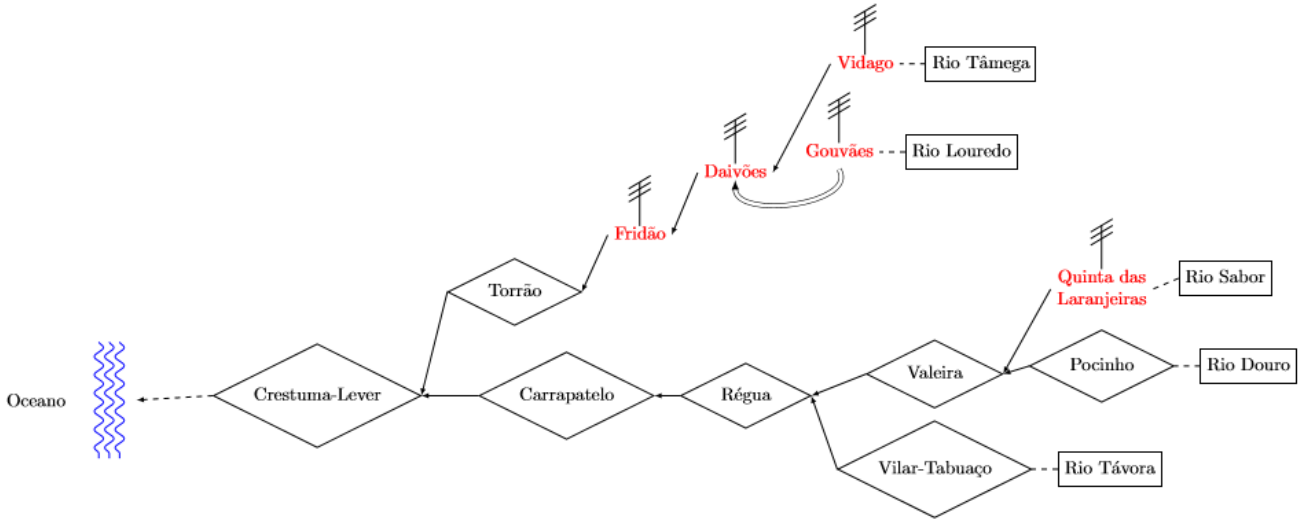


# Formulação de séries de afluências sintéticas aos aproveitamentos hidroelétricos existentes e futuros

André Gustavo Pereira Moutinho
Mestrado em Engenharia Matemática
Departamento de Matemática
2016

**Orientadora**
Maria do Carmo Guedes, Professora Auxiliar, FCUP

**Coorientadora**
Rute Almeida, Investigadora Auxiliar, FCUP

**Supervisor Local**
Sónia Vilela

– Página em branco –

# Resumo

A previsão do mercado nacional de energia é uma tarefa com custo computacional elevado que envolve simulações que têm em conta os diversos elementos produtores, o consumo nacional e o custo associado à energia produzida. Uma das fontes de produção de energia elétrica mais utilizadas em Portugal são os aproveitamentos hidroelétricos abastecidos por rios nacionais e internacionais.

Por vezes existem falhas de informação, que é necessário colmatar para uma correta análise do sistema elétrico. Neste trabalho pretende-se corrigir uma dessas falhas, que se estende ao longo dos anos de 2011 até 2015, e corresponde a locais onde está planeada a construção de novos aproveitamentos hidroelétricos. Os dados utilizados são séries temporais mensais desde Janeiro de 1956 até Dezembro de 2010, obtidas pela medição nos locais de afluências hidrológicas, em vários pontos de interesse.

As séries foram analisadas com modelos lineares de estrutura autorregressiva (ARIMA), em conjunção com variáveis exógenas. Os parâmetros do modelo foram estimados linearmente com os modelos ARIMA, e não linearmente com redes neuronais. Algoritmos adequados, em ambiente `R`, foram utilizados para a identificação dos modelos e estimação dos parâmetros, em modelos de várias ordens, e utilizou-se minimização do AIC, nos modelos lineares, e validação cruzada, nos modelos não lineares, para prevenir *overfitting* dos modelos. No fim, compararam-se os modelos e previram-se os cinco anos em falta, com aquele que apresentava a medida de erro mais baixa, para o conjunto de teste.

**Palavras-chave:** modelos ARIMA, redes neuronais, hidrologia, afluência, séries temporais, modelação, previsão.

– Página em branco –

# Abstract

Predicting the national energy market is a computational expensive task involving simulations which take into account production sources, national consumption and costs associated with produced energy. One of the most common sources for producing electric energy in Portugal are hydroelectric power dams powered by national and international rivers.

Sometimes there are lapses of information, which need to be addressed for a correct analysis of the electric system. This work aims to fix one of those lapses from the years 2012 through 2015, corresponding to locations where is planned to build new hydroelectric power dams. The data used are time series from January, 1956 to December, 2010 made of measurements on hydrologic inflows locations, in several places of interest.

The series were analyzed with autorregressive structure linear models (ARIMA), along with exogenous variables. The model's parameters were linearly estimated with ARIMA models, and non linearly estimated with neural networks. Appropriate algorithms, in `R` environment, were used for model identification and parameter estimation, in models of several orders, along with AIC minimization, for linear models, and cross validation, for non linear models, to prevent overfitting of the models. Lastly the models were compared and the missing five years were predicted with the model with the lowest error metric for the test set.

**Keywords:** ARIMA models, neural networks, hydrology, inflow, time series, modelling, prediction.

– Página em branco –

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

– Página em branco –

# 1 Introdução

O Sistema Hidroelétrico Nacional é constituído por diversos subsistemas de rios e afluentes interligados ou geograficamente próximos. Para prever as necessidades da rede elétrica nacional e aumentar a capacidade de satisfazer o consumo é importante analisar a produção de cada subsistema, e prever o impacto da sua expansão. Além dos aproveitamentos hidroelétricos (AH) existentes, há AH planeados, em locais dos rios equipados com postos de medição, que foram utilizados para determinar a sua relevância energética.

No capítulo 1 pode ler-se a introdução ao problema e uma breve descrição dos sistemas analisados adiante, utilizados na produção hidroelétrica em Portugal. No capítulo 2 são descritos os modelos utilizados: ARIMA e redes neuronais; tal como ferramentas importantes ao tratamento dos dados e escolha dos modelos: ACF e PACF, periodograma, Q-Q plot, AIC, Validação Cruzada, MAE, MAPE e MASE e transformações logarítmicas. No capítulo 3 é descrito o processo de análise, modelação e previsão dos dados, bem como a distinção entre os dois sistemas estudados, e os problemas que foi necessário ultrapassar para prever as séries. No capítulo 4 são mencionadas sugestões para melhorar os resultados obtidos e aplicar modelos alternativos.

## 1.1 Enquadramento do Problema

Foram estudadas séries de afluências de dois subsistemas a Norte do país. As séries de afluências são medições da quantidade de água, por unidade de tempo, num determinado local e quando nos referimos a afluências diárias falamos do volume de água que chega a cada AH por dia. Ao longo do trabalho utilizaram-se séries de afluências diárias, semanais e mensais, que são as medições da quantidade de água por dia, semana ou mês de observação, respetivamente.

