TÉCNICO LISBOA

INSTITUTO SUPERIOR TÉCNICO

MESTRADO EM MATEMÁTICA E APLICAÇÕES

Níveis de Altura do Mar Báltico e Índice do Mercado de Acções NASDAQ

Séries Temporais

Catarina Padrela Loureiro [80868]
Matilde Tristany Farinha [81240]

Professor: Manuel Scotto

2ºSemestre 2017/2018

Conteúdo

1 Introdução

O primeiro conjunto de dados com que se vai trabalhar consiste das medições diárias dos níveis de altura do Mar Báltico feitas desde o início do mês de Janeiro de 1979 até ao fim do mês de Dezembro de 2005. Estas medições foram feitas para sete cidades banhadas pelo Mar Báltico, nomeadamente, as cidades de Furuögrund (FUR), Hornbaek (HOR), Ölands Norra Udde (OLA), Stockholm (STO), Gedser (GED), Kungsholmsfort (KUN) e Ratan (RAT). Nos dias de hoje, devido às alterações climáticas que se têm verificado, estudar as consequências que tais alterações provocam no meio que habitamos e tentar prever a evolução deste torna-se um assunto de grande importância. Isto, na medida em que poderemos estar prevenidos destas alterações de modo a sermos capazes de lidar com as suas consequências, evitando possíveis catástrofes.

O segundo conjunto de dados com que se vai trabalhar consiste no índice de mercado de acções chamado NASDAQ. As medições que temos são feitas diariamente e foram registadas desde 23 de Fevereiro de 2015 até 22 de Fevereiro de 2018. Este conjunto de dados situa-se na área de finanças em que é extremamente importante prever a evolução do mercado e o impacto no futuro deste devido a alterações correntes e passadas.

Em relação ao primeiro conjunto de dados procurou-se ajustar um modelo $SARIMA(p,q,d) \times (P,D,Q)_S$ com período S ($d, D \in \mathbb{N}$). Neste modelo considera-se (Y_t) um processo $SARIMA(p,q,d) \times (P,D,Q)_S$ (B é o *backwards operator*) se a série diferenciada $X_t = \nabla^d \nabla_S^D Y_t \equiv (1-B)^d (1-B^S)^D Y_t$ é um processo ARMA causal que satisfaz $\Psi(B)\Phi_P(B^S)X_t = \theta(B)\Theta_Q(B^S)Z_t$, onde $Z_t \sim WN(0, \sigma_Z^2)$, $\Psi(B)$ e $\theta(B)$ são polinómios de graus p e q, respetivamente, e $\Phi_P(B^S)$ e $\Theta_Q(B^S)$ são polinómios de grau P e Q, respetivamente, e com período S. Relembre-se que (X_t) é um processo $ARMA(p,q)$ se $X_t = \sum_{i=1}^{p} \psi_i X_{t-i} + \sum_{j=1}^{q} \theta_j Z_{t-j} + Z_t$, onde $Z_t \sim WN(0, \sigma_Z^2)$, $\psi_p \neq 0$ e $\theta_q \neq 0$.

Por outro lado, sendo que o segundo conjunto de dados são de cariz financeiro, a estes tenta-se ajustar um modelo de uma família diferente do anterior. Isto porque no caso de modelos da família ARMA o que se tenta modelar é o valor esperado condicionado do processo dado o passado, sendo que se assume que a variância condicionada ao passado se mantém constante. Mas e se de facto se verifica que a volatilidade não é constante? Ou seja, caso a variância condicionada ao passado varie, a família anterior de modelos não se adequa. Este é na verdade o caso das séries de cariz financeiro. Deste modo, introduz-se a família de modelos GARCH que já acomoda o efeito ARCH (heteroscedasticidade condicional auto-regressiva). Relembre-se que um modelo $GARCH(p,q)$ está definido como:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i X_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2 \end{cases}$$

onde $(Z_t) \sim \mathcal{N}(0, \sigma_Z^2 = 1)$, $a_o > 0$, $a_1, \ldots, a_{p-1} \geqslant 0$, $a_p > 0$, $b_1, \ldots, b_{q-1} \geqslant 0$ e $b_q > 0$.

Antes de se proceder à análise de dados relembre-se as seguintes medidas de selecção de modelos (medem a adequação dos modelos aos dados, sendo que um modelo mais adequado é o que tem valores AIC e BIC mínimos):

$$AIC := -2log(L(\hat{\psi}, \hat{\theta}, \hat{\sigma_Z^2})) + 2(p+q+1)$$

$$BIC := nlog(\hat{\sigma_Z^2}) + klog(n)$$

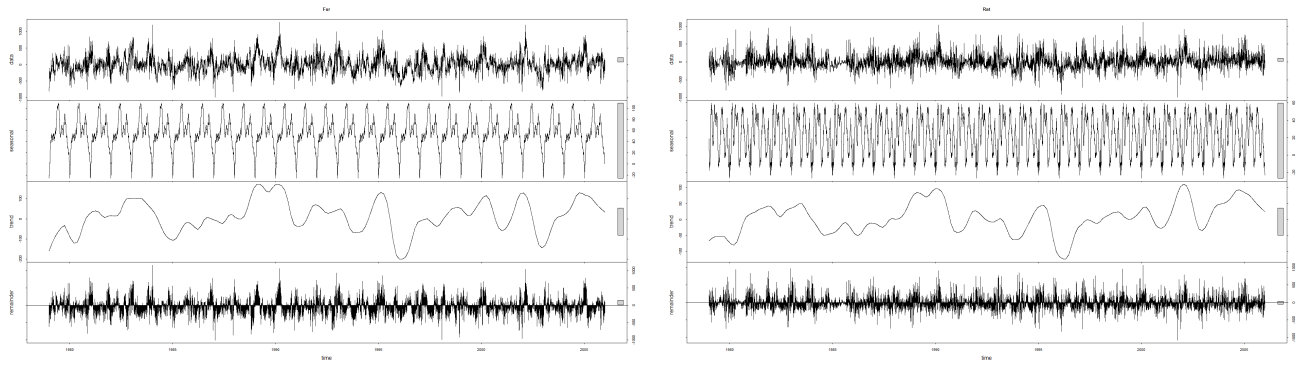
Mais ainda, todos os resultados aqui apresentados foram obtidos recorrendo ao software *R*, recorrendo em especial aos pacotes `forecast`, `fGarch`, `car`, e `aTSA`.

