**Relatório Do Trabalho Prático**

**Modelação e previsão de uma serie temporal**

José Pedro Ribeiro Ferreira Pinto – 201603713

Faculdade de Ciências da Universidade do Porto

Series Temporais e Previsão

**1-Resumo**

Neste trabalho foram analisadas as características de uma serie temporal de inspeções alimentares em Chicago de 2010 a 2019. Vários modelos ARIMA e de amaciamento exponencial foram testados, tentando escolher os melhores paramentos, com o objetive de descrição e previsão da serie. Após escolher os melhores modelos, previsões do último ano dos dados foram efetuados.

**1-Descrição do problema**

Este projeto tem como objetivo a compreensão das características e a previsão de valores futuros do número de inspeções alimentares em Chicago. A compreensão da evolução desta serie permitira analisar a eficácia dos serviços de saúde alimentar, identificar falhas, e no caso de ser necessário aplicar correções, seja isto através de uma alteração na legislação ou uma modificação no orçamento.

**1.1-Base de dados**

A base de dados obtida foi uma base de dados de inspeções alimentares encontrada no kaggle <https://www.kaggle.com/chicago/chicago-food-inspections#food-inspections.csv>.

A base de dados é atualizada continuamente, contendo registos até ao dia de hoje. Na altura de obtenção da base de dados está apenas continha valores até ao fim de outubro de 2019.

Esta base de dados é constituída por um conjunto de variáveis explicativas, de entre elas a data da inspeção.

Uma vez que os dados não são originalmente uma serie temporal, mas sim uma base de dados, diversas operações foram efetuadas para se poder trabalhar num contexto de series temporais.

As inspeções foram agregadas por mês e contadas, sendo o número de inspeções em cada mês os valores apresentados na serie temporal.

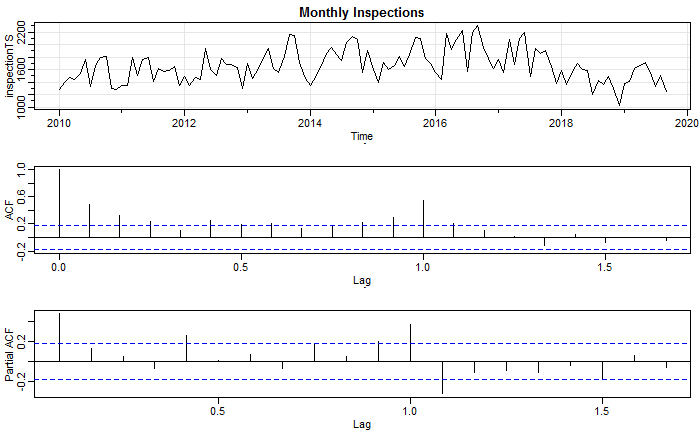
**1.1-Serie temporal**

As serie temporal tem um comprimento de 116 meses, começando a 1/2010 e acabando em 9/2019. Devido ao processamento prévio tem uma amostragem mensal.

Devido á atualização permanente da base de dados, o último mês não estava completo, como tal foi removido.

Houve uma alteração de legislação no início de 2018. Esta alteração está presente nos dados, alterando a distribuição e dificultando a previsão.

Abaixo é apresentado o gráfico da serie temporal que representa o número de inspeções alimentares efetuadas por mês, a função de autocorrelação e a função de autocorrelação parcial.



**Figura 1:** *Gráfico da serie temporal, função de autocorrelação e função de autocorrelação parcial.*

Na figura anterior (figura 1) é possível detetar a alteração dos dados no início de 2018. Também é possível observar nas funções de autocorrelação uma provável sazonalidade de 12 meses.

**2-Processamento da serie temporal**

Antes de se poder avançar com modelação algum reprocessamento da serie deve ser testado e efetuado em caso de ser necessário. Transformações para tornar a serie estacionaria, a variância constante ou simplificar.

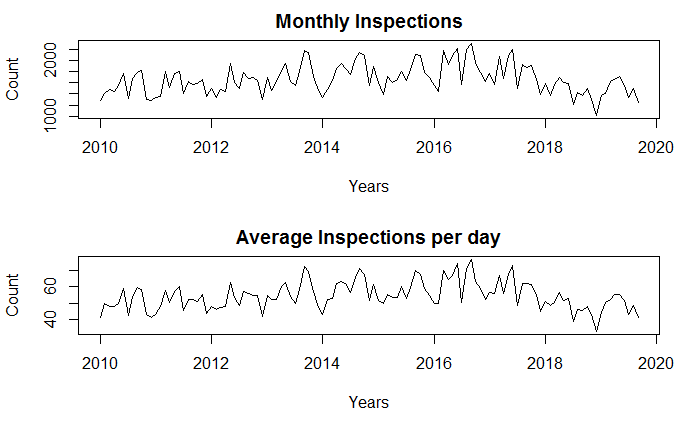
**2.1-Transformações**

Uma vez que a serie tem amostragem mensal, a divisão dos valores pelo número de dias do mês foi testada.