Para entender as interligações entre os AH e o comportamento das séries de afluências, analisamos o Sistema Lima-Cávado, constituído por oito AH, dois localizados no

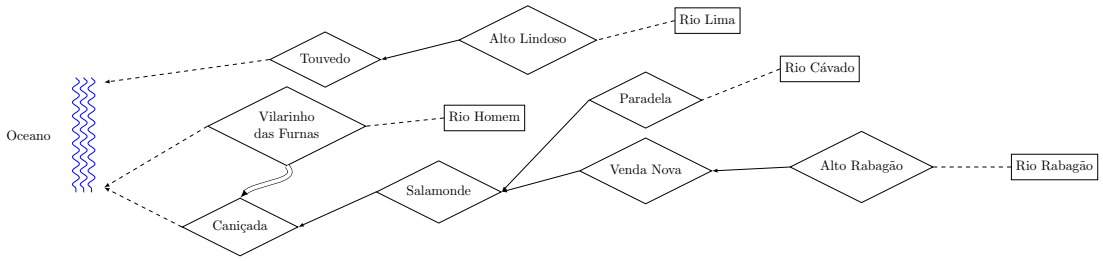


Fig. 1 – Esquema dos AH do Sistema Lima-Cávado

Lima e os restantes no Cávado. Em seguida, analisamos o Sistema Douro, onde estão planeados cinco AH futuros, os quais se pretende modelar, e prever, a partir da informação disponibilizada pelos seus postos de medição.

## 1.2 Sistema Lima-Cávado

O Sistema Lima-Cávado, esquematizado na figura 1, é composto por oito AH do tipo albufeira, ou seja, com reservatório de água para controlo das descargas. Os AH estão assinalados com losangos, e ligados entre si por linhas cheias, ou duas linhas quando se trata de uma ligação artificial. Além disso, temos os rios, em que estão instalados, assinalados com retângulos e ligados ao sistema por linhas a tracejado que representam o curso a montante dos AH. Estes também estão ligados por linhas a tracejado ao restante curso dos rios, que desagua no oceano.

Os rios Lima e Cávado não têm ligações físicas entre si, contudo devido à sua proximidade, partilham condições meteorológicas e geológicas com influência sobre as suas afluências. O rio Cávado tem AH instalados em dois afluentes, Homem e Rabagão, em que o Rabagão desagua a jusante de Venda Nova entre Paradela (montante) e Salamonde (jusante); e o Homem desagua a jusante dos AH do Cávado, mas Vilarinho das Furnas tem uma ligação artificial a Caniçada para efetuar bombagens.

Nas cabeceiras, e colocados a montante de todos os outros, estão os AH de Alto Lindoso, Vilarinho das Furnas, Paradela e Alto Rabagão, portanto estes AH não são influenciados pelos restantes.

## 1.3 Sistema Douro

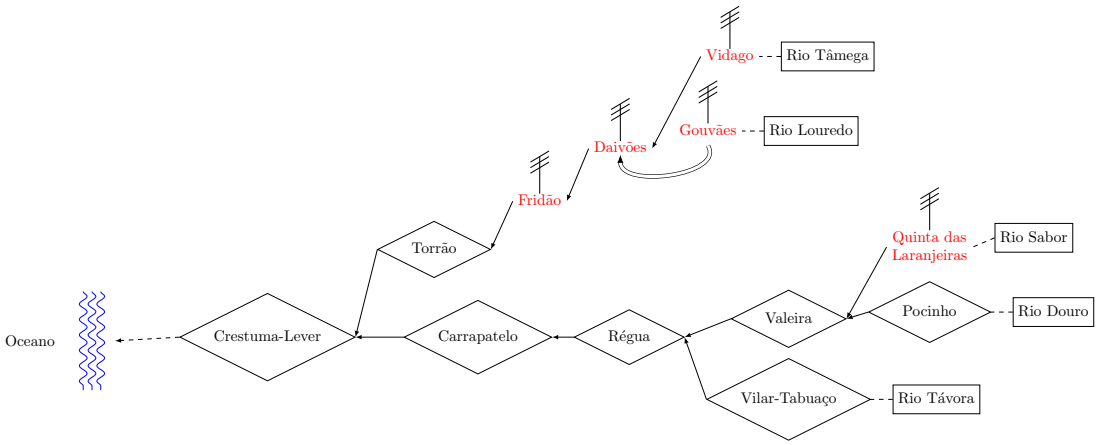


Fig. 2 – Esquema dos AH e postos de medição do Sistema Douro

O sistema Douro é constituído por vários AH sendo um dos rios com mais elevada produção hidroelétrica a nível nacional. Por esse motivo está planeado adicionar novos aproveitamentos, assinalados com $\overset{\Upsilon}{|}$ na figura 2, nos rios Tâmega e Sabor. Apenas estão esquematizados os AH a jusante de Pocinho, inclusive, porque os restantes não têm relevância para os postos de medição estudados. Além dos postos de medição, temos os AH existentes assinalados com losangos, e as ligações entre postos e AH assinaladas com linhas cheias, ou duas linhas quando se trata de uma ligação artificial. Também temos os rios em que estão instalados, assinalados por retângulos, e ligados por linhas a tracejado aos postos de medição ou AH. As linhas a tracejado assinalam ainda a ligação entre o AH mais a jusante, Crestuma-Lever, e o Oceano.

O rio Sabor, mais a montante do Douro, tem um posto de medição em Quinta das Laranjeiras, o local previsto para instalação de um futuro AH. O rio Tâmega tem quatro novos AH previstos, em cascata, com Vidago e Gouvães como cabeceiras da cascata nos rios Tâmega e Louredo, respetivamente, seguidos pelo Daivões a jusante e a terminar no Fridão a montante do AH de Torrão. É de salientar que, apesar de não ter afluência natural ao Daivões, o Gouvães está representado por uma ligação dupla devido ao canal

artificial que ligará os AH.

### 1.3.1 Dados de precipitação

Para prever corretamente as afluências do sistema Douro foram consideradas variáveis exógenas, mais concretamente dados de precipitação em torno dos locais estudados. Foram consultadas duas fontes, o SNIRH (Sistema Nacional de Informação de Recursos Hídricos) e a NASA (National Aeronautics and Space Administration). Devido ao número elevado de valores em falta nas medições do SNIRH foram somente utilizados os dados da NASA retirados do projeto MERRA (Modern Era Retrospective-Analysis for Research and Applications) acedidos em Abril de 2016 [1].

[1] `http://disc.sci.gsfc.nasa.gov/daac-bin/FTPSubset.pl`

# 2 Tópicos da teoria utilizada na modelação

Neste capítulo são introduzidos os modelos aplicados aos dados e respetivas ferramentas necessárias à escolha do modelo apropriado ao conjunto de dados.

## 2.1 Modelos de previsão

De acordo com Murteira et al. (1993), dadas $N$ variáveis aleatórias $X_1, X_2, ..., X_N$ com observações nos instantes temporais $t_1, t_2, ..., t_N$ representamos a realização $x(t_1)$, $x(t_2), ..., x(t_N)$ como $x(t),\ t = t_1, t_2, ..., t_N$. A este conjunto $x(t)$ de observações aleatórias indexadas no tempo, chamamos série temporal.

A partir daqui, para simplificar a representação de uma série temporal, vamos utilizar o operador atraso $B$, que se define $Bx(t) = x(t-1)$, ou seja, $B^j x(t) = x(t-j),\ j \in \mathbb{N}$. Assim, podemos representar uma série $x(1), x(2), ..., x(t)$ como $B^{t-1}x(t)$, $B^{t-2}x(t)$, ..., $B^0 x(t)$.

A forma mais simples de uma série temporal é o ruído branco $\epsilon_t,\ t = 1, 2, ..., N$. Quando a série tem média constante $E\{\epsilon_t\} = \mu_\epsilon$, variância constante $Var\{\epsilon_t\} = \sigma_\epsilon^2$ e covariância $\gamma_k = \sigma_\epsilon^2 \delta_k$, em que

$$\delta_k = \begin{cases} 1,\ k = 0 \\ 0,\ k \neq 0 \end{cases}$$

é a função $\delta$ de Dirac, estamos perante ruído branco.

Assim, quando temos uma série de observações $x(t) = \epsilon(t)$ estamos perante o caso mais simples em que as observações são puramente aleatórias. Contudo, isto raramente acontece na prática, e por isso podemos utilizar modelos como os apresentados a seguir, para explicar as séries em que as observações estão correlacionadas com o passado, das próprias observações ou dos erros.

