na safra 2021/2022, seguido de 120,71 milhões, provenientes dos EUA e 46,5 milhões produzidas na Argentina.

O maior importador de soja do mundo é, atualmente, a China. Segundo dados coletados em Statista (2022a), na safra 2021/2022 a China importou 97 milhões de toneladas de soja, mais de seis vezes a quantia importada pela União Européia como um todo, que ficou em segundo lugar, importando um total de 14,8 milhões de toneladas de soja.

A soja pertence à classe de ativos das commodities, que segundo Stonex-Brasil (2021), são produtos amplamente negociados no mercado internacional, ou seja, que possuem uma ampla gama de produtores e compradores, e não são facilmente perecíveis. Segundo o mesmo, a maior parte das commodities são matérias-primas, usadas para produção de outras mercadorias, e possuem baixo ou nenhum grau de industrialização.

Devido às características citadas, as commodities normalmente possuem cotações internacionais amplamente difundidas e muitas vezes são negociadas em bolsas de mercadorias (STONEX-BRASIL, 2021). Dessa forma, não é o produtor que define o preço do produto a ser vendido, mas sim uma complexa dinâmica mundial de oferta e demanda. Por isso, torna-se difícil prever o preço do grão de forma a definir o momento óptimo para sua venda ou compra, visando o lucro ou mesmo proteção financeira.

Esse momento ideal para a compra ou venda da soja costuma ser do interesse dos produtores rurais, considerando que os mesmos costumam investir um capital significativo para realizar o plantio e a colheita do grão, e que buscam, a depender de sua cotação, bem como previsão de variação do preço do mesmo, realizar determinadas operações, tais como a venda da colheita, o armazenamento da mesma (para venda posterior, quando o preço estiver mais favorável) ou mesmo operações no mercado financeiro. Da mesma forma, esse momento oportuno também é do interesse das partes compradoras de soja, tais como grandes indústrias alimentícias, que buscam comprar produtos a preços competitivos. Essas operações visam otimizar o lucro das partes interessadas, garantindo assim a viabilidade financeira de seus negócios.

Uma das formas de se otimizar as probabilidades de se obter lucro em operações envolvendo commodities, é através do uso de análise quantitativa. Com os recentes avanços nas áreas de

engenharia de finanças, computação e algoritmos numéricos, que possibilitam análises de Big Data cada vez mais rápidas e precisas, as estratégias de análise quantitativa tem ganhado cada vez mais popularidade, se tornando uma importantes no contexto de fundos de investimentos, e gestoras de ativos (LO; MAMAYSKY; WANG, 2000)(MAGMA-CAPITAL-FUNDS, 2022).

Baseando-se na idéia de que preços passados contém informações para prever retornos futuros (LO; MAMAYSKY; WANG, 2000), uma das possibilidades que a análise quantitativa traz é utilizar relações de correlação e causalidade entre séries temporais distintas para realizar previsões nas mesmas. Dessa forma, pode-se utilizar de variações em um ou mais conjuntos de dados para prever tendências em outro conjunto, desde que tenhamos sets de dados com relações estatisticamente comprovadas.

Visto que o dólar é a moeda internacionalmente aceita como padrão para a negociação da soja, e que a China é o maior importador de soja do mundo, título que mantém por uma grande margem, é interessante que sejam estudados os impactos da variação da taxa de câmbio entre o dólar e a moeda corrente da China, o yuan, na cotação internacional da soja. Dessa forma, caso se constate que as séries estão correlacionadas e que existe uma relação de causalidade comprovada entre elas, partes interessadas poderiam se utilizar dessas informações para incrementar modelos matemáticos que visam prever o preço da soja. Esse tema se torna mais pertinente à medida em que a China costumava adotar uma taxa de câmbio fixa entre o dólar e o yuan, e que passou, a partir de meados do fim de 2005, a flexibilizar gradualmente suas políticas, em busca de uma moeda mais flexível (DAS, 2019).

Considerando as informações destacadas, que mostram a importância da soja, tanto no cenário interno do país, quanto no contexto internacional, bem como no problema evidenciado, esse estudo visa estudar a existência de possíveis formas de correlação e causalidade entre um conjunto de dados contendo o preço da soja, e outro contendo a variação da taxa de câmbio entre o yuan e o dólar, quando plotados na forma de séries temporais.

Assim, é presumido que partes interessadas poderiam se utilizar dos resultados dessa pesquisa para auxiliar no estudo de relações entre a cotação internacional da soja e a variação da taxa de câmbio entre dólar e yuan, bem como na montagem de modelos matemáticos ou estratégias que complementariam processos de tomada de decisões relacionadas à negociação de soja e derivados, possibilitando um melhor gerenciamento de risco para as entidades interessadas na variação de preço da commodity.

Como objetivo geral, o artigo em questão busca a investigação de uma possível relação de correlação e causalidade entre a cotação internacional da soja, em dólar, e a taxa de câmbio entre o yuan e o dólar. Como objetivos específicos, o presente estudo busca atingir três objetivos. Primeiramente, estudar a utilização da linguagem R para análise de dados e estudo de correlações. Segundamente, encontrar e padronizar conjuntos de dados referentes à cotação internacional da soja e a taxa de câmbio entre dólar e yuan. Por último, aplicar os conceitos estudados nos conjuntos de dados encontrados, de forma a buscar evidenciar relações de correlação e causalidade entre as séries temporais.

Além desta primeira seção introdutória, este artigo contempla outras três seções. A

segunda seção trata do referencial teórico utilizado no estudo, incluindo informações relevantes ao contexto da soja, bem como ferramentas e técnicas pertinentes à proposta de pesquisa, tais como os conceitos de correlação e causalidade. Na terceira seção, é apresentada a metodologia utilizada para o desenvolvimento do trabalho, incluindo informações sobre as etapas de pesquisa e exploração sobre o tema, bem como a manipulação, tratamento e análise dos dados utilizando a linguagem R. Por fim, na quarta e última seção, são apresentadas algumas considerações finais, pertinentes ao estudo e pesquisas futuras relacionadas.

2 Referencial Teórico

Esta seção visa apresentar o referencial teórico utilizado na elaboração do estudo. Serão apresentados alguns conceitos relevantes ao contexto da soja no Brasil e no mundo, a linguagem R e conceitos estatísticos importantes para o entendimento das análises realizadas, tais como o conceito de correlação, gráficos Q-Q, teste de Shapiro-Wilks, coeficiente de correlação de Spearman e Causalidade de Granger.

2.1 Cultura da Soja

Embora a cultura da soja tenha sido introduzida no Brasil na metade do século dezenove, a sua devida exploração se iniciou apenas no início do século vinte, quando foi levada à região sul do país, onde o clima se mostrou mais propício para o desenvolvimento da mesma, de acordo com Dall'Agnol (2006). Segundo o mesmo, a soja foi crescendo de maneira tímida nas décadas seguintes, ganhando maior destaque a partir dos anos 80, quando, graças aos avanços tecnológicos conseguidos pelos cientistas brasileiros, a sua cultura foi introduzida com sucesso na região do Cerrado brasileiro.

Segundo Hirakuri (2020), a partir da década de 2000, com o crescimento econômico rápido dos países emergentes, e com isso, a elevação do poder de compra da população, criaram-se condições favoráveis para um aumento na demanda mundial por alimentos, com ênfase em proteína animal. Nesse cenário, segundo o autor, as variáveis determinantes para a formação do preço da soja na *Chicago Board of Trade* (CBOT) se tornaram, basicamente, a oferta e demanda do produto no mercado.