2 Resultados - Níveis de Altura do Mar Báltico

2.1 Análise Inicial dos Dados

Numa análise inicial dos dados começou-se por realizar a decomposição STL (Seasonal Trend decomposition based on Loess) das sete séries: $Y_t = T_t + S_t + X_t$. Na Figura 1, encontram-se os gráficos obtidos desta decomposição (para duas das séries, FUR e HOR, sendo que para as restantes os gráficos são semelhantes). Nestes gráficos tem-se a série original (Y_t), a sazonalidade (S_t), a tendência (T_t) e por fim o que resta após da decomposição (X_t). Analisando estes gráficos, conclui-se que as séries têm uma forte componente de sazonalidade, sendo que a tendência parece não exercer um papel relevante. Com efeito, perece haver evidência de que os dados apresentam sazonalidade anual.

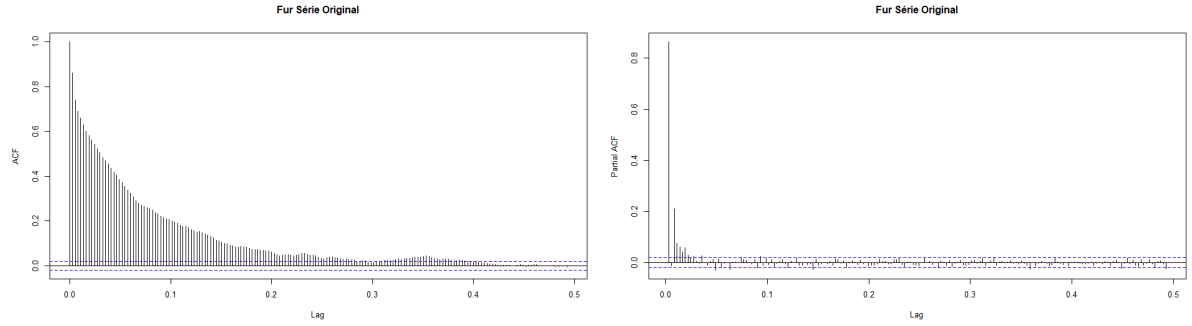
De seguida, pretende-se determinar se (X_t) é de facto estacionário e para tal realizou-se o teste de Dickey-Fuller aumentado (`adf.test`), onde a hipótese nula é (X_t) não ser estacionária (logo pretende-se rejeitar a hipótese nula). Para todas as séries, o teste devolveu um valor-p menor que 0.01, o que indica que se pode sempre rejeitar a hipótese nula para os níveis de significância usuais (1%, 5% e 10%). Assim, conclui-se que as séries (X_t) obtidas após a decomposição STL são estacionárias.



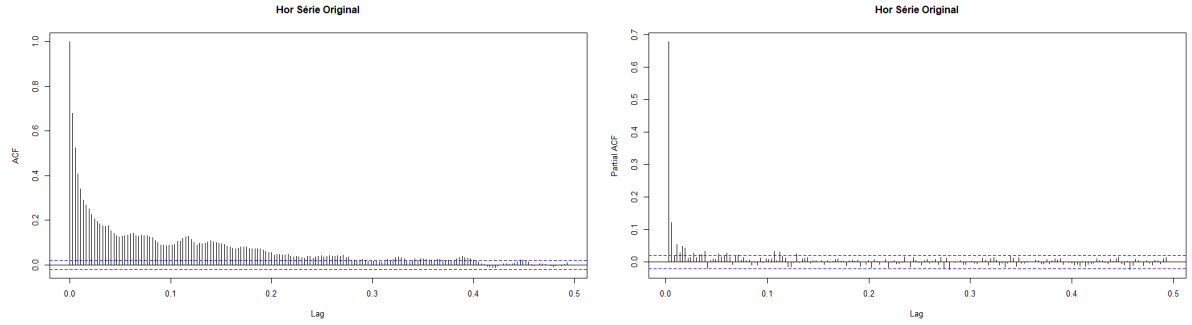
(a) Série FUR. (b) Série HOR.

Figura 1: Decomposição STL para duas das localizações (série original, tendência, sazonalidade e o restante respectivamente por esta ordem). Escolheu-se apresentar apenas para estas duas, visto que à primeira vista a decomposição é relativamente semelhante para as outras localizações.

Seguidamente, procedeu-se a uma análise dos gráficos da ACF (Auto-Correlation Function, obtida com o comando `acf`) e da PACF (Partial ACF, usando o comando `pacf`) de (Y_t). Na Figura 2, encontram-se a título de exemplo estes gráficos para as séries FUR e HOR (para as restantes séries foram obtidos gráficos muito semelhantes). Analisando os gráficos da ACF, observa-se que esta decresce de forma acentuada e aparentemente exponencial, indicando assim alguma evidência da necessidade de modelos de memória curta, nomeadamente modelos ARMA, em detrimento de modelos ARIMA (modelo de memória muito longa).



(a) Série FUR.



(b) Série HOR.

Figura 2: ACF e PACF de (Y_t) para as séries FUR (a) e HOR (b).