Abaixo encontram-se os gráficos da serie original e da serie transformada.

É possível observar que a transformação não teve o resultado pretendido, tendo ficado tanto ou mais complicada. Isto dá como indicação que o número de inspeções pouco varia com o número de dias do mês, e como tal que o número deve ser planeado desde o início.

Uma vez que a transformação não foi útil esta foi descartada, permanecendo, como tal a serie original.



**Figura 2:** *Gráfico da serie temporal original contra a serie temporal ajustada ao número de dias do mês.*

A serie não apresenta uma alteração da variância ao longo do período, como tal nenhuma transformação será testada com o objetivo de corrigir este problema.

**2.2-Diferenciação**

Através de uma análise gráfica a serie não aparenta ser estacionaria. Um teste de estacionaridade foi efetuado, o que confirmou a hipótese.

As funções de autocorrelação e autocorrelação parcial indicam uma sazonalidade de 12 meses, como tal, diferenciação de 12 meses (1 ano) foi testada.

A serie temporal obtida da diferenciação, tal como as suas funções de autocorrelação e autocorrelação parcial são apresentadas abaixo (figura 3).

A serie ainda não aparenta ser estacionaria, no entanto, nenhuma sazonalidade obvia é detetada.

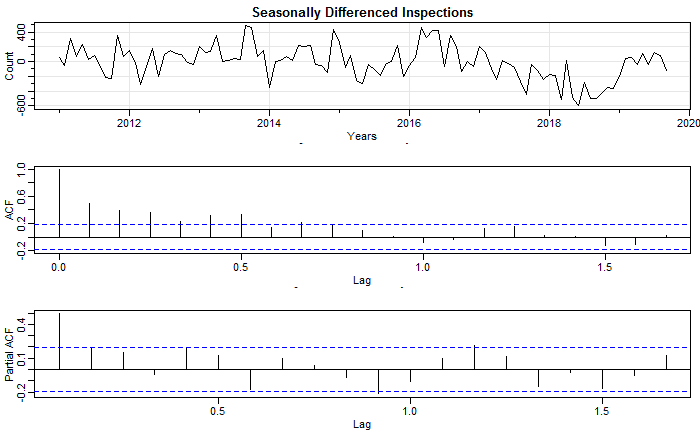
Outro teste foi efetuado para verificar a hipótese da serie não ser estacionaria, sendo esta hipótese confirmada.

Uma segunda diferenciação, desta vez normal, foi efetuada.

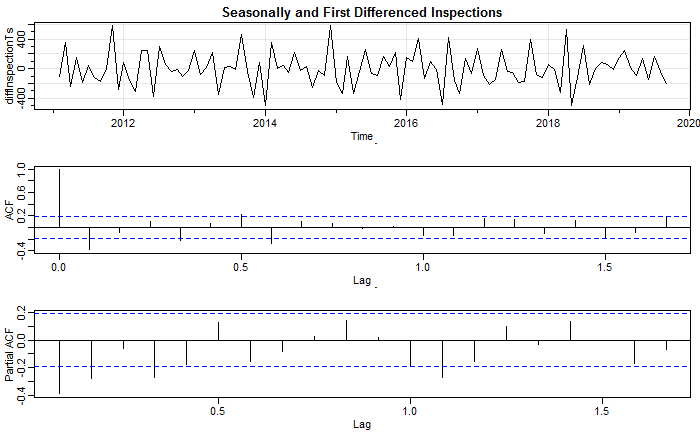
A serie temporal obtida desta segunda diferenciação, tal como as suas funções de autocorrelação e autocorrelação parcial são apresentadas abaixo (figura 4).

A serie obtida apos esta segunda diferenciação aparenta ser estacionaria, no entanto, para confirmar, tal como nos casos anteriores, um teste de estacionaridade foi efetuado.

O teste confirmou a ideia anteriormente apresentada.



**Figura 3:** *Gráfico, função de autocorrelação e função de autocorrelação parcial da serie com diferenciação sazonal de 12 meses (1 ano).*



**Figura 4:** *Gráfico, função de autocorrelação e função de autocorrelação parcial da serie com diferenciação sazonal de 12 meses (1 ano) e diferenciação normal.*

Uma vez que a serie era agora estacionaria e todas as transformações necessárias e uteis tinham sido aplicadas, procedeu-se e com a modelação.

**3-Modelação da serie temporal**

Com a serie estacionaria, todas as mudanças dos dados efetuadas e todas as transformações razoáveis testadas, começou-se com a modelação.

A primeira etapa foi a divisão da serie em duas secções diferentes, o último ano da serie constitui os dados de teste, enquanto que o resto dos dados (tudo exceto o último ano) constituem os dados de treino.