**Estacionariedade e Invertibilidade**

- Uma série é estacionária se as suas propriedades estatísticas forem invariantes no tempo. Contudo, esta definição de estacionária raramente ocorre na prática, por isso, é utilizada uma definição mais abrangente de estacionariedade de ordem $k$, que apenas restringe as propriedades estatísticas de ordem $k$. Em geral, é utilizada a estacionariedade de ordem 2, que exige que a média, variância e função de autocorrelação da série sejam invariantes no tempo.

- A condição de invertibilidade de um modelo é importante para que não haja uma perda de generalidade. Cada série pode ser explicada por vários modelos que a explicam com características distintas. Isto quer dizer que se o modelo não for invertível, não é possível identificar apenas um modelo adequado à série. Além disso, se o modelo não for invertível, isso pode afetar o cálculo dos resíduos (Hipel et al., 1994).

### 2.1.1 Família de modelos Box-Jenkins

Quando cada observação da série depende de mais do que uma observação aleatória, usamos um modelo de média móvel.

- MA$(q)$ - Média Móvel
  Modela uma série $x_t$ a partir de $q+1$ erros

$$x_t = \epsilon_t - \theta_1\epsilon_{t-1} - ... - \theta_p\epsilon_{t-q} = \epsilon_t - \sum_{j=1}^{q}\theta_j\epsilon_{t-j} \tag{2.1.1}$$

  Seja $\Theta_q(B) = 1 - \theta_1 B - ... - \theta_q B^q$ o polinómio de grau $q$ no operador atraso $B$ associado ao modelo $x_t = \Theta_q(B)\epsilon_t$. Para que a série seja invertível é necessário que as raízes do polinómio se encontrem fora do círculo unitário.

Existem outros casos em que cada observação da série está dependente das suas próprias observações anteriores, e nesses casos aplica-se um modelo autorregressivo.

- AR($p$) - Autorregressivo

  Modela uma série $x_t$ a partir das suas $p$ observações anteriores

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + \epsilon_t = \epsilon_t + \sum_{i=1}^{p} \phi_i x_{t-i} \tag{2.1.2}$$

  Seja $\Phi_p(B) = 1 - \phi_1 B - ... - \phi_p B^p$ o polinómio associado ao modelo $\Phi_p(B)x_t = \epsilon_t$. Para que a série seja estacionária é necessário que as raízes do polinómio se encontrem fora do círculo unitário.

Se cada observação da série depender de observações aleatórias, e observações anteriores, podemos conjugar os dois modelos e utilizar um modelo autorregressivo de média móvel.

- ARMA$(p, q)$ - Autorregressivo de Média Móvel

  Modela uma série $x_t$ a partir das suas $p$ observações anteriores e de $q + 1$ erros

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{2.1.3}$$

  Este modelo é uma composição dos dois mencionados anteriormente, equações (2.1.1) e (2.1.2), e está restrito a ambas as condições de invertibilidade e estacionariedade dos anteriores.

- I$(d)$ - Integrado

  Quando temos uma série não estacionária, como por exemplo, quando a podemos explicar com o modelo autorregressivo $x_t = x_{t-1} + \epsilon_t$, em que $\phi_1 = 1$ e não cumpre a restrição de estacionariedade, não podemos modelar com modelos autorregressivos ou de média móvel. Contudo podemos resolver o problema mediante aplicação de diferenciações, ou seja, do operador atraso $B$ e temos o modelo $(1 - B)^d x_t = \epsilon_t$, em que $d$ é o número de raízes unitárias presentes nos dados.

Novamente, é possível combinar o modelo integrado com os anteriores, para obter um único modelo para modelar as séries, que tenham as características de cada um dos anteriores. Assim, obtemos o modelo autorregressivo integrado de média móvel.

- ARIMA$(p,d,q)$ - Autorregressivo Integrado de Média Móvel

  Modela uma série não estacionária $x_t$, que após diferenciada, depende de um modelo ARMA$(p,q)$

$$(1-B)^d x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{2.1.4}$$

  Este modelo tem as mesmas restrições que o modelo ARMA, dado pela equação (2.1.3)

Analisar as observações consecutivas anteriores pode requerer modelos de elevada ordem para captar a autocorrelação das séries. Por isso, quando temos séries periódicas (ou sazonais), por exemplo, conseguimos obter mais informação se analisarmos as observações, intervaladas por um período $s$, ou seja, $t, t-s, t-2s, ....$ Para isso, utilizamos um modelo sazonal, obtido a partir de qualquer um dos anteriores. Vamos ilustrar a aplicação com o modelo ARIMA.

- S$(P,D,Q)_s$ - Sazonal

  É um modelo ARIMA$(P,D,Q)$ em que a série depende de $P$ observações anteriores periodicamente espaçadas, $Q+1$ erros anteriores periodicamente espaçados e tem $D$ aplicações do operador atraso sazonal $B^s$. Para um modelo sazonal com período $s$ temos

$$\begin{aligned}(1-B^s)^D x_t = \;& \phi_s x_{t-s} + \phi_{2s} x_{t-2s} + ... + \phi_{Ps} x_{t-Ps} + \\ & + \epsilon_t - \theta_s \epsilon_{t-s} - \theta_{2s} \epsilon_{t-2s} - ... - \theta_{Qs} \epsilon_{t-Qs} \\ = \;& \sum_{k=1}^{P} \phi_{ks} x_{t-ks} + \epsilon_t - \sum_{l=1}^{Q} \theta_{ls} \epsilon_{t-ls}\end{aligned} \tag{2.1.5}$$

  Ao considerarmos os polinómios sazonais $\Phi_P^s(B) = 1 - \phi_s B^s - ... - \phi_{Ps} B^{Ps}$ e $\Theta_Q^s(B) = 1 - \theta_s B^s - ... - \theta_{Qs} B^{Qs}$ podemos escrever de forma sucinta o modelo sazonal $(1-B^s)^D \Phi_P^s x_t = \Theta_Q^s \epsilon_t$.

O modelo SARIMA$(p,d,q) - (P,D,Q)_s$ tem a seguinte forma polinomial

$$(1-B)^d (1-B^s)^D \Phi_p \Phi_P^s x_t = \Theta_q \Theta_Q^s \epsilon_t \tag{2.1.6}$$

obtido da composição dos modelos atrás descritos.

Quando uma série não depende apenas das suas observações e erros, como no modelo ARIMA da equação (2.1.4), se dispusermos de séries independentes que contribuem para explicar o comportamento da série $x_t$, podemos utilizar uma regressão linear para juntar essa informação à nossa série, num modelo com variáveis exógenas.

- X$(r)$ - eXógeno

  Regressão de variáveis $\xi_t$ não presentes na série temporal $x_t$ que contribuem para explicar a aleatoriedade da série não explicada pelas suas observações anteriores.

  $$y_t = \sum_{i=0}^{r-1} \beta_i \xi_{t-i} + x_t \tag{2.1.7}$$

  Neste caso os resíduos correlacionados da regressão são $x_t$, a série resultante dos modelos anteriores, que podemos exemplificar a partir do modelo SARIMA, em que se obtém a série com a equação (2.1.6)

  $$x_t = \frac{\Theta_q \Theta_Q^s}{(1-B)^d (1-B^s)^D \Phi_p \Phi_P^s} \epsilon_t \tag{2.1.8}$$

  E as $r$ observações da variável exógena, $\xi_t, \xi_{t-1}, ..., \xi_{t-r+1}$, são utilizadas na regressão de $y_t$, e a otimização dos parâmetros é feita com mínimos quadrados ponderados (*weighted least squares*) (Shumway et al., 2010).

Se conjugarmos os modelos todos obtemos os modelos SARIMAX.