2.2 Transformações dos Dados

Os dados em estudo apresentam uma variância aproximadamente constante. Tal facto, em conjunto com a análise inicial dos dados, levou a que não se procedesse a uma transformação destes.

2.3 Identificação das Ordens de Dependência e do Grau de Diferenciação

Numa primeira análise, os dados aparentavam ter sazonalidade. Desse modo, recorreu-se ao uso do comando `auto.arima` para estimar as ordens de dependência e calcular a respetiva adequação dos modelos seleccionados às séries (AIC e BIC) de duas formas distintas. Na primeira fixou-se o termo de diferenciação sazonal como um ($D = 1$) (forçar o modelo a admitir que de facto existe uma sazonalidade) e aplicou-se sobre a série original (Y_t). Noutra, não se fixou o parâmetro de diferenciação e aplicou-se o comando sobre o restante da decomposição STL da série (X_t). Contudo, após correr o comando `auto.arima` com o parâmetro de sazonalidade forçado a diferenciar uma vez a série original (Y_t) (obtendo como output uma sazonalidade de 365 dias), verificou-se que pelas medidas de seleção de ordem dos modelos os valores de adequação do modelo como o AIC e o BIC tinham valores superiores aos obtidos pelo `auto.arima` aplicado à decomposição STL da série (X_t) sem forçar o parâmetro de sazonalidade. Desse modo, após alguma exploração, escolheu-se adoptar os modelos devolvidos pelo `auto.arima` após a decomposição STL da série conjuntamente com os respectivos parâmetros estimados já que estes apresentavam os menores valores das medidas de selecção como o AIC e o BIC. Recorrendo, então, à decomposição STL, obtivemos as ordens de dependência dos modelos para (X_t) associados a cada série representados na Tabela 1. Inicialmente, chegou-se a experimentar testar usar diretamente a série original, obtendo-se mesmo um modelo ARIMA para o caso da série HOR. No entanto, os valores AIC e BIC obtidos para estes modelos foram novamente superiores aos dos modelos que se decidiu adoptar posteriormente.

	Modelo	**AIC**	**BIC**
FUR	ARMA(2,2)	122305.1	122348.3
HOR	ARMA(3,2)	122304.8	122347.9
OLA	ARMA(3,0)	108968.6	109004.6
STO	ARMA(4,3)	98860.34	98925.1
GED	ARMA(2,3)	126845.7	126896
KUN	ARMA(2,2)	116754.8	116798
RAT	ARMA(3,2)	120342	120392.4

Tabela 1: Modelos ARMA ajustados às séries após decomposição STL (X_t) e respectivas medidas de selecção dos modelos.

Com o intuito de fazer uma pesquisa mais abrangente em relação às ordens de dependência e grau de diferenciação, voltou-se a correr o comando `auto.arima` no restante da decomposição STL (X_t) mas desta vez com `stepwise=FALSE`, o que se traduz numa pesquisa sobre todos os modelos, implicando uma quantidade de tempo superior necessário para obter resultados. Estes resultados não são aqui apresentados uma vez que foi considerado que o tempo que o comando demorava a correr não compensava a possível melhoria dos modelos já ajustados. Com isto, queremos dizer que ao correr o comando com `stepwise=FALSE` obtemos em quatro das séries, num total de sete, modelos ajustados diferentes dos da Tabela 1. Nomeadamente, foram obtidos modelos diferentes para as séries FUR, STO, KUN e RAT, cujas ordens de dependência obtidas foram $(1,3)$, $(2,3)$, $(1,3)$ e $(1,3)$, respectivamente. No entanto, relativamente às medidas de selecção (AIC e BIC), estes novos modelos apresentaram valores que se verificarem relativamente pouco menores ou relativamente pouco maiores quando comparados com os obtidos para os modelos na Tabela 1. Com isto quer-se dizer que em geral verificou-se uma diferença da ordem de uma unidade em medidas da ordem de 10^5 ou 10^4 unidades (máxima diferença verificada foi da ordem de 10 unidades). Tendo em conta isto, e o facto do grande aumento na quantidade de tempo necessária para obter os modelos, estes resultados não serão os considerados no projecto e os da Tabela 1 mantém-se como os modelos ajustados finais.

2.4 Estimação Paramétrica

A estimação dos parâmetros dos modelos seleccionados anteriormente com o uso do comando `auto.arima` é feita via *conditional sums of squares* (CSS). Os modelos ajustados são então:
FUR: $X_t = 1.1967X_{t-1} - 0.2340X_{t-2} - 0.3575Z_{t-1} - 0.2577Z_{t-2} + Z_t$, onde $Z_t \sim \mathcal{WN}(0, 14227)$;
HOR: $X_t = 0.9897X_{t-1} + 0.2377X_{t-2} - 0.2783X_{t-3} - 0.4127Z_{t-1} - 0.3786Z_{t-2} + Z_t$, onde $Z_t \sim \mathcal{WN}(0, 18950)$;
OLA: $X_t = 1.0009X_{t-1} - 0.1882X_{t-2} + 0.1236X_{t-3} + Z_t$, onde $Z_t \sim \mathcal{WN}(0, 3680)$;
STO: $X_t = 2.0285X_{t-1} - 1.1988X_{t-2} + 0.0413X_{t-3} + 0.1165X_{t-4} - 0.6991Z_{t-1} - 0.3299Z_{t-2} + 0.3793Z_{t-3} + Z_t$, onde $Z_t \sim \mathcal{WN}(0, 1320)$;
GED: $X_t = 0.8270X_{t-1} + 0.0215X_{t-2} - 0.2647Z_{t-1} - 0.1586Z_{t-2} - 0.0571Z_{t-3} + Z_t$, onde $Z_t \sim \mathcal{WN}(0, 22546)$;
KUN: $X_t = 1.0583X_{t-1} - 0.1198X_{t-2} - 0.3042Z_{t-1} - 0.1505Z_{t-2} + Z_t$, onde $Z_t \sim \mathcal{WN}(0, 8104)$;

RAT: $X_t = 1.2143X_{t-1} - 0.3486X_{t-2} + 0.0921X_{t-3} - 0.4031Z_{t-1} - 0.1277Z_{t-2} + Z_t$, onde $Z_t \sim \mathcal{WN}(0, 11658)$;

Note-se que $\mathcal{WN}$ é abreviação de *White Noise* onde o valor esperado é sempre zero e a variância é finita.