Vários métodos e modelos diferentes foram testados e validados. Estes métodos são, a escolha de modelos SARIMA com auxílio a analises gráficas e testes, modelos SARIMA escolhidos automaticamente e modelos de amaciamento exponencial escolhidos automaticamente e manualmente escolhidos.

**3.1-Seleção manual do modelo SARIMA**

Uma vez que os níveis de diferenciação tinham sido previamente obtidos, todos os modelos SARIMA usam essa diferenciação, como tal todos os modelos testados são do tipo SARIMA(x,1,y)(z,1,w)12.

Primeiro o parte sazonal do modelo foi considerada, alterando apenas a parte sazonal e mantendo a parte ARMA constante, com os valores de (0,0). Sem indícios fortes dos fatores a escolher, vários foram testados. Os fatores testados foram todos aqueles com os valores inferiores a 2, isto é, desde (0,0) até (2,2).

O modelo escolhido para proceder foi aquele que apresentou o melhor BIC (valor menor), tendo este modelo, como parte sazonal, (2,1).

Após se ter escolhido a parte sazonal, procedeu-se com a seleção da parte do modelo ARMA. Através de uma análise gráfica (figura 4), o modelo mais ajustado aparentou ser um ARMA (2,1). Todos os modelos ARMA com uma distância de 1 do anteriormente mencionado foram testados, isto é, os modelos deste (1,0) até (3,2).

O modelo com o melhor BIC (menor valor) foi escolhido, este sendo um ARMA (1,1).

Depois do modelo final ser escolhido, os resíduos foram testados, gráfica e estatisticamente, para correlação e homoestaticidade.

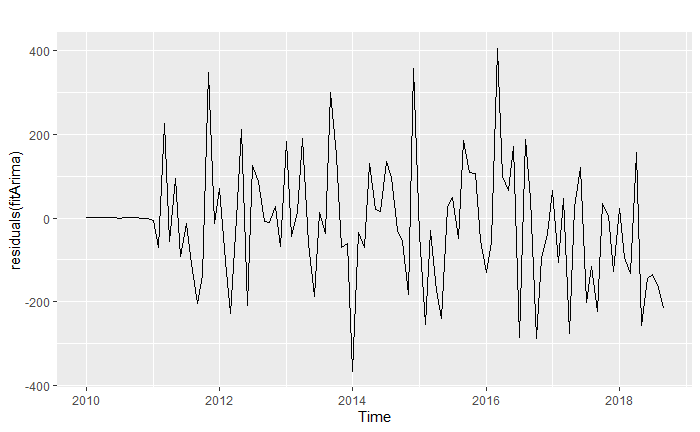
Os resíduos estão apresentados abaixo. Tanto o gráfico dos resíduos como as funções de autocorrelação e de autocorrelação parcial indicam resíduos independentes e normalmente distribuídos.

Um teste de box-pierce foi efetuado para verifica as conclusões. O valor-p obtido foi de 0.8652, o que nos leva a rejeitar a hipótese de correlação, confirmando a ideia anterior.

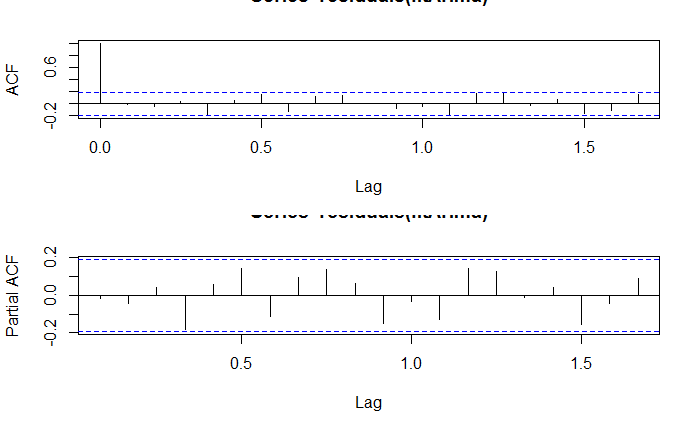
Box-Pierce test

data: resid(fitArima)

X-squared = 0.028811, df = 1, p-value = 0.8652



**Figura 5:** *Gráfico, dos resíduos do modelo SARIMA manualmente selecionado.*



**Figura 6:** *Função de autocorrelação e função de autocorrelação parcial dos resíduos do modelo SARIMA manualmente selecionado.*

O modelo final obtido foi SARIMA(1,1,1)\*(2,1,1)12.

Com a seleção manual do modelo SARIMA concluída procedeu-se para o método seguinte.

**3.2-Seleção automática do modelo SARIMA**

Para comparar as diferenças entre algoritmos de seleção automática de modelos e modelos escolhidos à mão, o segundo modelo escolhido foi obtido por seleção automática.

A função utilizada foi auto.arima que utiliza técnicas de otimização não linear para escolher os modelos. Esta função compara diferentes modelos ou com o AIC ou com o BIC. BIC foi escolhido por ter maior precisão.