2.2 Chicago Board Of Trade

A *Chicago Board of Trade* (CBOT) é uma exchange de commodities utilizada globalmente como referência para o preço de commodities. A CBOT foi criada em 1848 para ajudar fazendeiros a realizarem o gerenciamento de risco de suas plantações, possibilitando aos mesmos que firmassem, de maneira sistêmica e organizada, contratos para a compra e venda futura de produtos agrícolas, como trigo, milho e soja. Com o passar do tempo, a entidade passou a ofertar a possibilidade de se negociar vários outros produtos financeiros relacionados a commodities (CHEN, 2022a).

Segundo Chen (2022b), uma exchange de commodities é uma entidade legal que determina regras e procedimentos para a negociação de contratos padronizados de commodities e produtos financeiros relacionados. Nesse contexto, as partes interessadas não costumam negociar

commodities fisicamente, através de uma exchange. Ao invés disso, elas negociam contratos futuros.

A CBOT se mantém como o principal fórum no qual o preço internacional da soja é setado, detendo a maior fatia de operações de hedging e trading envolvendo mercados futuros (OLIVEIRA, 2016). Por ser uma entidade criada em solo americano, a Chicago Board of Trade utiliza o dólar para suas cotações, e, segundo Oliveira (2016), o dólar é a unidade padrão para o comércio internacional de soja.

2.3 Correlação

A emergência da correlação foi uma das principais revelações em estatística durante o final do século 19 (ALDRICH, 1995). Correlação é a medição estatística de dependência linear entre duas variáveis aleatórias (GEORGAKOPOULOS, 2015). Mais especificamente, ao estudar correlação, nos interessamos pela força dessa relação linear entre as variáveis, que será determinada pela estimativa de um coeficiente de correlação, sobre o qual se realizam inferências estatísticas, de forma a analisar como a associação entre duas variáveis afeta um determinado experimento (DOWDY; WEARDON; CHILKO, 2004).

O coeficiente de correlação geralmente possui um valor entre 1 e -1 e não possui unidades de medida associadas (DOWDY; WEARDON; CHILKO, 2004). Segundo o autor, baseando-se na magnitude do valor absoluto desse coeficiente, pode-se perceber a força de uma associação linear entre variáveis.

Considerando o coeficiente de correlação como “r”, em todos os casos, r é maior ou igual a -1, e menor ou igual a 1. Se r é igual a -1, existe uma relação negativa perfeita (DOWDY; WEARDON; CHILKO, 2004), e, à medida que os valores de uma variável crescem, os valores da outra variável decrescem, na mesma proporção. Se r é igual a 1, existe uma relação positiva perfeita, e a variação dos valores das variáveis têm comportamentos idênticos, em suas devidas proporções. À medida em que o valor de r se aproxima de zero, as associação entre variáveis diminui. Segundo Dowdy, Weardon e Chilko (2004), se r é igual a 0, as variáveis não estão correlacionadas.

É importante destacar que correlação não implica causalidade, visto que duas variáveis podem estar correlacionadas, mas a variação em uma das variáveis, não necessariamente irá causar uma mudança na outra. Isso pode ser mais facilmente constatado ao analisar os dados do site Spurious Correlations, disponíveis em Vigen (2022), que agrega uma grande quantidade de variáveis correlacionadas, mas sem uma relação de causalidade necessária.

Existem diversas metodologias para se calcular a correlação entre duas ou mais variáveis. As três formas populares de se estimar correlação são o coeficiente de produto-momento de Pearson, o coeficiente de correlação de Kendall e o coeficiente de correlação de rank de Spearman (GEORGAKOPOULOS, 2015), que foi utilizado no presente estudo.

2.4 Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman é um método estatístico não paramétrico proposto por Charles Spearman como uma medida de força de uma associação entre duas variáveis. É uma medida de uma associação monótona que é utilizada quando a distribuição dos dados faz com que o coeficiente de correlação de Pearson seja indesejado ou errôneo (HAUKE; KOSSOWSKI, 2011). Para se calcular o coeficiente de Spearman, ao invés de se realizar o cálculo do coeficiente de correlação utilizando os valores em si, como em outros métodos, o mesmo realiza, primeiramente, um ranking dos dados, e após, calcula o coeficiente com base no ranking.

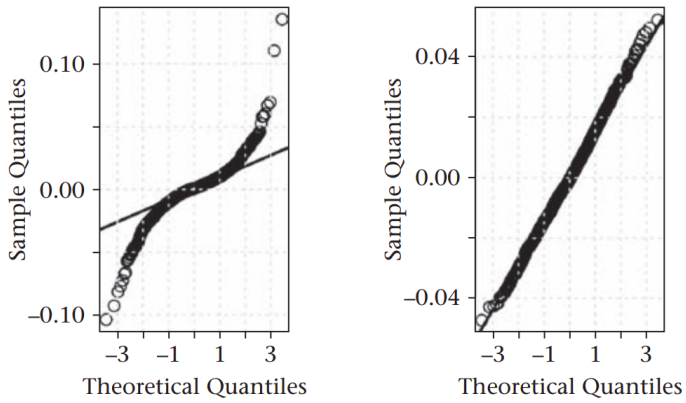
O coeficiente de correlação de Spearman é usualmente adotado quando não se pode assumir que as duas variáveis seguem uma distribuição normal (ARTUSI; VERDERIO; MARUBINI, 2002), já que, segundo Akoglu (2018) em distribuições não normais, coeficientes de correlação devem ser calculados pelo ranqueamento dos dados, e não por seus valores.

2.5 Gráfico Q-Q

Segundo Georgakopoulos (2015), é interessante que se realizem testes e suposições comparando sets de dados a serem estudados com distribuições (gaussianas) normais, já que realizar isso pode direcionar nossas análises e simplificar cálculos. Segundo o mesmo, uma boa forma de se visualizar a diferença entre dados empíricos e a distribuição normal teórica, é através de gráficos Q-Q (ou Quantil-Quantil).

A interpretação dos gráficos é relativamente simples: o desvio dos dados em relação à linha reta que cruza o gráfico, significa um desvio em relação à normalidade (GEORGAKOPOULOS, 2015). Assim, se todos os pontos acabassem sobre a linha reta, os dados em questão seriam distribuidos normalmente. Para exemplificação, pode-se observar, na Figura 1, dois gráficos Q-Q: um à esquerda, contendo dados (denotados através de círculos) que se desviam da distribuição normal (linha reta), e outro à direita, onde os dados acompanham a linha reta, e, portanto, seguem uma distribuição normal.

Figura 1 – Exemplos de gráficos Q-Q seguindo, à esquerda, uma distribuição não normal, e à direita, uma distribuição normal



Fonte: Georgakopoulos (2015)

2.6 Teste de Shapiro-Wilks

Segundo Georgakopoulos (2015), além do uso de gráficos Q-Q, que demonstram de forma mais visual um desvio em relação à normalidade, existem vários testes estatísticos para a determinação da normalidade e ou desvio em relação à normalidade, sendo um desses testes, o Teste de Shapiro-Wilks. Segundo o autor, o teste retorna um valor de "p". Esse valor resultante especifica a probabilidade de os dados terem se originado de uma distribuição normal. Se o valor de "p" for menor que 0.05, a hipótese de que os dados estão normalmente distribuidos deve ser rejeitada. Se o valor da variável for maior que "p", a hipótese de teste é validada, e os dados muito provavelmente seguem uma distribuição normal.