- SARIMAX$(p,d,q) - (P,D,Q)_s - (r)$

  Podemos escrever sucintamente, a partir das equações (2.1.7) e (2.1.8)

  $$y_t = \sum_{i=0}^{r-1} \beta_i \xi_{t-i} + \frac{\Theta_q \Theta_Q^s}{(1-B)^d (1-B^s)^D \Phi_p \Phi_P^s} \epsilon_t \tag{2.1.9}$$

### 2.1.2 Redes Neuronais

Inspiradas no sistema nervoso, as redes neuronais evoluíram ao longo dos últimos anos, e distanciaram-se do conceito original, ao reter as ferramentas e conceitos matemáticos introduzidos. Nas redes neuronais, cada neurónio é um número, e cada

ligação entre os neurónios é um peso. Um neurónio tem várias ligações, as de entrada que vêm de todos os neurónios de uma camada anterior, e as de saída que vão para todos os neurónios da camada seguinte. Isto pode ser visto na figura 3, em que cada neurónio está assinalado com um círculo e cada ligação por uma linha a cheio.

Cada neurónio soma os seus *inputs* e acrescenta-lhes a constante de viés, *bias*, para constituir o *input* $v_{i,k-1}$ e originar o *output* $v_{j,k}$. Isto quer dizer que numa camada $k$ o neurónio $j$ é dado por

$$v_{j,k} = f_k\left(b_{k-1} + \sum_{i=1}^{I} w_{ij} v_{i,k-1}\right) \tag{2.1.10}$$

Em que $b_{k-1}$ é o viés da camada anterior, $v_{i,k-1},\ i = 1, ..., I$ os neurónios da camada anterior, $w_{ij}$ o peso da ligação que vai de $v_{i,k-1}$ até $v_{j,k}$ e $f_k$ a função de ativação do neurónio $v_{j,k}$.

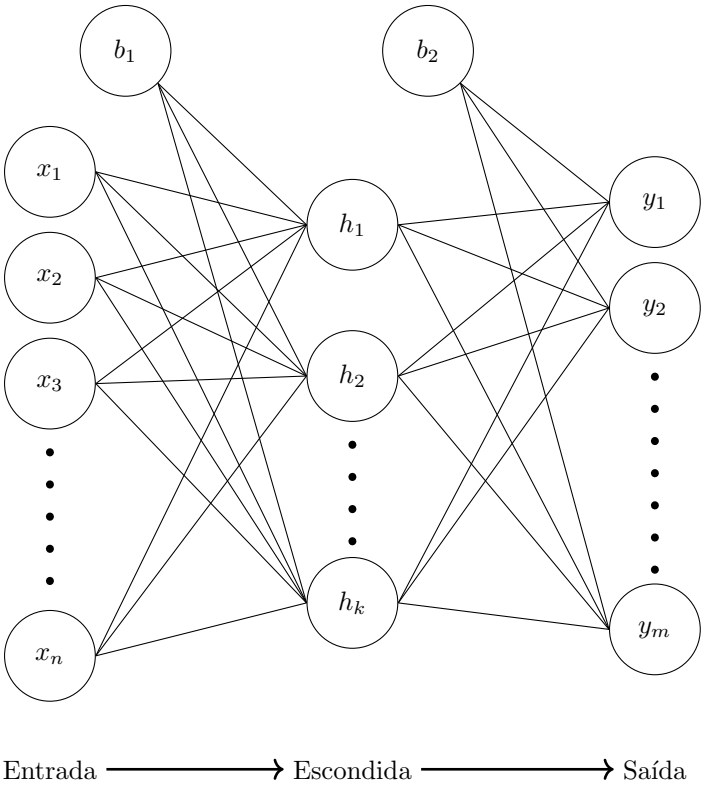


Fig. 3 – Estrutura de uma rede neuronal

Os neurónios de *input*, na camada inicial, estão presentes apenas para distribuir os *inputs* para a camada seguinte, portanto têm $f_i \equiv 1$, enquanto $f_j$ e $f_l$ representam as funções de ativação dos neurónios das camadas escondida e de saída, respetivamente. Assim, podemos exprimir a rede com a seguinte função

$$y_l = f_l\left(b_2 + \sum_{j\to l} w_{jl}\, \underbrace{f_j\left(b_1 + \sum_{i\to j} w_{ij}x_i\right)}_{h_j}\right) \tag{2.1.11}$$

Em que $y_l$ são os neurónios de saída, $x_i$ são os neurónios de entrada, e os neurónios da camada escondida são substituídos pela sua função de ativação, ou seja, são $h_j = f_j\left(b_1 + \sum_{i\to j} w_{ij}x_i\right)$. Além disso, temos os pesos $w$ correspondentes à ligação entre os neurónios de uma camada para a seguinte, e os neurónios de viés das camadas de entrada e escondida, $b_1$ e $b_2$ respetivamente, que têm um comportamento semelhante ao da constante de uma regressão linear.
As funções $f$ são habitualmente lineares, logísticas ( $f(x) = \frac{e^x}{1+e^x}$ ) ou função degrau de Heaviside ($f(x) = I(x > 0)$). Também se pode utilizar como variação a tangente hiperbólica ($f(x) = \tanh x = \frac{e^x-1}{e^x+1}$). (Ripley et al., 1995)

**Modelos NNSARX**$(p, P, r, k)_s$

Se construirmos uma rede neuronal com um único *output*, $x_t$, e utilizarmos como *inputs* as observações anteriores, tal como nos modelos autorregressivos, obtemos uma regressão com estrutura semelhante a um modelo SAR (Sazonal AutoRregressivo) e pesos estimados por funções não lineares (figura 4). Nestes modelos as séries neuronais têm $p + P + r$ *inputs* obtidos das $p$ observações anteriores, das $P$ observações anteriores com sazonalidade $s$ e das $r$ variáveis exógenas dadas ao modelo para explicar o *output* $x_t$. Além disso, podemos variar o tamanho $k$ da camada escondida, em geral para valores entre $1$ e $p + P + r$. Uma vantagem destes modelos é que sem camada escondida, como o modelo NNSARX$(p, P, 0, 0)_s$, temos o equivalente a um modelo ARIMA$(p, 0, 0) - (P, 0, 0)_s$ sem as restrições de estacionariedade (Hyndman e Athanasopoulos, 2014).

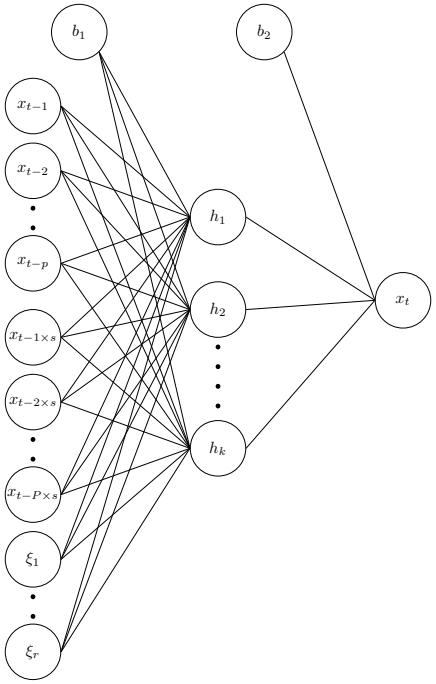


Fig. 4 – Estrutura de uma rede neuronal equivalente a modelos autorregressivos

## 2.2 Ferramentas de avaliação e diagnóstico

Para garantir uma escolha de modelo adequada aos dados e bons resultados na previsão foram utilizadas diversas ferramentas.

### 2.2.1 ACF e PACF

A função de autocorrelação amostral (*AutoCorrelation Function* - ACF) mostra a correlação entre uma série $x_t$ e os seus atrasos $x_{t-1}, x_{t-2}, \ldots, x_{t-m}$, enquanto a função de autocorrelação parcial amostral (*Partial AutoCorrelation Function* - PACF) mostra a correlação entre uma série $x_t$ e os seus atrasos $x_{t-k},\ k = 1, \ldots, m$, sem a influência dos atrasos $x_{t-1}, x_{t-2}, \ldots, x_{t-k+1}$.