No fim o modelo obtido foi um SARIMA(0,1,1)\*(2,1,0)12. Bastante diferente do modelo escolhido manualmente. Tal como anteriormente os resíduos foram testados graficamente e estatisticamente.

Abaixo encontram-se os gráficos dos resíduos tal como as funções de autocorrelação e de autocorrelação parcial.

Tanto do gráfico como das funções de autocorrelação e de autocorrelação parcial se observa a inexistência de correlação dos resíduos. Tal comprova o facto dos dados e toda a informação terem sido corretamente capturados.

Para verificar o que tinha sido constatado, também foi efetuado um teste de independência e normalidade.

O valor obtido pelo teste de box-pierce confirma o que tinha sido afirmado anteriormente, com um valor-p de 0.9738, não se pode rejeitar a inexistência de correlação dos resíduos.

Box-Pierce test

data: resid(fitArimaAuto)

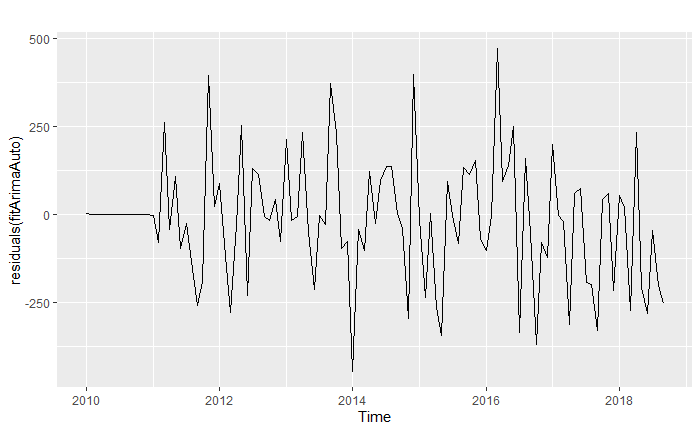
X-squared = 0.0010824, df = 1, p-value = 0.9738

De facto, o valor-p obtido é superior ao obtido anteriormente (do modelo SARIMA manualmente selecionado), o que pode indicar que o modelo automático é melhor que o modelo manualmente selecionado.

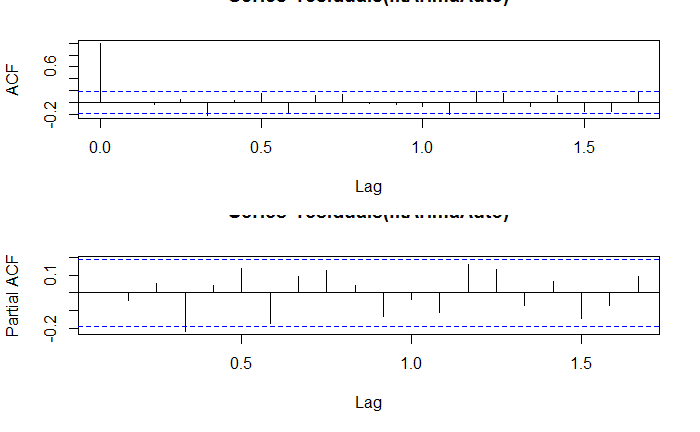
A obtenção de um modelo automático com o critério de comparação sendo o AIC também foi testada, no entanto o modelo obtido foi o mesmo do obtido pelo BIC (método anteriormente mencionado). Como tal foi descartado do código.

Depois de obter dois modelos diferentes alguns padrões emergem. Os resíduos de 2010 a 2011 em ambos o modelo tem um valor de quase 0. Também, nos dois modelos, apesar de estes serem bastante diferentes, os resíduos são extremamente semelhantes. Com valores elevados e salientes ligeiramente antes de 2012, um pico negativo em 2014 e outro pico positivo pouco depois de 2016.

Dado o facto de os modelos serem tão diferentes, e, no entanto, os resíduos tão semelhantes, isto poderá querer dizer que os padrões nos dados foram corretamente obtidos, mantendo apenas valores aleatórios.



**Figura 7:** *Gráfico dos resíduos do modelo SARIMA automaticamente selecionado.*



**Figura 8:** *Função de autocorrelação e função de autocorrelação parcial dos resíduos do modelo SARIMA automaticamente selecionado.*

Uma vez obtidos os dois modelos SARIMA, por modelação manual e automática, passou-se para a modelação de um novo tipo de modelos, modelos de amaciamento exponencial.

**3.3-Amaciamento exponencial**

Ao contrário dos modelos SARIMA, anteriormente obtidos, os modelos de amaciamento exponencial não foram diretamente abordados nas aulas, como tal uma explicação teórica e pratica com mais algum detalhe será fornecida.

Numa primeira instância, devido há falta de compreensão e intuição teórica e pratica sobre a seleção de modelos de amaciamento exponencial, uma vez que estas não foram abordados nas aulas, procedesse com apenas a seleção automática do modelo.