2.5 Diagnóstico dos Resíduos e Validação dos Modelos Escolhidos

Com o intuito de analisar a adequação dos modelos aos dados, fez-se a análise dos resíduos destes modelos por forma a verificar se são de facto *white noise* (ou seja se são independentes, não correlacionados e normalmente distribuídos).

Primeiro, começou-se por testar a hipótese de serem independentes. Para tal foi usado o comando `Box.test` do tipo `Ljung`, que realiza o teste de hipóteses *Ljung-Box* em que a hipótese nula (H_0) é a independência da série temporal que se dá como *input*. Recorrendo à respectiva documentação, sabe-se que se deve ter $fitdf = p + q$ e que $lag > fitdf$. Além disso, é recomendado o uso de $lag = ln(n)$, sendo n o número de observações da série. Neste caso, sendo que $n = 9862$ e $lag = ln(9862) = 9.2$, considerou-se para todos os modelos que $lag = 10$. Aplicando, então, o referido teste à série de resíduos dos modelos anteriormente definidos obteve-se os valores-p que se encontram na Tabela 2 (idealmente gostar-se-ia que estes valores-p fossem altos para que se pudesse aceitar a hipótese de os resíduos serem independentes com segurança).

FUR	HOR	OLA	STO	GED	KUN	RAT
0.1853	0.2287	0.07441	0.01466	0.5049	0.1858	0.3989

Tabela 2: Valores-p do teste de independência de *Ljung-Box* para os resíduos dos modelos anteriormente ajustados.

Analisando os valores-p na Tabela 2, pode-se concluir que de facto a hipótese dos resíduos serem independentes (H_0) é aceite para a maioria dos modelos, considerando os níveis de significância usuais de 1% e de 5% (níveis normalmente considerados na análise de resíduos). Para as séries FUR, HOR, OLA, GED, KUN e RAT tem-se que H_0 é aceite aos níveis de significância referidos. Para a série STO tem-se que H_0 é rejeitada aos níveis de significância referidos. Ou seja, o modelo ajustado à série STO não se adequa tão bem como os anteriores aos dados.

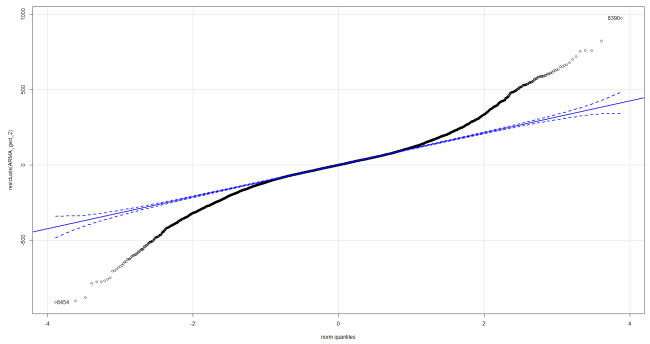


Figura 3: QQplot dos resíduos do modelo ajustado à série GED.

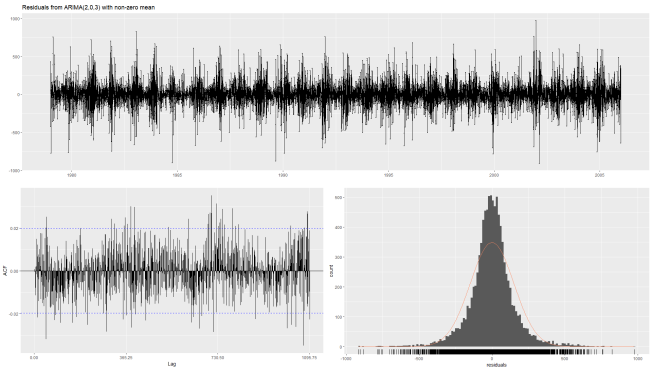


Figura 4: Gráfico dos resíduos do modelo ajustado à série GED ao longo do tempo (topo), da ACF destes (em baixo à esquerda) e da distribuição destes (em baixo à direita).

Na Figura 3, encontra-se o gráfico QQplot dos resíduos do modelo ajustado à série GED (série que apresentou maior valor-p aquando foi testada a hipótese de independência dos resíduos). Na Figura 4, encontra-se o resultado do comando `checkresiduals` aplicado à série de resíduos obtidos do modelo ajustado à série GED. Apresentam-se apenas estes gráficos, sendo que os das restantes séries são parecidos e a respetiva análise feita de modo semelhante. Analisando o QQplot dos resíduos (Figura 3) e a distribuição destes (Figura 4, em baixo à direita), pode-se concluir que os resíduos seguem uma distribuição aproximadamente normal (adequação destes à recta azul no primeiro e adequação à distribuição normal no segundo), o que por sua vez indica boa adequação do modelo aos dados. Note-se que o afastamento da distribuição destes nas extremidades em relação à recta azul evidencia o facto destes modelos possuírem *heavy-tails*. Para confirmar esta análise, efectuou-se também o teste de *Shapiro-Wilk* cuja hipótese nula é os dados de `input` serem normalmente distribuídos. Após alguma manipulação de dados, uma vez que este teste

(`shapiro.test`) está limitado a um `input` de dimensão máxima 5000, testou-se a hipótese dos resíduos serem normalmente distribuídos, sendo os valores-p obtidos razoavelmente altos, levando à aceitação da hipótese nula (de distribuição normal).