A ACF amostral $\widehat{\rho}_m$ obtém-se a partir da autocovariância amostral $\widehat{\gamma}_m$

$$\widehat{\gamma}_m = \frac{1}{n}\sum_{t=1}^{n-m}(x_{t+m} - \bar{x})(x_t - \bar{x}) \tag{2.2.1}$$

em que $\widehat{\gamma}_m = \widehat{\gamma}_{-m}$, para $m = 0, 1, ..., n-1$ e

$$\widehat{\rho}_m = \frac{\widehat{\gamma}_m}{\widehat{\gamma}_0} \tag{2.2.2}$$

é a função de autocorrelação amostral da série $x_t$ (Shumway et al., 2010).
A PACF amostral $\widehat{\phi}_k$ é calculada através de sucessivos modelos autorregressivos AR$(p)$, com $p$ a variar de 1 até $k$

$$\tilde{x}_t = \phi_{p1}\tilde{x}_{t-1} + ... + \phi_{pp}\tilde{x}_{t-p} + \epsilon_t \tag{2.2.3}$$

em que $\tilde{x}$ é a série subtraída da média amostral, e o parâmetro da regressão $\phi_{pp}$ é a estimativa da função de autocorrelação parcial amostral, $\widehat{\phi}_p$, para o atraso $p$.
No gráfico das funções ACF e PACF costuma-se também representar a banda de confiança para a autocorrelação a uma significância de 5%. Se o valor da autocorrelação se encontrar dentro do intervalo $\pm 2/\sqrt{n}$, assumimos que esta não é significativa (Shumway et al., 2010).

### 2.2.2 Periodograma cumulativo

Além da análise temporal, podemos determinar as frequências relevantes para uma série através da sua análise espectral. No caso de dados com período $P$, a magnitude da frequência $\frac{1}{P}$ será maior do que as restantes magnitudes.
Assim, podemos definir o periodograma $I(\omega_j)$ a partir da transformada de Fourier dos dados $d(\omega_j)$

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t} \tag{2.2.4}$$

em que $x_t$ é a série dos dados e $\omega_j = \frac{j}{n},\ j = 0, \ldots, n-1$ são as frequências fundamentais (Shumway et al., 2010). O periodograma define-se a partir da transformada de Fourier dos dados

$$I(\omega_j) = |d(\omega_j)|^2 \tag{2.2.5}$$

Finalmente temos o periodograma cumulativo $I_c(\omega_j)$ a partir da soma cumulativa do periodograma

$$I_c(\omega_j) = \frac{\sum_{k=0}^{j} I(\omega_k)}{\sum_{k=0}^{\lfloor n/2 \rfloor} I(\omega_k)} \tag{2.2.6}$$

Assim, o periodograma cumulativo é uma curva crescente no intervalo $[0, 1]$ que varia em frequência no intervalo $[0, 0.5]$. Além disso, podemos utilizar bandas de confiança em torno da recta $y = 2x$ e se a série representada for ruído branco estará contida nessa banda. Por outro lado, se a série for periódica apresentará um degrau acentuado na frequência associada ao seu período (Salas, 1980).

### 2.2.3 Q-Q plot e P-P plot

O Q-Q plot (gráfico de quantis-quantis) representa os quantis empíricos da distribuição dos dados contra os quantis teóricos da distribuição com a qual queremos comparar, enquanto o P-P plot (gráfico de percentis-percentis) representa os percentis empíricos da distribuição dos dados contra os percentis teóricos da distribuição a comparar. Em geral são utilizados os quantis, ou percentis, teóricos de uma distribuição normal com média $0$ e variância $1$, $N(0, 1)$. Se as distribuições teórica e empírica forem semelhantes os pontos das observações estão alinhados com a recta $y = x$.

### 2.2.4 Previsão

Para prevermos uma série $x_t$ assumimos que o comportamento da série pode ser explicado pelo modelo que melhor se ajusta aos dados. Assim, se tivermos um modelo M, tal $x_t = M_{t-1,t-2,\ldots}(x) + \epsilon_t$, em que este modelo tem em conta as observações e/ou erros anteriores a $x_t$ e assumirmos que $\epsilon_{t+1} = E\{\epsilon_t\} = 0$, podemos calcular a previsão $\hat{x}_{t+1}$ como $\hat{x}_{t+1} = M_{t,t-1,t-2,\ldots}(x)$.

Para alargar o horizonte de previsão e prever $N$ valores da série, podemos assumir que $\hat{x}_{t+k},\ k = 1, \ldots, N-1$ é o valor da série no instante $t + k$ e calcular recursiva-

mente $\hat{x}_{t+k+1} = M_{t+k,t+k-1,\ldots,t-1,t-2,\ldots}(\hat{x}, x)$ a partir da série $x_t$ e das suas previsões $\hat{x}_{t+1}, \ldots, \hat{x}_{t+k}$.

### 2.2.5 Métodos de seleção de modelos

Para modelar os dados com os modelos apresentados anteriormente, foram utilizados três algoritmos presentes no *software* `R`: `arima` para os modelos ARIMA e `nnet` e `nnetar` para as redes neuronais. Cada um destes algoritmos utiliza métodos diferentes para selecionar o modelo mais adequado que são apresentados a seguir

- AIC (Critério de Informação de Akaike)

  O critério de informação definido por Akaike para relacionar a verosimilhança do modelo com o número de parâmetros é calculado da seguinte forma

  $$AIC = \log \widehat{\sigma}_k^2 + \frac{n + 2k}{n} \tag{2.2.7}$$

  em que $\widehat{\sigma}_k^2$ é a variância dos resíduos do modelo estimado por máxima verosimilhança, $k$ é o número de parâmetros do modelo e $n$ é o número de observações do conjunto de dados.

  O AIC foi utilizado para selecionar o modelo que o minimizasse. Cada modelo estimado apresenta no `R` a sua log-verosimilhança, e o AIC calculado a partir dela. Com isto, e porque ao aumentar os parâmetros do modelo, em geral, também se aumenta o ajuste do modelo aos dados, vamos utilizar o AIC para penalizar o número de parâmetros, e obter um modelo que explique tanto ou mais que os restantes, sem ter de aumentar os parâmetros (Shumway et al., 2010).

- Validação cruzada para observações independentes

  Para efeitos de previsão, os dados são divididos em conjunto de treino e teste para testar a eficácia do modelo. Contudo, modelos sobreajustados (*overfitting*) ao conjunto de treino vão ter fracas previsões do conjunto de teste. Se dividirmos o conjunto de treino em subconjuntos e fizermos a previsão de cada um a partir dos restantes obtemos um modelo com boa capacidade preditiva. O ideal é fazer

validação cruzada utilizando para previsão cada uma das observações do conjunto de dados (*leave-one-out*), mas devido ao custo computacional de repetir $N$ modelos é comum fazer validação cruzada para $k$ subconjuntos aleatórios com $k = 10$ um valor razoável para obter um bom modelo em função do número de modelos treinados (Hastie et al., 2001).

- Validação cruzada para séries temporais
  Contrariamente às observações independentes, em séries temporais os modelos e as previsões geradas dependem da correlação entre observações, portanto não podemos seguir a estratégia utilizada acima e dividir o conjunto de dados em subconjuntos aleatórios. O que podemos fazer, é selecionar o subconjunto mínimo necessário à modelação e prever a observação seguinte, e aumentar o subconjunto uma observação de cada vez, e calcular a métrica dos erros de previsão. Uma alternativa quando se tem dados sazonais, é fazer a previsão de um período completo para reduzir o custo computacional. Se tivermos dados com período $s$, selecionamos o subconjunto mínimo para modelar os dados e prevemos $s$ valores, em seguida acrescentamos as $s$ observações e continuamos o processo até esgotarmos o conjunto de treino (Hyndman e Athanasopoulos, 2014).