Devido a más previsões do modelo automático, após comunicação com o docente, outro modelo, desta vez escolhido há mão foi testado.

**3.3.1-Descrição de amaciamento exponencial**

Amaciamento exponencial designa um conjunto de métodos de modelação e previsão de series temporais que são essencialmente não paramétricos, no sentido em que não se encontra um modelo para os dados, que é posteriormente iterado até obter previsões. Este tipo de modelos também não requere que sejam assumidos fatores como, por exemplo, estacionaridade. As previsões são obtidas através dos dados com pesos que decaem exponencialmente.

Existem bastantes variações de modelos de amaciamento exponencial. A mais simples apenas tem um parâmetro alfa. Com valores mais elevados de alfa as observações mais distantes ganhão mais peso, enquanto que com valores baixos, estas observações rapidamente perdem influência. O valor de alfa tem de estar entre 0 e 1 para o modelo ser invertível. Com um valor de 0 todas as previsões são 0, enquanto que com o valor de um todas as previsões são a media de todas as anteriores.

A fórmula básica para a previsão é:

y^T+1|T = αyT + α(1 − α)yT−1 + α(1 − α)^2yT−2 + ⋯

Os modelos que serão utilizados não são este modelo mais simples, mas modelos com componentes sazonais e componentes de tendência(*trend*). Nestes modelos, mais dois parâmetros tem de ser estimados, beta e gama, o primeiro para a tendência e o segundo para a sazonalidade.

**3.3.2-Modelo de amaciamento exponencial escolhido automaticamente**

Para a seleção do modelo e parametrização do mesmo foi utilizada a função ets, que automaticamente escolhe o modelo e obtém os valores dos parâmetros alfa, beta e gama.

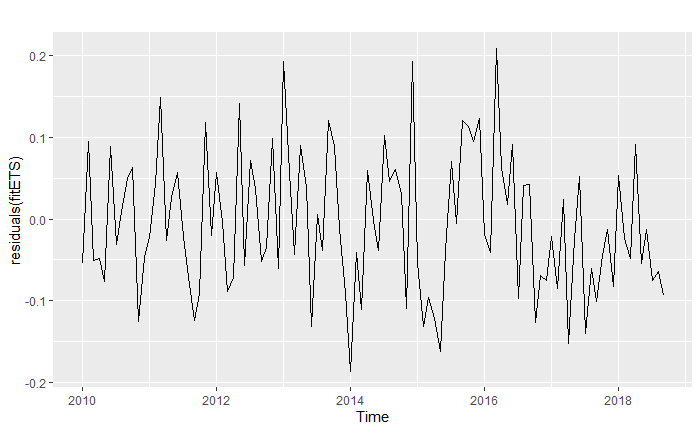
O modelo final obtido por este método é um modelo ETS(M,A,M), tendo este uma componente de decaimento multiplicativa, uma componente de tendência aditiva e uma componente de sazonalidade multiplicativa.

Os testes típicos sobre os resíduos foram efetuados, para determinar a existência ou não de correlação e homoestaticidade. Os gráficos dos resíduos, das funções de autocorrelação e de autocorrelação parcial são apresentadas abaixo. Um teste de box-pierce, tal como anteriormente também foi efetuado.

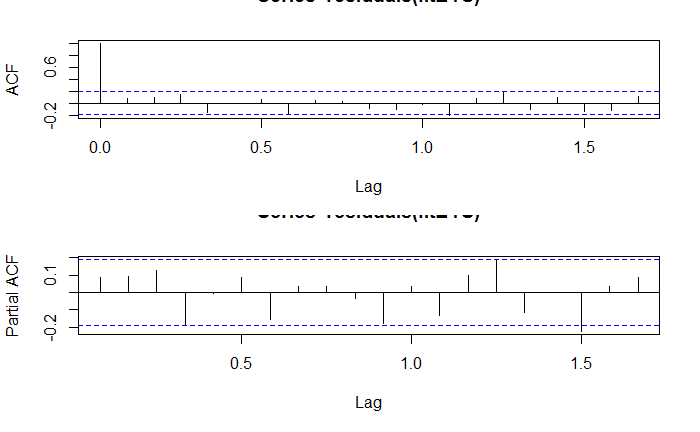
Box-Pierce test

data: resid(fitETS)

X-squared = 0.81668, df = 1, p-value = 0.3662



**Figura 9:** *Gráfico dos resíduos do modelo ETS automaticamente selecionado.*



**Figura 10:** *Função de autocorrelação e função de autocorrelação parcial dos resíduos do modelo ETS automaticamente selecionado.*

Do gráfico, das funções de autocorrelação e de autocorrelação parcial e do teste efetuado podemos concluir que a informação foi corretamente captada e que como tal os resíduos são não correlacionados.

Como tal procedemos agora para o próximo e último modelo.