Analisando a ACF dos resíduos pretende-se avaliar a aleatoriedade destes que, por sua vez, caso esteja presente, evidencia que estes são de facto não correlacionados. Analisando o gráfico da ACF dos resíduos representado na Figura 4 em baixo à esquerda, note-se que a autocorrelação não apresenta valores significativamente diferentes de zero (todos contidos numa abertura de 0.04, sendo que a grande maioria se encontra contida numa abertura de 0.02), o que leva a concluir que de facto os resíduos são não correlacionados. Esta mesma análise (de autocorrelação e distribuição normal) é igual para as séries de resíduos ajustadas às outras séries (cujos gráficos não são aqui representados). O que indica que além da hipótese de independência dos resíduos (*Ljung-Box*) ser aceite para para a maioria das séries de resíduos dos modelos ajustados, a hipótese dos resíduos não serem correlacionados e de possuírem distribuição normal são também aceites. No seguimento de verificar se os resíduos são de facto não correlacionados e normalmente distribuídos, estuda-se a variabilidade da variância, ou seja, se existem subconjuntos dos resíduos dos modelos que apresentem variâncias diferentes. Isto implica então fazer um teste de homoscedasticidade aos resíduos, neste caso chamado teste de *ARCH Engle's*, cuja hipótese nula é os resíduos do modelo ARIMA serem homoscedásticos. Os valores-p obtidos foram todos altos o suficiente para que a hipótese nula fosse aceite aos níveis de significância usuais. Conclui-se deste modo que os resíduos apresentam homoscedasticidade (o que evidencia uma maior qualidade dos modelos seleccionados).

2.6 Previsão dos Quatro Períodos Seguintes

Para obter as previsões dos quatro períodos seguintes e respetivos intervalos de previsão a 95% de confiança ($IC_{95\%}$), recorremos ao comando `forecast.stl`, com $h = 4$, `level=95`, e `method="arima"`. Este comando por definição recebe a decomposição STL da série temporal e faz a previsão baseada no modelo obtido através do comando `auto.arima` aplicado a (X_t) (o que resta após a decomposição STL), que é por sua vez o modelo que decidimos adotar para cada uma das séries. Os resultados obtidos encontram-se seguidamente na Tabela 3.

Tempo	**FUR**	**HOR**	**OLA**	**STO**
01/01/2006	$38.39 \in [-195.99, 272.78]$	$-98.35 \in [-369.84, 173.13]$	$51.53 \in [-67.68, 170.75]$	$42.08 \in [-29.31, 113.47]$
02/01/2006	$-8.27 \in [-315.07, 298.52]$	$-31.33 \in [-346.41, 283.73]$	$60.99 \in [-108.20, 230.19]$	$60.97 \in [-58.09, 180.05]$
03/01/2006	$-19.61 \in [-349.89, 310.65]$	$-20.56 \in [-358.76, 317.63]$	$56.66 \in [-139.08, 252.40]$	$56.21 \in [-89.68, 202.10]$
04/01/2006	$-17.15 \in [-362.17, 327.87]$	$-8.53 \in [-358.40, 341.33]$	$52.38 \in [-163.85, 268.63]$	$44.71 \in [-125.01, 214.44]$

Tempo	**GED**	**KUN**	**RAT**
01/01/2006	$112.80 \in [-182.34, 407.95]$	$88.17 \in [-88.77, 265.12]$	$32.61 \in [-179.58, 244.80]$
02/01/2006	$84.82 \in [-254.58, 424.23]$	$97.98 \in [-124.26, 320.23]$	$-15.53 \in [-289.56, 258.48]$
03/01/2006	$86.20 \in [-267.44, 439.86]$	$97.08 \in [-144.64, 338.82]$	$-21.18 \in [-316.69, 274.33]$
04/01/2006	$87.46 \in [-273.40, 448.33]$	$93.54 \in [-162.53, 349.63]$	$-19.98 \in [-330.01, 290.04]$

Tabela 3: Previsão para os quatro períodos seguintes e os respetivos $IC_{95\%}$.

Note-se que quanto maior o valor de h (previsão a maior prazo), maior será a amplitude dos intervalos de confiança (maior a incerteza da previsão). Tal seria de esperar uma vez que a incerteza de uma previsão à partida aumenta para futuros mais distantes. Além disso, a função `forecast.stl` do R assume que os resíduos dos modelos são de facto não correlacionados e normalmente distribuídos e tal pode não ser sempre verdade para estes modelos, como é visto na subsecção anterior. Além disso, não é tido em conta a incerteza das estimações dos parâmetros usados nos modelos, aumentando ainda mais o nível de incerteza deste resultados.

3 Resultados - Valores de Fecho do Índice NASDAQ

3.1 Série Temporal de Log-returns

De seguida encontra-se na Figura 5 o gráfico da série temporal dos log-returns associados aos valores de fecho do índice NASDAQ, $X_t = log(P_t) - log(P_{t-1})$, onde P_t é o valor do fecho do índice NASDAQ.