Georgakopoulos (2015) salienta que esse tipo de análise puramente estatística não deve ser considerado isoladamente, e que é importante entender, antes de tudo, a estrutura dos dados, ponto no qual, tanto análise visual dos dados quanto gráficos Q-Q auxiliam.

2.7 Causalidade de Granger

Clive Granger, em 1969, elaborou um framework matemático para descrever uma forma de causalidade, que mais tarde recebeu o nome de Causalidade de Granger. Dadas duas variáveis estocásticas X e Y, se diz que existe uma relação causal (de Granger) entre elas se as observações passadas de Y ajudam a predizer o estado atual de X, e vice-versa. Se sim, diz-se que Y causa, por Granger, X (LIMA et al., 2020).

Assim como o teste de Shapiro-Wilks, o teste de causalidade retorna uma variável "p", e, a depender do seu valor, uma determinada hipótese é aceita, ou invalidada. Se "p"for menor que 0.05, a hipótese de que conhecer o valor de uma série X em um determinado instante é útil para se predizer valores futuros de uma série temporal Y, é validada. Se "p"for menor que 0.05, a hipótese é invalidada.

2.8 Linguagem de Programação R

Para a importação e análise de dados das séries temporais, foi utilizada a linguagem de programação R. A linguagem R é um projeto open-source que oferece um ambiente para cálculo estatístico e criação de gráficos. Ela fornece uma vasta gama de ferramentas de análise estatística, modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, entre outros (R-PROJECT, 2022).

Ela foi a linguagem escolhida como objeto de estudo devido à sua aplicabilidade no campo da estatística e análise de séries temporais; por ser um software livre; devido à facilidade em utilizá-la para criar gráficos de qualidade e por possuir diversas bibliotecas contendo métodos estatísticos para estudos de correlações.

Também possui uma comunidade em crescimento, e vários livros exemplificando seus usos em estudos de correlação e causalidade, como o de Georgakopoulos (2015), bem como seus usos em séries temporais, tal como o de Tsay (2013)

3 Metodologia

Esta seção visa explicar a metodologia que foi utilizada para o desenvolvimento do estudo de caso, explicando todas as etapas percorridas e atividades realizadas para atingir os objetivos desejados, desde os processos de pesquisa, importação e tratamento dos dados, até a aplicação de funções estatísticas relevantes ao objeto de estudo.

Após compreensão do contexto geral da comercialização mundial do ativo, bem como as entidades envolvidas em sua negociação, foi escolhida uma variável para ser estudada, juntamente com a cotação da soja. A variável escolhida foi a variação da taxa de câmbio entre dólar e yuan. Foram então definidos os pacotes de dados, e as fontes padrão a serem utilizados para realizar o download dos sets de dados no ambiente de desenvolvimento em R.

Os critérios que foram utilizados para a escolha das fontes para importação das variáveis foram os que seguem. Primeiramente, ambos os sets de dados deveriam estar disponíveis na internet de forma gratuita. Segundamente, os dados deveriam estar disponíveis em um formato ou pacote possível de ser importado diretamente para o R, ou possuirem a possibilidade de conversão para um formato compatível.

Foi então realizada uma investigação acerca dos métodos de análise estatística relevantes à estudos de correlação e causalidade, de forma a compreender quais deles seriam mais adequados para explorar uma correlação entre a variação da taxa de câmbio entre dólar e yuan, e a cotação internacional da soja. Isso foi feito através de pesquisa exploratória, bem como testes práticos a fim de validar o uso dos métodos pesquisados, utilizando a linguagem de programação R.

3.1 Importação e Preparo dos Dados

Nesta subseção, serão descritos os passos percorridos para realizar a importação e preparo dos dados das séries temporais, de forma a se obter dados possíveis de serem analisados pelas funções dos pacotes em R, bem como encontrar e corrigir quaisquer inconsistências presentes nos mesmos.

3.1.1 Dados da Soja

Os dados referentes à cotação mensal da soja em dólares foram retirados do IPEA. O Instituto de Pesquisa Econômica Aplicada (IPEA) é um Instituto criado em 1964, está atualmente vinculado ao Ministério da Economia, e tem como missão institucional aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro, por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas (IPEA, 2022).

A entidade, por meio de seu Plano de Dados aberto, disponibiliza, por meio da plataforma Ipeadata, dados abertos e atualizados diariamente (IPEA, 2022). Os dados disponíveis na plataforma podem ser importados diretamente na linguagem R, por meio do pacote “ipeadatar”, disponível na rede CRAN.

Para descobrir as séries presentes no pacote, foi utilizada a função avaliable_series(), disponível no pacote, e o resultado foi atribuido à uma variável de nome “series_av”:

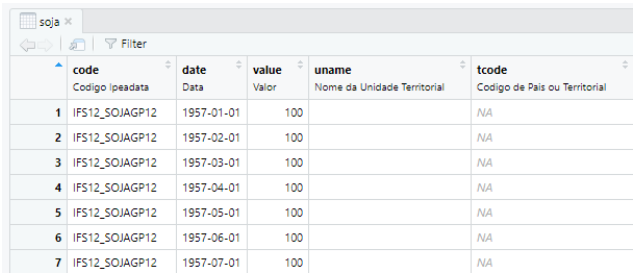
```
series_av <- available_series(language = c("en", "br"))
```

Após análise dos sets de dados disponíveis, a série temporal escolhida foi a referente à cotação internacional (ou seja, em dólar) mensal da commodity soja em grão, de código "IFS12_SOJAGP12". Após escolha, foi importada a série temporal desejada para uma variável "soja", no formato de data.frame, através do código do set de dados, utilizando o seguinte comando:

```
soja <- ipeadata(code = "IFS12_SOJAGP12", language = "br")
```

Obteve-se assim o conjunto de dados disponível na Figura 2, armazenado na variável de nome “soja”:

Figura 2 – Dados referentes à cotação da soja, importados do pacote ipeadatar



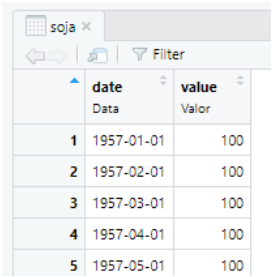
soja × | Filter

	code Codigo Ipeadata	date Data	value Valor	uname Nome da Unidade Territorial	tcode Codigo de Pais ou Territorial
1	IFS12_SOJAGP12	1957-01-01	100		NA
2	IFS12_SOJAGP12	1957-02-01	100		NA
3	IFS12_SOJAGP12	1957-03-01	100		NA
4	IFS12_SOJAGP12	1957-04-01	100		NA
5	IFS12_SOJAGP12	1957-05-01	100		NA
6	IFS12_SOJAGP12	1957-06-01	100		NA
7	IFS12_SOJAGP12	1957-07-01	100		NA

Fonte: Elaborado pelo autor.

Foram removidas as colunas "uname", "tcode" e "code" do data.frame importado, já que não seriam relevantes para o estudo, originando uma coluna “value”, de tipo numérico, contendo o valor da cotação da soja em dólar em um dado mês, e uma coluna de nome “date”, de tipo "Date", contendo a data de referência do valor correspondente, como pode ser observado na Figura 3. A série foi então plotada utilizando o comando plot(), e pode ser visualizada na Figura 4.