### 2.2.6 Medidas do erro de previsão

Dados os resíduos $\epsilon_i = x_i - \hat{x}_i,\ i = 1, ..., n$, em que $x_i$ são as observações registadas e $\hat{x}_i$ são as observações previstas pelo modelo, podemos definir algumas medidas do erro como as que se seguem (Hyndman, 2006):

- Erros dependentes de escala MAE (Mean Absolute Error)

$$mae(\epsilon_t) = \frac{1}{n}\sum_{i=1}^{n} |\epsilon_i|$$

- Erros Percentuais MAPE (Mean Absolute Percent Error)

$$mape(\epsilon_t) = \frac{1}{n}\sum_{i=1}^{n}\left|100\frac{\epsilon_i}{y_i}\right|$$

- Erros relativos MASE (Mean Absolute Scaled Error)

$$mase(\epsilon_t) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\epsilon_i}{\frac{1}{n-1}\sum_{j=2}^{n}|y_j - y_{j-1}|}\right|$$

Na secção seguinte estas medidas de erro serão utilizadas para escolher entre os modelos selecionados por minimização do AIC ou validação cruzada.

## 2.3 Transformação das variáveis

Devido à elevada assimetria dos valores de afluências, com caudas pesadas à direita, é necessário aplicar uma transformação para que a sua distribuição apresente simetria. Mesmo que os valores negativos sejam tomados como nulos, a transformação de Box-Cox não lida com os nulos devido ao logaritmo, o que não permite a sua aplicação. Assim são apresentadas duas transformações logarítmicas acrescidas de uma constante, da família $\log(x+c)$, utilizadas para reduzir a assimetria das séries de afluências.

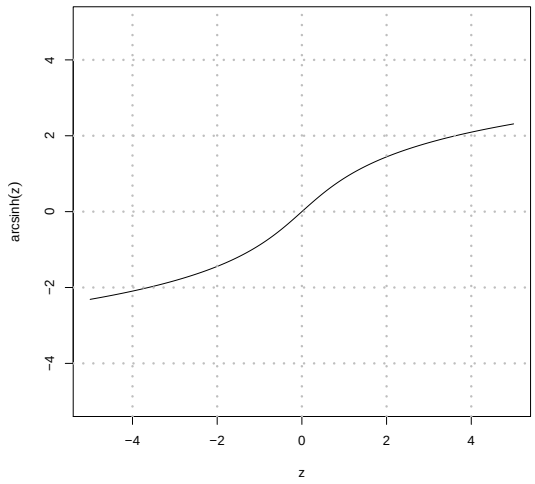


Fig. 5 – Gráfico da função $\sinh^{-1} z$, ou seja, a Transformação IHS com $\theta = 1$

### 2.3.1 Transformação de Box-Cox

Dada uma série $x$ e o parâmetro da transformação $\lambda$ temos

$$x^{(\lambda)} = \begin{cases} \log(x), \text{ se } \lambda = 0 \\ \frac{x^{\lambda}-1}{\lambda}, \text{ se } \lambda \neq 0 \end{cases} \tag{2.3.1}$$

em que $x^{(\lambda)}$ é a série transformada.

### 2.3.2 Transformação IHS (Inversa do Seno Hiperbólico)

Dada uma série de afluências $x$ e o parâmetro da transformação $\theta$ define-se a transformação IHS

$$\sinh^{-1}(\theta x)/\theta = \log\left(\theta x + \sqrt{1+(\theta x)^2}\right)/\theta \tag{2.3.2}$$

que podemos ver na figura 5. A transformação pouco altera os valores mais baixos de $x$ e reduz os valores mais altos o que os condensa e diminui o peso da cauda direita. (Hyndman, 2010)

### 2.3.3 Transformação de Wilson-Hilferty

Como definida no artigo de Raman et al. (1995)

$$x_{\nu,\tau} = \log\left(x_{\nu,\tau} + c\bar{x}_{\tau}\right) \tag{2.3.3}$$

Neste caso as observações $x_{\nu,\tau}$ correspondem ao ano $\nu = 1, ..., 55$ e ao mês $\tau = 1, ..., 12$, com $c = a/g^2$ em que $a$ é o parâmetro adimensional obtido por análise de regressão mencionado no artigo (Ochoa-Rivera et al., 2002) como $0.35$ e $g$ é o coeficiente de assimetria calculado para a série; e $\bar{x}_{\tau}$ é a média do mês $\tau$ ao longo dos anos.

# 3 Trabalho desenvolvido

O trabalho é constituído por duas partes, cuja componente computacional se realizou com R. Na primeira foram analisadas séries de afluências diárias do Sistema Lima-Cávado, para compreender o funcionamento dos AH, e de que maneira estão interligados. Além disso, estudou-se a distribuição das afluências, e propuseram-se ferramentas para superar os obstáculos inerentes aos dados diários.

Na segunda parte, o foco voltou-se para o Sistema Douro, no qual se encontram planeados novos AH. Esses AH futuros é que foram objeto de estudo, pois era pretendido modelá-los e captar as correlações entre as observações, de modo a prever as suas afluências nos anos de 2011 até 2015.

## 3.1 Análise das afluências diárias

Os dados de afluências são medições da quantidade de água, por unidade de tempo, num determinado local e quando nos referimos a afluências diárias falamos do volume de água que chega a cada AH por dia. Claramente as afluências são valores não negativos, mas podem ser nulas, como acontece nos dias de seca, por outro lado as afluências não são infinitas. Contudo, as afluências não são constantes e o caudal dos rios é muito volátil e pode oscilar rapidamente entre períodos de seca e períodos de cheia, ou vice versa. Isto faz com que as afluências limitadas por 0 tenham também valores extremos e uma distribuição assimétrica com uma cauda de valores positivos pesada. Por estas razões é geral assumir-se que uma variável aleatória das afluências siga uma distribuição de Pearson tipo III (Kottegoda, 1980).

No entanto, as técnicas de medição de afluências do curso de um rio não são 100% fiáveis e por vezes são medidos, ou até calculados, valores importantes para a gestão dos AH, mas que não se adequam à teoria. Quando isso acontece, os dados têm valores incoerentes com a teoria, negativos neste caso. Em períodos de seca, como nos meses de Verão, quando estamos perante poucos valores negativos, podemos

| Aproveitamento Hidroeléctrico | Número de Observações | Negativos (%) | | Zeros (%) | |
|---|---|---|---|---|---|
| | | 2004 | Total | 2004 | Total |
| Alto Lindoso | 8035 | 0.75 | 0.37 | 0 | 0.15 |
| Alto Rabagão | 13879 | 15.70 | 4.55 | 0.42 | 18.06 |
| Caniçada | 13879 | 13.24 | 3.83 | 0.17 | 22.70 |
| Paradela | 13879 | 17.15 | 4.96 | 3.16 | 15.19 |
| Salamonde | 13879 | 1.14 | 0.33 | 0.05 | 2.31 |
| Touvedo | 7670 | 23.35 | 12.23 | 0.27 | 8.68 |
| Vilarinho das Furnas | 13879 | 6.99 | 2.02 | 0.80 | 15.66 |
| Venda Nova | 13879 | 9.98 | 2.89 | 0.17 | 9.50 |

Tabela 1 – Tabela de zeros e negativos para as afluências do Sistema Lima-Cávado

admitir que são todos nulos correspondentes a dias de seca e prosseguir com a análise das afluências.