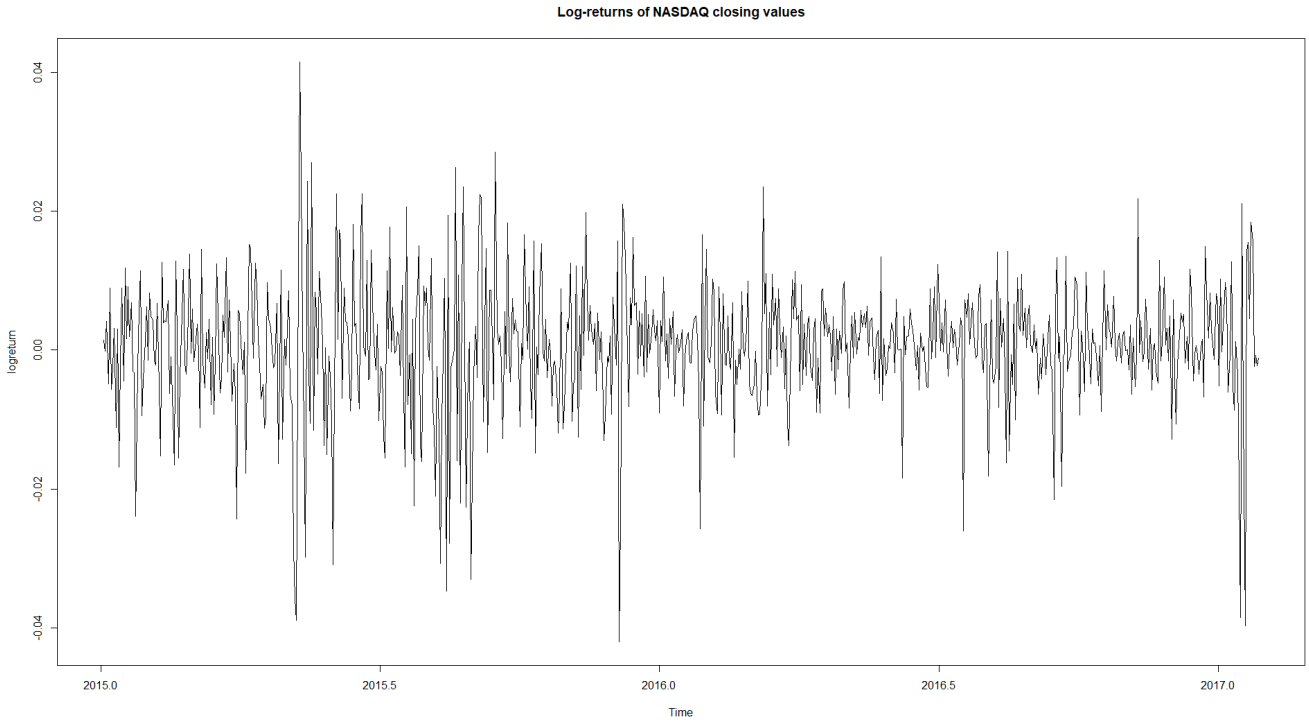


Figura 5: Log-returns dos valores de fecho do índice composto NASDAQ.

3.2 Análise Inicial dos Log-Returns

Inicialmente, tentou-se ajustar um modelo $ARMA(p, q)$ à série dos log-returns, por forma a eliminar alguma dependência linear nos dados. Contudo, após correr o comando `auto.arima`, obteve-se que o modelo que mais se adequava tinha ordens de dependência $p = q = 0$. Assim, concluiu-se que seria desnecessária a utilização de um modelo ARMA nestes dados.

Seguidamente, procurou-se, então, evidência de efeitos ARCH nos dados. Para tal, primeiramente optou-se por fazer um teste de independência (*Ljung-Box*) aos log-returns cujo resultante valor-p de 0.93 indica que a hipótese nula deve ser aceite. Deste modo, sendo que os dados aparentam não ter dependência serial, correu-se novamente o teste de *Ljung-Box* desta vez em relação aos resíduos quadrados padronizados, obtendo-se um valor-p de 2.2×10^{-16}. Assim, deve-se rejeitar a hipótese nula (a de independência), querendo dizer que de facto as autocorrelações dos resíduos quadrados dos log-returns são colectivamente significativamente grandes em magnitude (diferentes de zero). Por outras palavras, há indícios de efeito ARCH.

De seguida, avaliou-se algumas características que os modelos de variância condicionada apresentam. Por exemplo, a média dos log-returns da série é aproximadamente zero (média= 0.000495) e a variância é de ordem menor que 10^{-4} (variância= 8.721256×10^{-5}). Analisando agora a Figura 6, vemos que a ACF é insignificante em quase todos os lags com exepção apenas do primeiro, como era já de esperar para este tipo de dados. Por outro lado, a ACF dos valores absolutos ou do quadrado dos log-returns da série é significativamente diferente de zero para um grande número de lags e sempre positiva. Todas estas observações levaram à conclusão de que, neste caso, um modelo multiplicativo de facto se adequa. Confrontando com os resultados anteriores obtidos nos testes de *Ljung-Box*, confirma-se que de facto os log-returns são não correlacionados (primeiro teste) e que o quadrado dos log-returns são correlacionados (segundo teste), o que evidencia a existência de efeito ARCH, sendo por isso agora necessário modelar a equação que rege a volatilidade.