**3.3.3-Modelo de amaciamento exponencial escolhido manualmente**

Tal como mencionado anteriormente, os resultados de previsão obtidos pelo modelo de amaciamento exponencial escolhido automaticamente foram bastante maus e como tal, um novo modelo escolhido há mau foi testado.

Apos alguma ponderação sobre a análise gráfica e estatística previamente efetuada, o modelo que aparenta melhor se adaptar aos dados foi escolhido.

Este modelo escolhido foi um ETS(A,A,A).

Também para este modelo foi usada a função ets, mas neste caso o modelo foi fixado, não permitindo a seleção automática.

Mais uma vez, tal como para os outros modelos, foi efetuada uma análise gráfica dos resíduos e um teste de corelação dos mesmos. O gráfico dos resíduos e as funções de autocorrelação e autocorrelação parcial encontram-se abaixo.

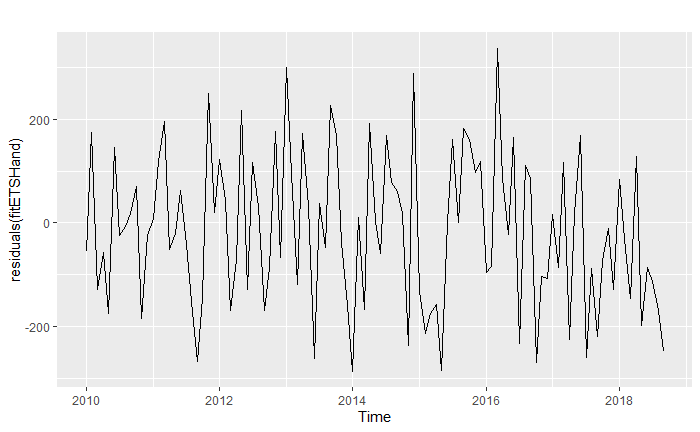
Um teste de box-pierce foi efetuado.

Box-Pierce test

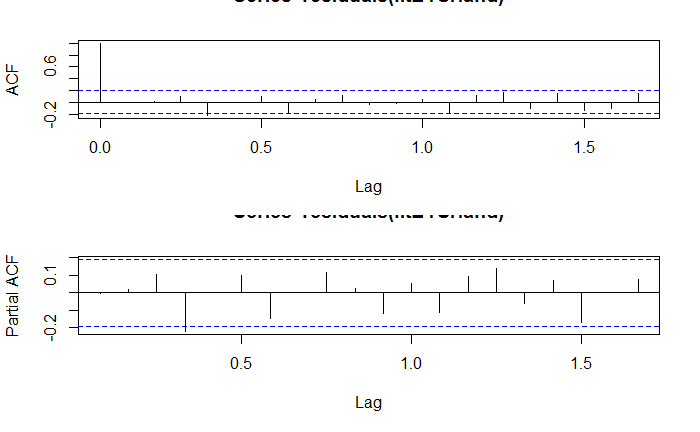
data: resid(fitETSHand)

X-squared = 0.0037095, df = 1, p-value = 0.9514

Mais uma vez, tanto os resíduos e os gráficos confirmam que a informação foi corretamente captada e que os resíduos não são correlacionados.



**Figura 11:** *Gráfico dos resíduos do modelo ETS manualmente selecionado.*



**Figura 12:** *Função de autocorrelação e função de autocorrelação parcial dos resíduos do modelo ETS manualmente selecionado.*

**4-Previsão da Serie Temporal**

Tendo todos os modelos escolhidos parametrizados e testados procedeu-se com a previsão da serie temporal.

Tal como referido anteriormente a serie foi dividida em duas partes, a primeira com toda a serie exceto o último ano e a segunda com apenas este último ano. Todos os modelos foram treinados utilizando a primeira parte e, nesta secção procedeu-se às previsões da segunda parte.

Em todos os gráficos nesta secção a azul encontram-se os valores previstos e os intervalos de confiança, enquanto que a preto os valores observados.

**4.1-Previsões do modelo SARIMA escolhido manualmente**

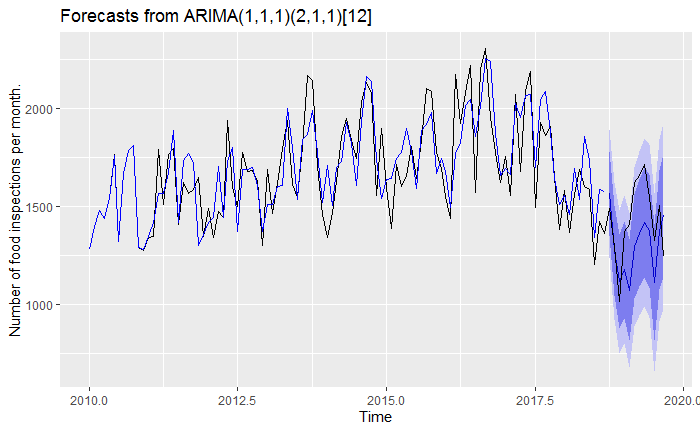
O gráfico contendo os valores ajustados e as previsões encontrasse abaixo.