Uma análise superficial dos dados revela uma elevada percentagem de observações negativas e nulas como representado na Tabela 1. É de assinalar o ano de 2004 como o ano da implementação de novos métodos de medição de afluências que passaram a contemplar mais valores nulos como negativos, aumentando a proporção de negativos e diminuindo a proporção de zeros, face às séries completas.

Devido à presença dos negativos a distribuição dos dados foi comparada com a distribuição logística para cada uma das séries, como se pode ver ilustrado na figura 6. O que acontece é que os valores estão muito concentrados e próximos de zero, com os eventuais valores elevados de afluência a serem difíceis de distinguir no histograma, e a divergirem dos quantis teóricos no gráfico de quantis.

Para solucionar esta situação, a melhor opção é transformar os dados de modo que os valores mais elevados sejam diminuídos e os pequenos mantidos. E isso, em geral, consegue-se com uma transformação de Box-Cox mas, neste caso, os dados não são somente positivos, por isso a função logaritmo não é aplicável. Nesse sentido, escolheu-se uma transformação que tivesse o mesmo efeito e permitisse aplicar a da-

dos com zeros e negativos (O'Hara et al., 2010). Por este motivo foi utilizada a transformada IHS (2.3.2) e os dados comparados com a distribuição normal (figura 7). Contudo, a elevada presença de medições nulas mantém-se, e os dados apresentam dois picos no seu histograma, sendo o primeiro dos zeros. O efeito dos zeros também é facilmente observável pela sua repetição, com aspeto de degrau, nos gráficos de probabilidade acumulada, quantis e percentis.

A presença elevada de zeros com um pico distinto na distribuição dos dados não nos permite modelar corretamente os dados diários. Uma alternativa, é reduzir a presença de negativos e zeros ao somar os dados diários de modo a analisar e modelar dados semanais ou até mensais. Se os dados forem agregados por semanas, a percentagem de negativos, apesar de menor, continua a não permitir a sua modelação. Por outro

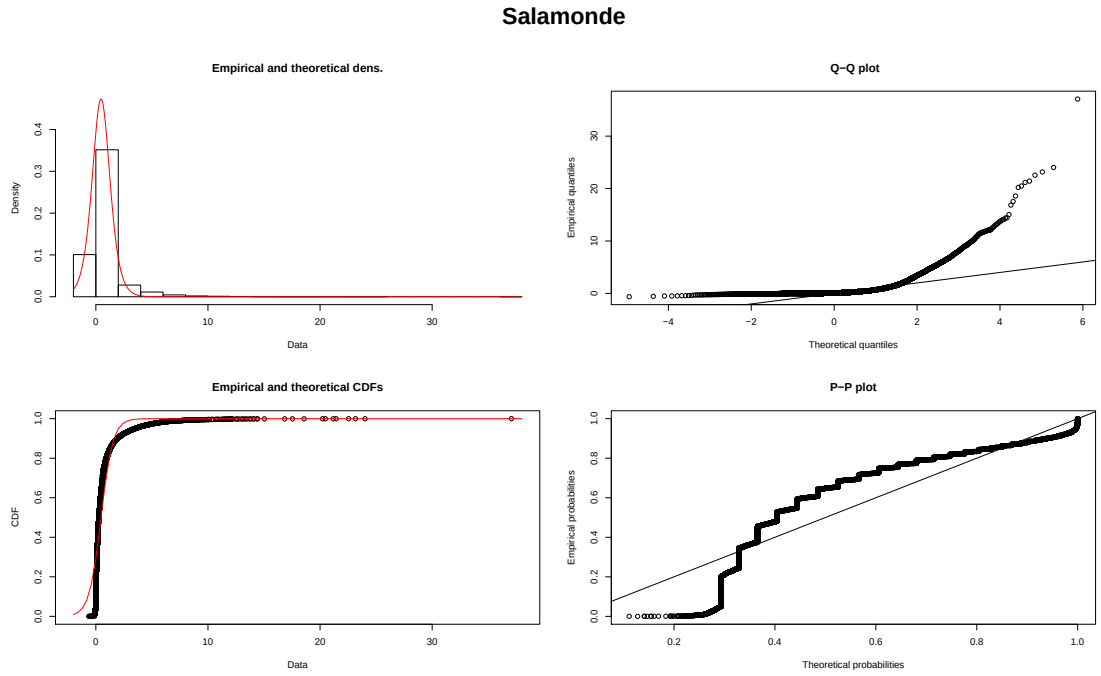


Fig. 6 – Ajuste de uma distribuição logística às afluências originais do AH de Salamonde.
Cima: Esquerda, Histograma das afluências e densidade do ajuste de uma distribuição logística; Direita, Q-Q plot contra uma distribuição logística.
Baixo: Esquerda, Função de distribuição acumulada das afluências contra a teórica de uma distribuição logística; Direita, P-P plot contra uma distribuição logística.

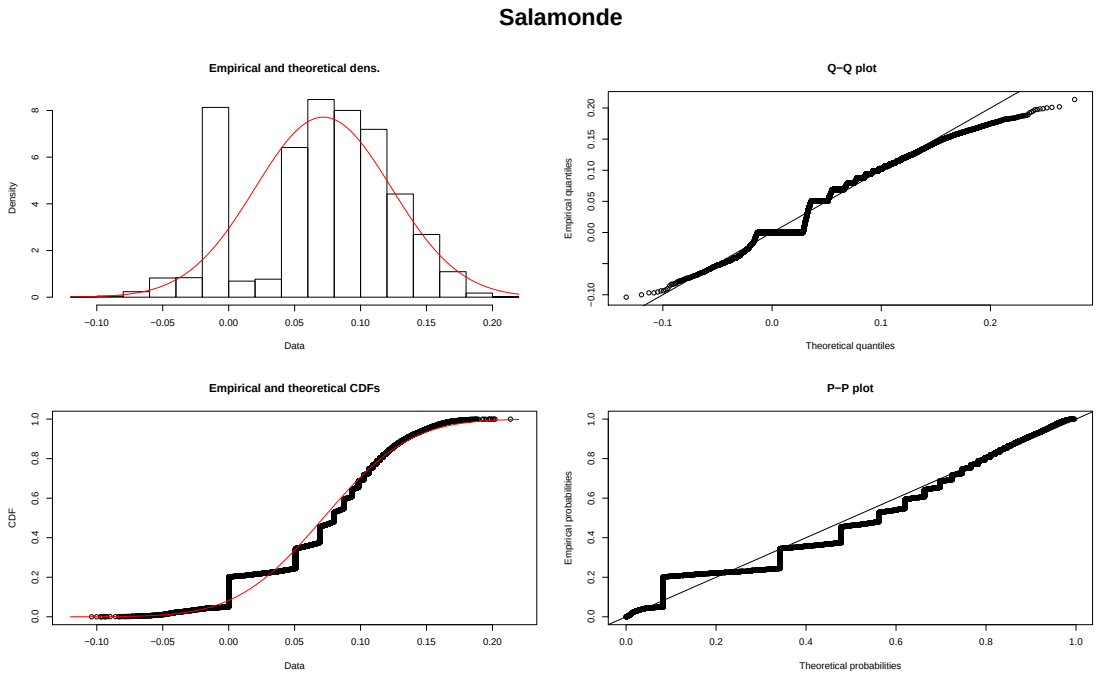


Fig. 7 – Ajuste de uma distribuição normal às afluências transformadas do AH de Salamonde.
Cima: Esquerda, Histograma das afluências transformadas e densidade do ajuste de uma distribuição normal; Direita, Q-Q plot contra uma distribuição normal.
Baixo: Esquerda, Função de distribuição acumulada das afluências transformadas contra a teórica de uma distribuição normal; Direita, P-P plot contra uma distribuição normal.

lado, a agregação mensal continua a ter valores negativos, que na sua maioria corresponde a períodos de seca. Quando esses negativos não correspondem a períodos de seca e estiverem isolados, podemos assumir que se devem a erros de medição e substituí-los com técnicas de interpolação.