Figura 3 – Dados referentes à contação da soja, após tratamento

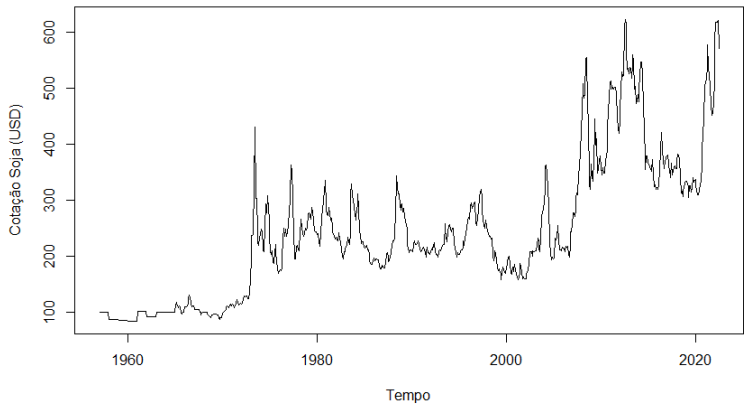


soja × | Filter

	date Data	value Valor
1	1957-01-01	100
2	1957-02-01	100
3	1957-03-01	100
4	1957-04-01	100
5	1957-05-01	100

Fonte: Elaborado pelo autor.

Figura 4 – Cotação mensal da soja de 1957 até 2022, em dólares



Fonte: Elaborado pelo autor.

Foi realizada, então, a soma de valores nulos presentes na série, de forma a encontrar a presença de possíveis falhas no data.frame. Para isso, foi utilizada a função "sum()", e passada como parâmetro outra função que retorna os registros nulos em um data.frame: "is.na()", como segue: "sum(is.na(nome_do_dataframe))". Obteve-se um valor de zero registros nulos, confirmando que não existem valores nulos na série.

Para realizar algumas operações e testes, além de um objeto do tipo data.frame, também foi necessário a construção de um objeto de série temporal de tipo "time series". A série temporal contida no data.frame foi transformada em um objeto desse tipo utilizando a função ts(), contida no pacote “tseries”, do R. Para a função, é passada a coluna com os dados, as datas de início e fim da série, e a frequência dos dados. Para a criação de uma série com dados mensais, a frequência passada como parâmetro foi de 12.

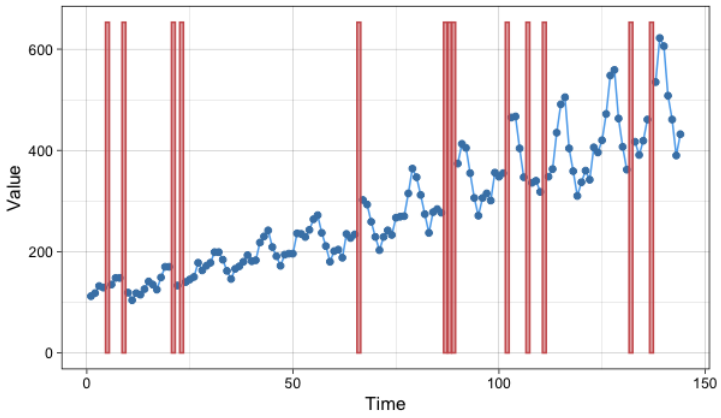
```
sojats<- ts(soja$value, start=c(1957, 1),
          end=c(2022, 7),frequency=12)
```

Mais uma vez, foram buscadas inconsistências nos dados do objeto criado. Por ser um objeto do tipo "time series", foi utilizada a biblioteca "imputeTS", e aplicada a seguinte função, proveniente da mesma:

```
ggplot_na_distribution(sojats)
```

A função "ggplot_na_distribution()"retorna um gráfico da série temporal, contendo os valores da mesma plotados em relação ao tempo, com falhas nos dados marcadas em vermelho, caso existam, como pode ser observado na Figura 5, retirada da documentação da função em Moritz (2022):

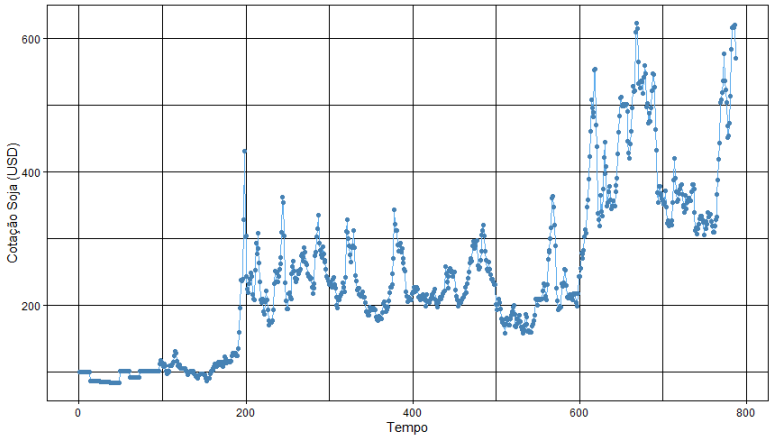
Figura 5 – Exemplo de aplicação de função "ggplot_na_distribution()"em série com valores faltantes



Fonte: Moritz, Steffen (2022).

Ao aplicar a função na série de dados da soja, obteve-se a Figura 6. Mediante análise da figura, percebe-se que que a série temporal foi criada com sucesso, e sem inconsistências.

Figura 6 – Gráfico resultante da aplicação da função "ggplot_na_distribution()"na série de dados da soja



Fonte: Elaborado pelo autor.

Também foi criado um objeto de série temporal contendo o preço da soja de janeiro de 2006 até julho de 2022. Para isso, primeiramente foi criado outro dataframe com os cortes de dados nos períodos corretos, feitos da seguinte forma:

```
soja2006 <- soja[589:787, ]
```

Esse data.frame foi então transformado em um objeto do tipo “time series” com o seguinte comando:

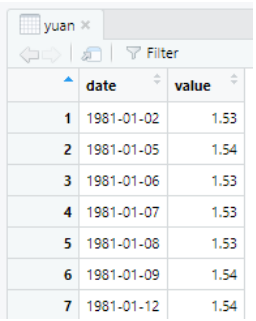
```
sojats<- ts(soja2006$value, start=c(2006, 1),
            end=c(2022, 7), frequency=12)
```

3.1.2 Dados da Taxa de Câmbio entre Yuan e Dólar

A taxa de câmbio utilizada para o estudo foi a de yuan para dólar, ou seja, quantos yuans são equivalentes a um dólar, em um dado momento. Ao se realizar a busca no pacote “ipeadatar”, bem como no site do Ipeadata (buscas foram realizadas em 10/11/2022), foi percebido que os dados referentes à taxa de câmbio entre dólar e yuan estavam desatualizados, já que a taxa de câmbio mensal possuia dados apenas até o ano de 2017. De forma alternativa, foram encontrados dados diários atualizados na plataforma Macrotrends (2022), que disponibiliza uma grande variedade de dados, que variam de taxas de câmbio, até cotações de metais preciosos e taxas de juros, e são possíveis de serem exportados em diversos formatos.

O download do set de dados foi feito diretamente pelo site, em formato “.csv”, e o mesmo foi importado no RStudio, em uma variável data.frame, utilizando a função "read.csv()". Como pode ser observado na figura 7, a série continha uma coluna “value”, contendo o valor da taxa de câmbio em uma determinada data, e uma coluna “date”, contendo a data de referência.