É possível observar um bom ajustamento dos dados, com os valores previstos seguindo de perto a distribuição verdadeira.

Quase todos os valores se encontram dentro das bandas de confiança, como se deveria esperar para um modelo bem ajustado aos dados.

No entanto as bandas são bastante largas, quando tendo em consideração as flutuações dos valores da serie. Isto poderá ser um problema, diminuindo o valor das previsões e a sua utilidade.

Não se pode, no entanto, esquecer a previamente referida mudança de legislação em 2018 que causou uma disrupção nos dados. É bastante possível e provável que a obtenção de mais dados depois desta disrupção permita obter um melhor modelo com um intervalo de confiança mais pequeno.

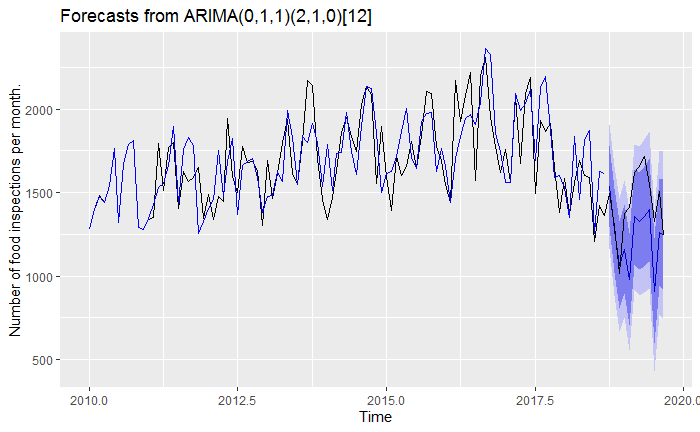


**Figura 12:** *Gráfico dos valores previstos pelo modelo SARIMA manualmente selecionado contra os valores verdadeiros da serie temporal.*

**4.2-Previsões do modelo SARIMA escolhido manualmente**

O gráfico contendo os valores ajustados e as previsões encontrasse abaixo.

É possível observar um bom ajustamento dos dados, com os valores previstos seguindo de perto a distribuição verdadeira.



**Figura 13:** *Gráfico dos valores previstos pelo modelo SARIMA automaticamente selecionado contra os valores verdadeiros da serie temporal.*

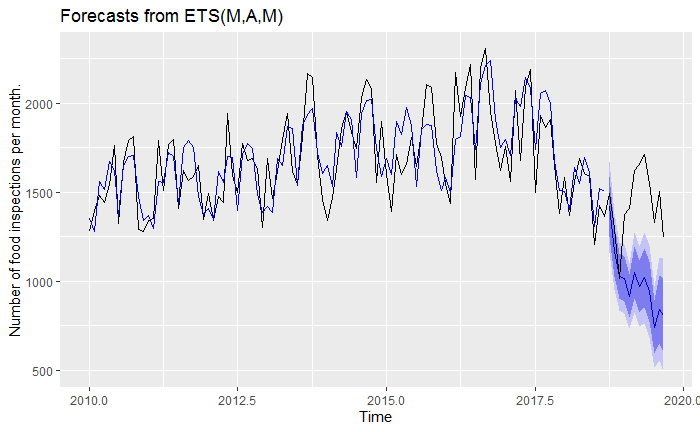
Quase todos os valores se encontram dentro das bandas de confiança, como se deveria esperar para um modelo bem ajustado aos dados.

No entanto as bandas são bastante largas, quando tendo em consideração as flutuações dos valores da serie. Isto poderá ser um problema, diminuindo o valor das previsões e a sua utilidade.

Não se pode, no entanto, esquecer a previamente referida mudança de legislação em 2018 que causou uma disrupção nos dados. É bastante possível e provável que a obtenção de mais dados depois desta disrupção permita obter um melhor modelo com um intervalo de confiança mais pequeno.

**4.3-Previsões do modelo de amaciamento exponencial escolhido automaticamente**

O gráfico contendo os valores ajustados e as previsões encontrasse abaixo.



**Figura 14:** *Gráfico dos valores previstos pelo modelo de amaciamento exponencial automaticamente selecionado contra os valores verdadeiros da serie temporal.*

Ao contrário dos modelos SARIMA, o modelo de amaciamento exponencial obtido automaticamente tem muito, mas previsões, com os valores previstos extremamente longe dos valores verdadeiros.

Não só os valores previstos estão muito longe, mas também, grande parte dos valores verdadeiros não estão dentro das bandas de confiança. Tal indica um modelo muito mal ajustado aos dados.