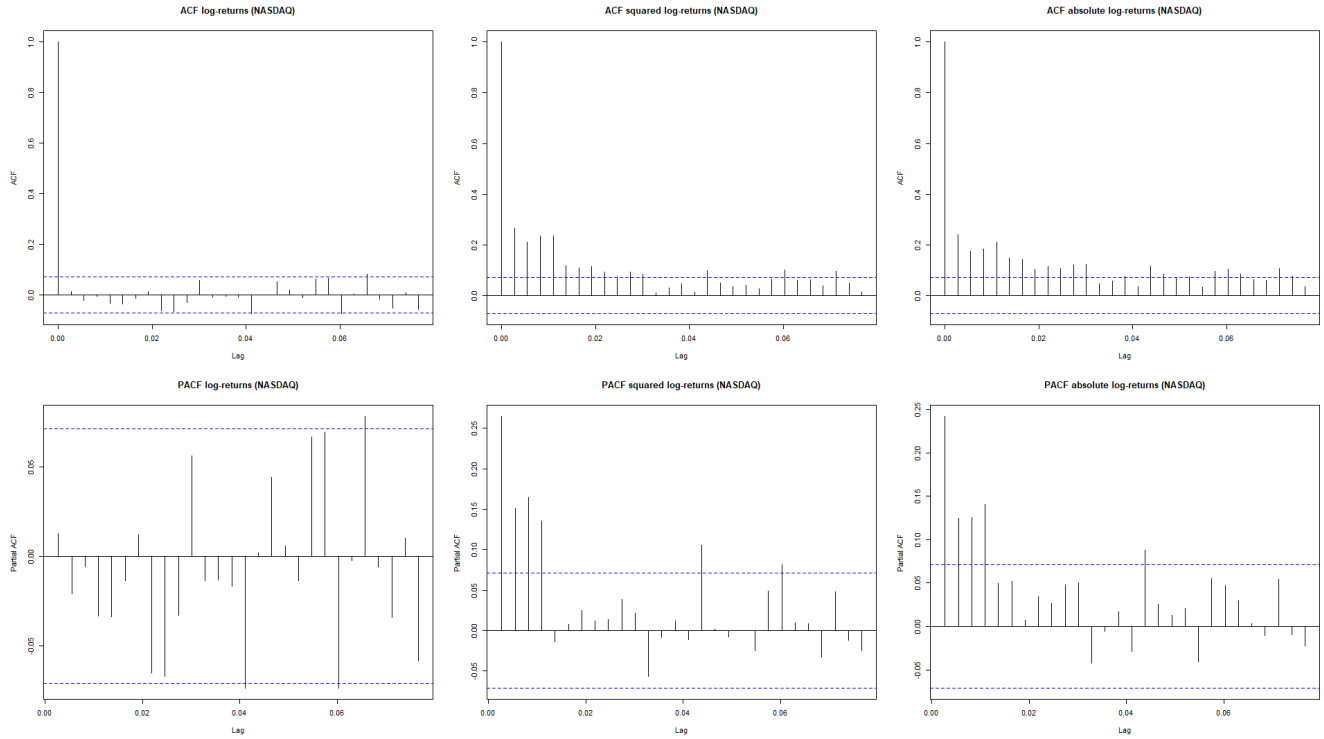


Figura 6: Gráficos das ACF (em cima) e das PACF (em baixo) dos log-returns (esquerda), do quadrado dos log-returns (centro) e do valor absoluto dos log-returns (direita).

3.3 Ajustamento de Modelos do Tipo GARCH à Série Temporal de Log-returns

De forma a encontrar o modelo GARCH que melhor se adequa aos log-returns, começou-se por fazer uma *"grid-search"* para várias ordens de dependência (variando parâmetros p e q), recorrendo ao comando `garchFit`. Os resultados obtidos para os valores AIC e BIC encontram-se na Tabela 4. Com base nestes valores (e outros omitidos neste relatório devido a serem superiores), escolheu-se adoptar o modelo que apresenta melhores medidas de selecçao (menores valores de AIC e BIC), nomeadamente o modelo GARCH(1,1).

	GARCH(1,1)	**GARCH(1,0)**	**GARCH(1,2)**	**GARCH(2,1)**	**GARCH(2,2)**
AIC	-6.685	-6.597	-6.689	-6.682	-6.687
BIC	-6.660	-6.579	-6.659	-6.651	-6.650

Tabela 4: Resultados das medidas AIC e BIC para vários valores de p e q no modelo GARCH(p,q).

3.4 Estimação Paramétrica

Usando a estimação paramétrica que o comando `garchFit` calculou para o modelo GARCH(1,1) ajustado, obtemos:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = 1.060 \times 10^{-5} + 0.1838 X_{t-1}^2 + 0.69307 \sigma_{t-1}^2 \end{cases}$$

onde $(Z_t) \sim \mathcal{N}(0,1)$. Note-se que o parâmetro μ foi omitido por ser considerado não significativo devido ao seu reduzido valor, $\mu = 8.111 \times 10^{-4}$ ($X_t = \mu + \sigma_t Z_t$). O valor de a_0 é também muito reduzido ($a_0 = 1.060 \times 10^{-5}$) mas é indicado no modelo uma vez que faz parte da definição deste ($\sigma_t^2 = a_0 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2$, a_0, a_1, $b_1 > 0$).

3.5 Diagnóstico dos Resíduos e Validação do Modelo Escolhido

Após a selecção do modelo GARCH que melhor se adequa aos log-returns, procede-se ao diagnóstico de resíduos (idealmente quer-se que estes sejam *white noise*) e consequente validação do modelo. Primeiro, usou-se um teste

de *Ljung-Box* para determinar se os resíduos são de facto independentes. O valor-p obtido de 0.15 indica-nos que de facto aos níveis de significância usuais a hipótese de independência dos resíduos é aceite. De seguida, analisando o QQplot presente na Figura 7 e a distribuição destes que se encontra na Figura 8 (em baixo à direita) pode-se concluir que os resíduos seguem uma distribuição aproximadamente normal (adequação destes à recta azul no primeiro e adequação à distribuição normal no segundo). Para confirmar este resultado, efectuou-se também o teste de *Shapiro-Wilk* cuja hipótese nula é a série dada como `input` ser normalmente distribuída. De novo foi necessária alguma manipulação de dados pelas razões referidas anteriormente. O valor-p obtido foi relativamente alto, indicando que a hipótese de distribuição normal deve ser aceite aos níveis de significância usuais. Na Figura 8 (em baixo á esquerda), encontra-se a ACF dos resíduos. Analisando este gráfico conclui-se que, estando a grande maioria das auto-correlações todas contidas numa abertura de 0.075, as auto-correlações não apresentam valores significativamente diferentes de zero, o que leva a concluir que de facto os resíduos aparentam ser não correlacionados. Note-se ainda, que, para este modelo ajustado, GARCH(1,1), o valor-p do teste *Multiplicador de Lagrange* (`LM Arch Test`) aplicado aos resíduos padronizados é muito alto (0.936), evidenciando o facto do modelo se adequar bem aos dados, uma vez que os resíduos apresentam, então, homoscedasticidade.