Figura 7 – Dados importados referentes à taxa de câmbio entre dólar e yuan



	date	value
1	1981-01-02	1.53
2	1981-01-05	1.54
3	1981-01-06	1.53
4	1981-01-07	1.53
5	1981-01-08	1.53
6	1981-01-09	1.54
7	1981-01-12	1.54

Fonte: Elaborado pelo autor.

Ao se utilizar o comando "str()"para análise do conjunto de dados, pôde-se constatar que a coluna “date” estava em formato “chr” (caractere). Para a transformação desse set em uma série temporal, era necessário que estivesse no formato “Data”, mesmo formato em que se encontrava a série com a cotação da soja. Para essa transformação foi utilizado o seguinte comando:

```
cambio$date <- as.Date(cambio$date, "%Y-%m-%d")
```

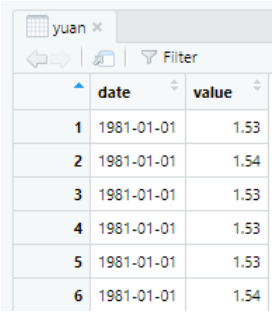
Foi realizada a soma de valores nulos na série obtida utilizando a função "sum(is.na())", e não foram encontrados valores nulos nela. Contudo, após inspeção visual dos dados, foram percebidas algumas falhas nos dados diários da taxa de câmbio, onde faltavam os dados de alguns dias. Também era necessário que os dados da taxa de câmbio estivessem com frequência mensal, assim como a série da soja. Para solucionar ambos os problemas, os dados diários

foram convertidos para uma média mensal com o valor da taxa de câmbio de yuan para dólar. Para realizar essa operação, premeiramente foi utilizada a função floor_date(), da biblioteca "lubridate", e as datas dos dados foram arredondadas para datas mensais:

```
cambio$date <- floor_date(cambio$date, "month")
```

Após essa operação, foi originado o set de dados visível na Figura 8.

Figura 8 – Dados da taxa de câmbio após uso de função floor_date()



yuan

Filter

	date	value
1	1981-01-01	1.53
2	1981-01-01	1.54
3	1981-01-01	1.53
4	1981-01-01	1.53
5	1981-01-01	1.53
6	1981-01-01	1.54

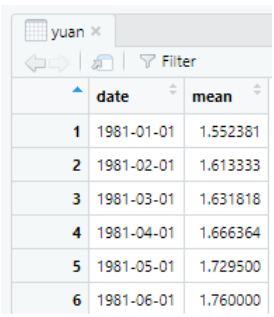
Fonte: Elaborado pelo autor.

Foi então utilizada a biblioteca “dplyr” para agrupar os valores por data, já que todos os valores de um dado mês estavam setados com a mesma data, e foi extraida a média deles, utilizando o seguinte comando:

```
cambio <- cambio%>%
group_by(date) %>%
        summarize(mean = mean(value))
```

Isso originou uma média mensal da taxa de câmbio dolar x yuan, como pode ser visto a seguir. Mais uma vez, foram buscados por valores nulos no data.frame, e não foram encontrados. A série resultante pode ser observada na Figura 9, e foi plotada utilizando o comando "plot()", cujo resultado pode ser visualizado na Figura 10.

Figura 9 – Dados mensais médios da taxa de câmbio entre yuan e dólar



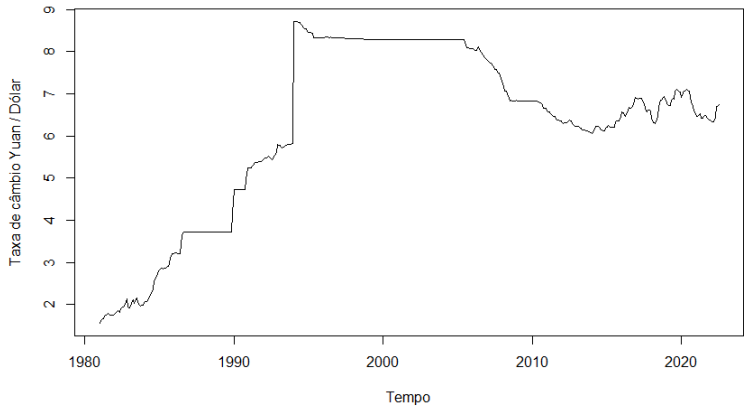
yuan

Filter

	date	mean
1	1981-01-01	1.552381
2	1981-02-01	1.613333
3	1981-03-01	1.631818
4	1981-04-01	1.666364
5	1981-05-01	1.729500
6	1981-06-01	1.760000

Fonte: Elaborado pelo autor.

Figura 10 – Taxa de câmbio mensal entre dólar e yuan, de 1981 até 2022



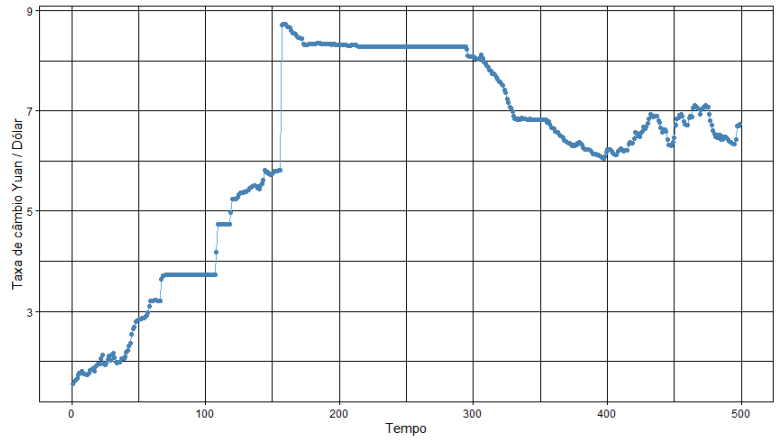
Fonte: Elaborado pelo autor.

Por fim, a série foi transformada em um objeto do tipo ts, utilizando a seguinte função:

```
cambiots<- ts(cambio$mean, start=c(1981, 1),
           end=c(2022, 7), frequency=12)
```

E foi buscada, mais uma vez, por falhas e valores nulos nos dados do objeto de série temporal utilizando a função ggplot_na_distribution(), como pode ser observado na Figura 11.

Figura 11 – Gráfico resultante da aplicação da função "ggplot_na_distribution()"na série de dados da taxa de câmbio



Fonte: Elaborado pelo autor.

Também foi criado um objeto de série temporal contendo os valores da taxa de câmbio de janeiro de 2006 até julho de 2022. Para isso, primeiramente foi criado outro dataframe com os cortes de dados nos períodos corretos, feitos da seguinte forma:

```
cambio2006 <- cambio[301:499, ]
```

Esse data.frame foi então transformado em um objeto do tipo "time series" com o seguinte comando:

```
cambiots<- ts(cambio2006$mean, start=c(2006, 1),
              end=c(2022, 7), frequency=12)
```

3.2 Análise dos Dados

Nesta subseção será descrita como foi realizada a análise dos dados tratados na subseção anterior, o que inclui análises gráficas qualitativas e aplicação de testes estatísticos de correlação e causalidade entre as séries. Como pode ser observado no gráfico da Figura 6, existem alguns momentos em que a taxa de câmbio entre dólar e yuan se manteve relativamente estacionária. Isso se deve à regulações feitas pela China, como demonstrado por Das (2019).