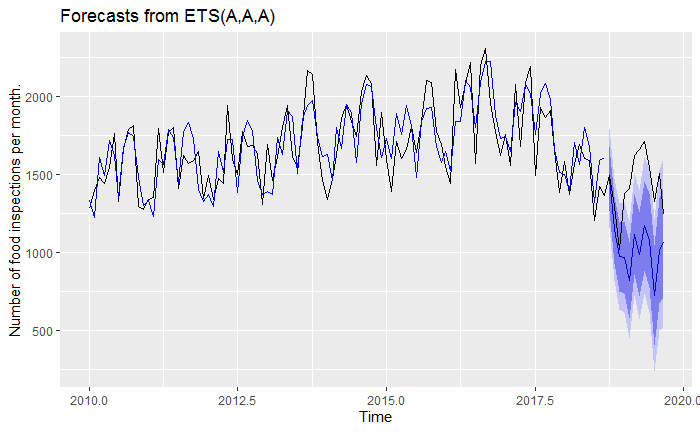
Não se pode, no entanto, esquecer a previamente referida mudança de legislação em 2018 que causou uma disrupção nos dados. É bastante possível e provável que a obtenção de mais dados depois desta disrupção permita obter um melhor modelo com os valores previstos mais próximos dos verdadeiros.

**4.4-Previsões do modelo de amaciamento exponencial escolhido manualmente**

Devido às más previsões obtidas anteriormente, no modelo de amaciamento exponencial escolhido automaticamente, foi testado um modelo de amaciamento exponencial escolhido manualmente.

O modelo testado foi um ETS(A,A,A).

O gráfico contendo os valores ajustados e as previsões encontrasse abaixo.



**Figura 15:** *Gráfico dos valores previstos pelo modelo de amaciamento exponencial manualmente selecionado contra os valores verdadeiros da serie temporal.*

O novo modelo apresenta melhores resultados que o modelo de amaciamento exponencial escolhido automaticamente, no entanto continua a ser substancialmente pior que os modelos SARIMA.

Tal com o modelo anterior, não só os valores previstos estão muito longe dos verdadeiros, mas também, grande parte dos valores verdadeiros não estão dentro das bandas de confiança. Tal indica um modelo mal ajustado aos dados.

Não se pode, no entanto, esquecer a previamente referida mudança de legislação em 2018 que causou uma disrupção nos dados. É bastante possível e provável que a obtenção de mais dados depois desta disrupção permita obter um melhor modelo com os valores previstos mais próximos dos verdadeiros.

**5-Resultados e Conclusões**

Nesta secção são apresentados os resultados finais, o quão próximo se ficou dos valores verdadeiros, e as conclusões sobre os mesmos resultados e o que foi alcançado.

**5.1-Resultados**

Dos modelos testados os modelos SARIMA são sem dúvida os com melhores resultados, no entanto apenas da análise gráfica não é possível saber qual dos dois é melhor. Como tal, para selecionar o melhor modelo foi utilizada a precisão, sendo o melhor modelo aquele com melhor precisão.

Os valores das diferentes métricas são apresentados nas duas tabelas abaixo.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil's U |
| Training set | -16.55065 | 143.28 | 106.43 | -1.583 | 6.320 | 0.5988 | -0.01656 | NA |
| Test set | 132.80 | 220.04 | 196.66 | 8.217 | 13.477 | 1.106 | 0.42002 | 0.9828694 |

**Tabela 1:** *Tabela das diferentes métricas de precisão do modelo SARIMA manualmente selecionado.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil's U |
| Training set | -19.00285 | 174.33 | 127.27 | -1.764 | 7.520 | 0.7160 | 0.00321 | NA |
| Test set | 194.33 | 257.99 | 210.73 | 12.844 | 14.202 | 1.1855 | 0.33272 | 1.150284 |

**Tabela 2:** *Tabela das diferentes métricas de precisão do modelo SARIMA automaticamente selecionado.*

Das métricas apresentadas acima é possível concluir que o modelo SARIMA selecionado manualmente é melhor para estes dados que o modelo selecionado automaticamente.

Todas as medidas apresentadas vão a favor desta conclusão.

**5.2-Conclusões**

No fim de todo o processo não foi possível obter os resultados pretendidos. Para os resultados poderem ser usados para o seu propósito original é necessário que estes sejam bastante precisos, no entanto tal não provou ser o caso. O melhor modelo obtido, apesar de estar modelado para os dados, apresenta intervalos de confiança demasiado elevados e uma distância dos valores previstos aos valores reais também demasiado elevada.

Foi possível concluir que os modelos de amaciamento exponencial não funcionam bem para este tipo de dados, dado que todos os modelos testados apresentam péssimos resultados.

Os modelos SARIMA, apesar de não conseguirem fornecer os resultados necessários, funcionarão bem com estes dados, fornecendo previsões razoáveis.

Devido ao tipo de resultados obtidos, é esperado que para melhorar seja necessário ou modelos mais complexos ou mais dados sobre a serie.