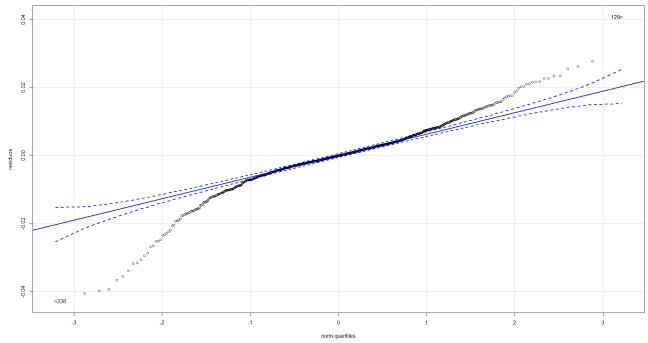


Figura 7: QQplot dos resíduos do modelo GARCH(1,1) ajustado.

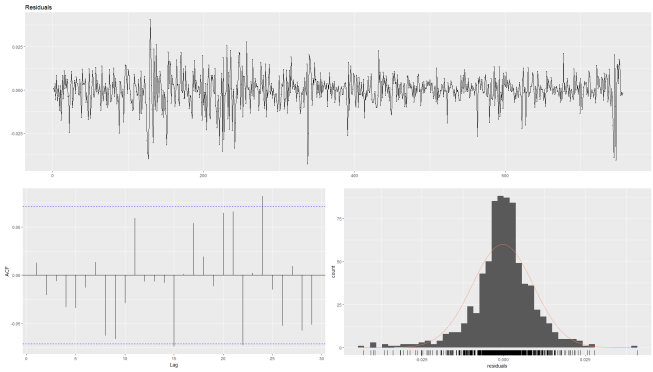


Figura 8: Gráfico dos resíduos do modelo GARCH(1,1) ajustado ao longo do tempo (topo), da ACF destes (em baixo à esquerda) e da distribuição destes (em baixo à direita).

4 Conclusões

Com a realização deste projecto, conseguiu-se, na primeira parte, ajustar modelos lineares a dados relativos ao nível do mar em cidades banhadas pelo Mar Báltico, enquanto que, na segunda parte, conseguiu-se ajustar um modelo não linear aos dados relativos ao índice de mercado de acções NASDAQ. Assim, os objectivos propostos para este projecto foram alcançados.

Como tal, em ambos os casos, começou por se realizar uma análise preliminar dos dados facultados com o intuito de identificar características que dessem indícios de que tipo de modelo se adequaria mais a cada um dos casos. De seguida, procedeu-se ao ajustamento dos respectivos modelos e à procura dos que se adequassem melhor, seguido por uma análise das séries de resíduos dos modelos ajustados e por conseguinte a validação desses mesmos modelos. Procedeu-se ainda, para o primeiro conjunto de dados, à previsão de valores futuros para os quatro períodos seguintes, calculando-se também os respectivos intervalos de confiança a 95%.

Os resultados obtidos, nomeadamente os modelos ajustados, foram relativamente satisfatórios como se verificou pela análise das séries de resíduos. Contudo, algumas suposições no ajustamento dos modelos não foram satisfeitas tais como a independência da série de resíduos para um dos casos, autocorrelações dos resíduos ligeiramente superior para os modelos GARCH que para os modelos ARMA, entre outras. Note-se também que, na primeira parte, originalmente pressupunha-se ajustar modelos SARIMA, mas que acabou por se revelar mais adequado aos dados ajustar simplesmente modelos ARMA.

Para concluir, deve-se referir que este projecto proporcionou uma oportunidade de aprendizagem sobre ambas as famílias de modelos e também adquirir alguma prática em relação ao procedimento a ter na análise de resíduos e na validação de modelos.

Referências

[1] Scotto, M. Notas de apoio à cadeira Séries Temporais. Lisboa: Instituto Superior Técnico.

[2] Data Science. 2018. *Introduction to Forecasting with ARIMA in R*. [ONLINE] Disponível em: `https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials`. [Acedido a 24 de Junho de 2018].

[3] r-statistics.co. 2018. *Time Series Analysis*. [ONLINE] Disponível em: `http://r-statistics.co/Time-Series-Analysis-With-R`. [Acedido a 24 de Junho de 2018].

[4] Statistical forecasting: notes on regression and time series analysis. 2018. *Identifying the numbers of AR or MA terms in an ARIMA model*. [ONLINE] Disponível em: `https://people.duke.edu/~rnau/411arim3.htm`. [Acedido a 24 de Junho de 2018].

[5] Hyndman, R. *Forecasting using R: Non-seasonal ARIMA models*. Disponível em: `https://robjhyndman.com/talks/RevolutionR/9-Nonseasonal-ARIMA.pdf`.

[6] Hyndman, R. *Forecasting using R*. Disponível em: `https://s3.amazonaws.com/assets.datacamp.com/production/course_3002/slides/ch5.pdf`.

[7] Lindberg, J, 2016. *Applying a GARCH Model to an Index and a Stock*. Bachelor Thesis in Mathematical Statistics. Estocolmo: Stockholm University. Disponível em: `https://kurser.math.su.se/pluginfile.php/20130/mod_folder/content/0/Kandidat/2016/2016_04_report.pdf`.