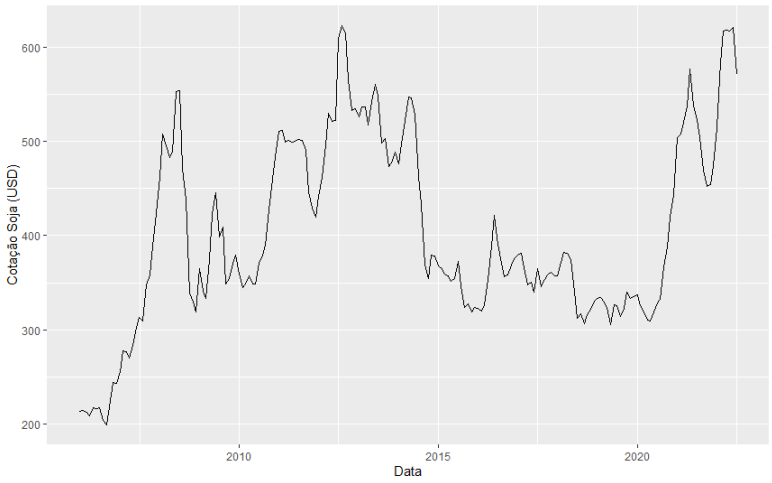
Ainda segundo Das (2019), a partir de julho de 2005, a China anunciou uma mudança de uma taxa de câmbio fixa, para uma flexibilização gradual das políticas do país em relação às flutuações da taxa de câmbio entre o dólar e a sua moeda oficial, embora, segundo o mesmo, a estabilidade da taxa de câmbio ainda se manteve, de certa forma.

Por esse motivo, e visto que estacionaridades poderiam influenciar nos testes estatísticos a serem aplicados, foram utilizados os dados de ambas as séries de janeiro de 2006 em diante, para o presente estudo. A partir daqui, ao mencionarmos as séries temporais, estamos falando de dados compreendidos entre janeiro de 2006 e julho de 2022.

Foram então criados dois objetos de séries temporais, derivados dos objetos criados nas subseções anteriores, um contendo a cotação internacional da soja, e outro, contendo a taxa de câmbio de yuan para dólar, ambas as séries, de janeiro de 2006 até julho de 2022.

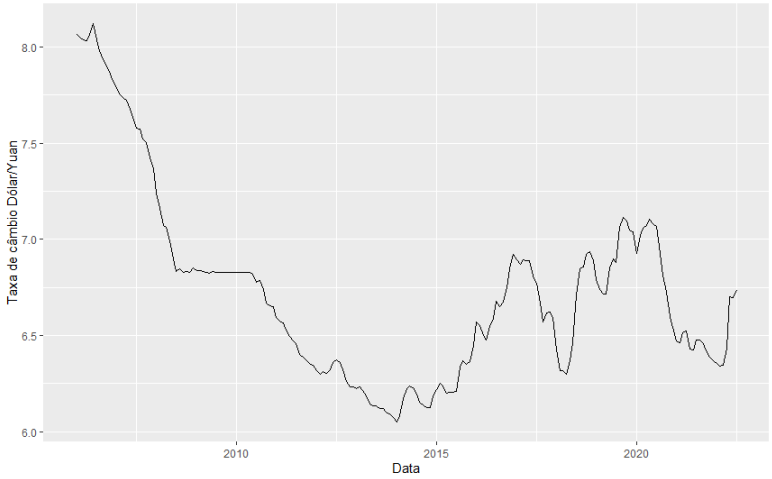
Foi realizada a importação da biblioteca tidyverse, que contém outros pacotes incluidos, como o ggplot2. Utilizando o último, pode-se plotar então os gráficos das séries, a partir de 2006, como pode ser observado nas Figuras 12 e 13.

Figura 12 – Série temporal da cotação da soja em dólares, de janeiro de 2006 até julho de 2022, plotada utilizando a biblioteca ggplot2



Fonte: Elaborado pelo autor.

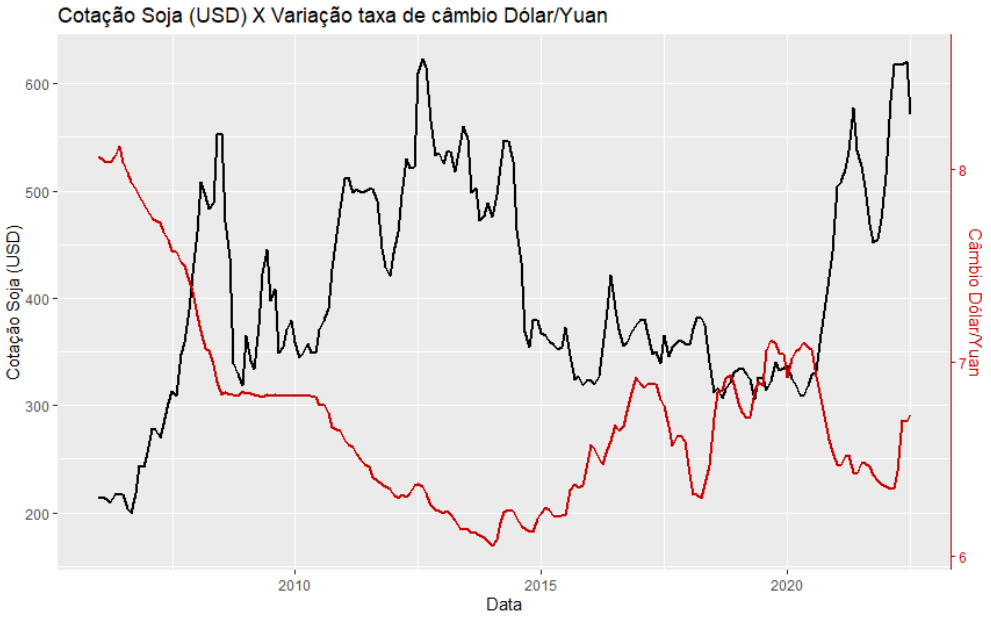
Figura 13 – Série temporal da taxa de câmbio entre dólar e yuan, de janeiro de 2006 até julho de 2022, plotada utilizando a biblioteca ggplot2



Fonte: Elaborado pelo autor.

Para facilitar a inspeção visual, foi utilizada a biblioteca ggplot2 para plotar os 2 gráficos em uma mesma figura. Para isso, também foram realizadas as importações dos pacotes "patchwork"e "hrbrthemes". As escalas das variáveis foram ajustadas, de forma a ajudar a evidenciar uma relação entre elas, visualmente, como pode ser visto na Figura 14.

Figura 14 – Cotação da Soja (USD) x Variação da taxa de câmbio, no período



Fonte: Elaborado pelo autor.

Mediante análise visual, identifica-se alguma forma de correlação entre as séries. Percebe-se que, no período entre 2006 e 2007, aproximadamente, houve uma grande valorização percentual na moeda chinesa em relação à moeda americana. Nesse mesmo período, o preço da soja subiu consideravelmente. Da mesma forma, a taxa de câmbio se manteve estacionária entre junho de 2008 até junho de 2010, devido à uma reclassificação momentânea da mesma como fixa por parte da China (DAS, 2019),e a cotação da soja não teve variações expressivas no período.

Entre 2010 e 2014, houve mais uma grande valorização percentual do yuan em relação ao dólar, e, enquanto isso, a cotação internacional da soja passou de aproximadamente 350 dólares, para mais de 600 dólares, o que seria mais um possível indício de correlação negativa entre as séries. Também percebe-se a presença de um período em que as séries apresentaram tendências comuns, de fevereiro de 2016 até julho de 2018. Percebe-se que de 2014 até 2016, as séries possuem tendências inversas, o mesmo acontecendo brevemente no ano de 2020.

Seguindo a metodologia para inferência de normalidade como apresentada por Georgakopoulos (2015), foram construidos 2 gráficos Q-Q, um para a cotação da soja (Figura 15) e outro para a taxa de câmbio (Figura 16). Para isso, foi utilizada a função "ggqqplot()", disponível no pacote "ggpubr".

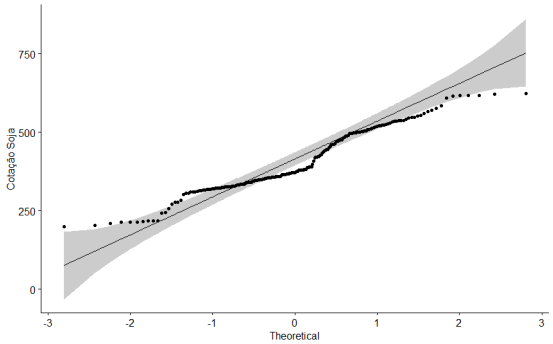


Figura 15 – Gráfico Q-Q - Soja

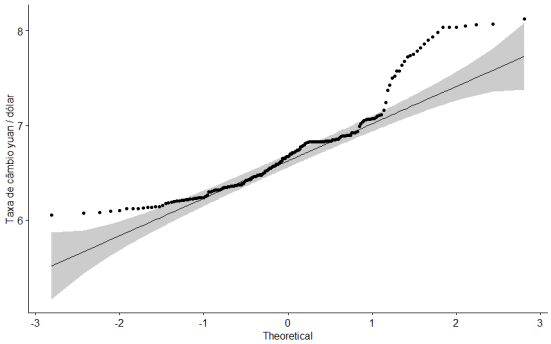


Figura 16 – Gráfico Q-Q - Câmbio

Mediantes análise dos gráficos gerados, percebe-se que ambas as séries temporais se desviam da normalidade, visto que não acompanham a linha ao centro, que representa uma distribuição de dados normal. Para complementar a análise visual dos gráficos, foi realizado o teste de Shapiro-Wilks, por meio da função "shapiro.test()", disponível no pacote "stats". Os resultados do teste podem ser observados nas Figuras 17 e 18.

```
> shapiro_soja <- shapiro.test(soja2006$value)
> shapiro_soja

        Shapiro-Wilk normality test

data:  soja2006$value
W = 0.95837, p-value = 1.39e-05
```

Figura 17 – Teste de Shapiro - Cotação da Soja

```
> shapiro_cambio <- shapiro.test(cambio2006$mean)
> shapiro_cambio

        Shapiro-Wilk normality test

data:  cambio2006$mean
W = 0.8936, p-value = 1.081e-10
```

Figura 18 – Teste de Shapiro - Taxa de Câmbio

Percebe-se, pelo resultado do teste de Shapiro-Wilks, que os dois valores "p" são menores que 0.05, o que implica que a hipótese de teste (no caso, de que os dados estão normalmente distribuidos) deve ser rejeitada. Dessa forma, mediante análise dos gráficos Q-Q, e dos resultados do teste de Shapiro-Wilks, considera-se que a distribuição dos dados de ambas as séries não segue uma distribuição gaussiana. Isso implica que uma boa forma de se estimar a correlação das variáveis seria através do Coeficiente de Correlação de Spearman. Ele foi realizado usando a função "cor.test()", do pacote "stats", passando como parâmetro o nome do método desejado, no caso, "spearman", e os resultados podem ser observaodos na Figura 19.

Figura 19 – Resultados do teste de Correlação de Spearman

```
> spearman_result <- cor.test(soja_cambio$soja,
+                             soja_cambio$cambio,
+                             method = "spearman",
+                             exact = FALSE)
>
> spearman_result

        Spearman's rank correlation rho

data:  soja_cambio$soja and soja_cambio$cambio
S = 2137636, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.6275589
```

Fonte: Elaborado pelo autor.

Ao se analisar os resultados do teste de correlação de Spearman, percebe-se que o coeficiente de correlação "rho" gerado é de aproximadamente -0.628, o que implica uma correlação negativa entre as variáveis estudadas. Dessa forma, se o valor de uma das variáveis aumenta, a tendência parece ser de que o valor da outra diminua. O coeficiente de correlação é estatisticamente relevante, e está alinhado com a hipótese a ser validada, e com a análise visual entre as séries.

Para os testes de Causalidade de Granger, foi utilizada a função "grangertest()", que está contida no pacote "lmtest". A função recebe duas variáveis do tipo time series, separadas por um "∼". A hipótese a ser testada é a de que a série à esquerda do símbolo é causada pela série à direita do símbolo. O resultado do teste pode ser observado na Figura 20.

Figura 20 – Resultados do teste de Causalidade de Granger entre a variação da taxa de câmbio e a cotação da soja

```
> grangertest(soja ~ cambio, soja_cambio)
Granger causality test

Model 1: soja ~ Lags(soja, 1:1) + Lags(cambio, 1:1)
Model 2: soja ~ Lags(soja, 1:1)
  Res.Df Df      F Pr(>F)
1    195
2    196 -1 0.2481  0.619
```

Fonte: Elaborado pelo autor.

O valor p resultante do teste foi de 0.619. Sendo que o valor não é menor do que 0.05, não se pode rejeitar a hipótese nula. Dessa forma, não se pode inferir que a variação da taxa de câmbio entre dólar e yuan causa uma mudança no preço da soja. Isso seria um indício de que, embora as séries estejam correlacionadas, uma mudança na taxa de câmbio entre dólar e yuan não necessáriamente causaria uma mudança na cotação da soja.

4 Considerações Finais

Por fim, pode-se inferir que este estudo conseguiu, de forma metódica e organizada, investigar relações de correlação e causalidade entre a cotação internacional da soja, em dólar, e a taxa de câmbio entre o yuan e o dólar. Foram realizadas as importações e padronização dos dados, bem como aplicados conceitos estatísticos estudados, por meio de pacotes da linguagem R, de forma a evidenciar relações estatísticas entre as séries.

Os dados foram padronizados e tratados, de forma a possibilitar, primeiramente, uma análise visual qualitativa da relação entre as séries, que, em primeiro momento, apresentavam indícios de uma possível correlação negativa. Após, foram realizados testes estatísticos, utilizando pacotes e funções disponíveis para a linguagem, de forma a compreender, por meio da análise visual de gráficos Q-Q e análise quantitativa dos resultados dos testes de Shapiro-Wilks, que ambas as séries seguem um padrão não normal, e que, devido a isso, uma boa forma de se estimar a correlação entre elas seria através do Coeficiente de Correlação de Spearman.

Foi realizado o então cálculo do Coeficiente de Correlação de Spearman e o mesmo indicou a presença de uma correlação negativa entre as séries, estando em concordância com a hipótese

apresentada, e a análise visual das séries. Por fim, realizou-se um teste de causalidade, por meio da metodologia de Causalidade de Granger. Este teste, contudo, não apresentou indícios de que exista uma relação de causalidade entre as séries temporais. Dessa forma, em primeira análise, não parece ser interessante adicionar essa correlação em modelos matemáticos preditivos do preço da soja, que seriam do interesse das partes interessadas.

O presente estudo possuia como intuito apenas o estudo de relações de correlação e causalidade entre as séries. Fatores econômicos técnicos, externos ao contexto de pesquisa, não foram levados em consideração. Para estudos futuros, seria interessante que fossem estudadas as relações entre as variáveis, em termos econômicos, e o porque de existir tal correlação.

Também seria de utilidade o estudo das séries em termos de acontecimentos históricos e de mercado, de forma a entender as variações que ocorreram, tanto na série de dados contendo a cotação da soja, como na série contendo a taxa de câmbio. Vale salientar que embora essa correlação entre a taxa de câmbio entre dólar e yuan e a cotação da soja em dólar exista, atualmente, não se pode afirmar que continuará existindo. Dinâmicas e políticas monetárias podem vir a interferir na variação da taxa de câmbio entre dólar e yuan, por exemplo.

Referências

AKOGLU, H. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, v. 18, n. 3, p. 91–93, 2018. ISSN 2452-2473. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2452247318302164>. Citado na página 6.

ALDRICH, J. Correlations genuine and spurious in pearson and yule. *Statistical Science*, Institute of Mathematical Statistics, v. 10, n. 4, p. 364–376, 1995. ISSN 08834237. Disponível em: <http://www.jstor.org/stable/2246135>. Citado na página 5.

ARTUSI, R.; VERDERIO, P.; MARUBINI, E. Bravais-pearson and spearman correlation coefficients: Meaning, test of hypothesis and confidence interval. *The International Journal of Biological Markers*, v. 17, n. 2, p. 148–151, 2002. PMID: 12113584. Disponível em: <https://doi.org/10.1177/172460080201700213>. Citado na página 6.

CHEN, J. *Chicago Board of Trade (CBOT)*. 2022. Disponível em: <https://www.investopedia.com/terms/c/cbot.asp>. Acesso em 07 de Julho de 2022. Citado na página 4.

CHEN, J. *Commodities Exchange*. 2022. Disponível em: <https://www.investopedia.com/terms/c/commoditiesexchange.asp>. Acesso em 07 de Julho de 2022. Citado na página 4.

DALL'AGNOL, A. Iv congresso brasileiro de soja - resumos. In: _____. [S.l.]: Londrina: Embrapa Soja, 2014., 2006. cap. Prefacio. Citado na página 4.

DAS, S. China's evolving exchange rate regime. *International Monetary Fund*, 2019. Disponível em: <https://www.imf.org/en/Publications/WP/Issues/2019/03/07/Chinas-Evolving-Exchange-Rate-Regime-46649>. Citado 3 vezes nas páginas 3, 15 e 17.

DOWDY, S.; WEARDON, S.; CHILKO, D. *Statistics for Research*. [S.l.]: John Wiley & Sons, Inc, 2004. Citado na página 5.

EDUCAçãO, B. *Histórico - B3*. 2020. Disponível em: <https://ri.b3.com.br/pt-br/b3/historico/>. Acesso em 07 de Julho de 2022. Nenhuma citação no texto.

GEORGAKOPOULOS, H. *Quantitative trading with R: understanding mathematical and computational tools from a quant's perspective*. [S.l.]: Springer, 2015. Citado 4 vezes nas páginas 5, 6, 7 e 17.

HAUKE, J.; KOSSOWSKI, T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, De Gruyter Poland, v. 30, n. 2, p. 87, 2011. Citado na página 6.

HIRAKURI, M. H. Tecnologias de produção de soja. In: _____. [S.l.]: Londrina: Embrapa Soja, 2014., 2020. cap. Cap. 1 - O contexto econômico da produção de soja. Citado na página 4.

HIRAKURI, M. H.; LAZZAROTTO, J. J. O agronegócio da soja nos contextos mundial e brasileiro. Londrina: Embrapa Soja, 2014., 2014. Citado na página 2.

IPEA. *Plano de Dados Abertos*. 2022. Disponível em: <https://www.ipea.gov.br/portal/images/acesso-a-informacao/dados-abertos/anexo_port_188_pda_2021_2023_ipea.pdf>. Acesso em 10 de Novembro de 2022. Citado na página 8.

LIMA, V. et al. Granger causality in the frequency domain: derivation and applications. *Revista Brasileira de Ensino de Física*, 2020. Disponível em: <https://www.scielo.br/j/rbef/a/m4LwwHLvk7YwPNMhngNJQwp/?lang=en>. Citado na página 7.

LO, A. W.; MAMAYSKY, H.; WANG, J. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The journal of finance*, Wiley Online Library, v. 55, n. 4, p. 1705–1765, 2000. Citado na página 3.

MACROTRENDS. *Macrotrends*. 2022. Disponível em: <https://www.macrotrends.net/>. Acesso em 10 de Novembro de 2022. Citado na página 12.

MAGMA-CAPITAL-FUNDS. *What Is Quantitative Investing*. 2022. Disponível em: <https://magmacapitalfunds.com/what-is-quantitative-investing/>. Acesso em 10 de Julho de 2022. Citado na página 3.

MOBILIARIOS, C. de V. *Mercado Futuro*. 2022. Disponível em: <https://www.investidor.gov.br/menu/Menu_Investidor/derivativos/mercado_futuro.html>. Acesso em 07 de Julho de 2022. Nenhuma citação no texto.

MORITZ, S. *Gallery: Times Series Missing Data Visualizations*. 2022. Disponível em: <https://cran.r-project.org/web/packages/imputeTS/vignettes/gallery_visualizations.html>. Acesso em 10 de Novembro de 2022. Citado na página 10.

OLIVEIRA, G. de L. T. The geopolitics of brazilian soybeans. *The Journal of Peasant Studies*, Routledge, v. 43, n. 2, p. 348–372, 2016. Disponível em: <https://doi.org/10.1080/03066150.2014.992337>. Citado na página 5.

R-PROJECT. *What is R?* 2022. Disponível em: <https://www.r-project.org/about.html>. Acesso em 07 de Julho de 2022. Citado na página 7.

STATISTA. *Import volume of soybeans worldwide in 2021/22, by country*. 2022. Disponível em: <https://www.statista.com/statistics/612422/soybeans-import-volume-worldwide-by-country/>. Acesso em 30 de Agosto de 2022. Citado na página 2.

STATISTA. *Leading soybean producing countries worldwide from 2012/13 to 2021/22*. 2022. Disponível em: <https://www.statista.com/statistics/263926/soybean-production-in-selected-countries-since-1980/>. Acesso em 07 de Julho de 2022. Citado na página 2.

STONEX-BRASIL. *O que é commodity?* 2021. Disponível em: <https://www.mercadosagricolas.com.br/inteligencia/o-que-sao-commodities/>. Acesso em 10 de Julho de 2022. Citado na página 2.

TSAY, R. S. *Multivariate time series analysis: with R and financial applications*. [S.l.]: John Wiley & Sons, 2013. Citado na página 7.

VIGEN, T. *Spurious Correlations*. 2022. Disponível em: <https://www.tylervigen.com/spurious-correlations>. Acesso em 10 de Outubro de 2022. Citado na página 5.