## 3.2 Modelação das afluências mensais

A modelação de dados hidrológicos é um processo complexo, influenciado pela sazonalidade anual do ciclo hidrológico, pela ação humana e até pela ocorrência de anos secos ou húmidos. Tal como os dados de outras áreas, as afluências apresentam pouca variabilidade quando o seu valor é baixo, em anos secos, e muita variabilidade quando

os seus valores são elevados, em anos húmidos.

No caso do Sistema Douro, podemos considerar dois grupos distintos de postos de medição:

- o primeiro grupo constituído pelos postos sem outros a precedê-los, aqueles que estão isolados ou se apresentam a montante de uma cascata (Quinta das Laranjeiras, Vidago e Gouvães);
- o segundo grupo constituído pelos postos com postos anteriores, situados numa cascata a jusante de outros (Daivões e Fridão).

As séries de afluências têm 660 observações mensais dos anos de 1956 até 2010, mas para comparar os modelos sem variáveis exógenas com os modelos com variáveis exógenas, vamos selecionar os anos de 1974 até 2010, por compatibilidade com as variáveis exógenas de precipitação da NASA. Além disso, consideramos Agosto o mês inicial e Julho o mês final, de modo que cada 12 observações apresentem os valores mais baixos nos extremos e os mais elevados no centro. Ficamos assim com 432 observações desde Agosto de 1974 até Julho de 2010.

Por não haver algoritmos de modelos de média móvel com redes neuronais implementados, vamos só comparar modelos autorregressivos. Por esse motivo, a ordem $p$ dos modelos vai variar entre 0 e 11 na parte autorregressiva, e a ordem $P$ dos modelos vai variar entre 0 e 5 na parte autorregressiva sazonal. Ou seja, no máximo consideramos que as afluências dependem dos últimos onze meses e cinco anos de observações do mesmo mês.

Antes de prevermos as séries, temos de testar a capacidade preditiva dos modelos. Nesse sentido, vamos dividir os dados em conjunto de treino e teste, em que o conjunto de teste são os últimos quatro anos de observações, e o conjunto de treino são as observações anteriores. Assim, temos as observações desde Agosto de 1974 até Julho de 2006 como o nosso conjunto de treino, e as observações de Agosto de 2006 até Julho de 2010 como o nosso conjunto de teste.

Começamos pelo caso isolado do posto de Quinta das Laranjeiras, do primeiro grupo,

| | Séries Originais | | Transformação de Wilson-Hilferty | |
|---|---|---|---|---|
| | Sem exógenas | Com exógenas | Sem exógenas | Com exógenas |
| SARIMA | 1.6115 | 1.1244 | 1.2466 | 0.6893 |
| Redes Neuronais | X | X | X | 0.5528 |
| NNETAR | X | X | X | 0.6898 |

Tabela 2 – Tabela do MASE obtido nos modelos selecionados em cada algoritmo aplicado às afluências de Quinta das Laranjeiras

e estimamos os vários modelos SARIMA para as afluências mensais não transformadas do conjunto de treino, com o algoritmo `arima`. Como os resultados do AIC e da validação cruzada não são comparáveis, para comparar o desempenho dos modelos foi calculado o MASE para algoritmos escolhidos com AIC ou validação cruzada. Assim, o modelo com AIC mínimo foi escolhido para calcular o MASE, da primeira coluna da tabela 2. A previsão do modelo é fraca, porque este prevê a média dos valores, inflacionada pelos anos húmidos.

Na tentativa de resolver a média inflacionada, utilizamos a transformada de Wilson-Hilferty. O modelo, com AIC mínimo, diminuiu o MASE, na terceira coluna da tabela 2, mas a previsão continua fraca, por não captar os anos húmidos. A transformada retirou a inflação à média, mas o modelo continua sem captar a variabilidade das afluências.

Para captar a variabilidade introduziram-se variáveis exógenas de precipitação, de maneira a distinguir os anos secos, dos anos húmidos. Em cada posto foram selecionados os quatro valores de precipitação mais próximos, ou seja, o quadrado da grelha de valores que inclui o posto. Apesar de baixar o MASE, do modelo que minimiza o AIC, como se vê na segunda coluna da tabela 2, o modelo estimado nas afluências transformadas, com variáveis exógenas é o que obtém o MASE mais baixo para os modelos SARIMA, na quarta coluna da tabela 2.

Desta forma, porque as características das afluências são melhor modeladas com a transformada e as variáveis exógenas, e porque os modelos de redes neuronais demo-

ram mais a treinar, apenas vamos comparar a modelação das afluências transformadas com variáveis exógenas, entre modelos de diferentes algoritmos.

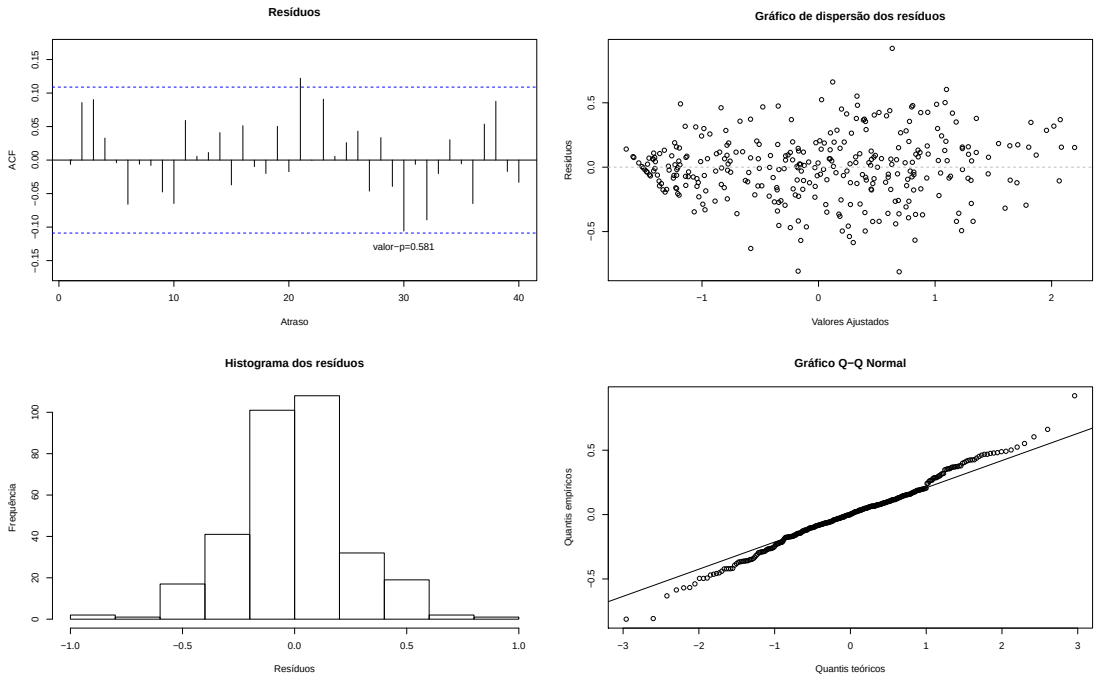


Fig. 8 – Resíduos da modelação de Quinta das Laranjeiras.
Cima: Esquerda, ACF dos resíduos; Direita, gráfico de dispersão.
Baixo: Esquerda, histograma de frequência absoluta; Direita, Q-Q plot contra distribuição normal.

De seguida, utilizamos o algoritmo `nnet` para modelar as afluências, e escolhemos a ordem do modelo pela minimização do erro de validação cruzada, dado pelo algoritmo no campo `error`. O MASE do modelo está representada na tabela 2 e teve um melhor desempenho para o conjunto de teste que o algoritmo `arima`.

Devido ao tempo de treino elevado e à preparação dos dados que o algoritmo necessita, pretende-se comparar o seu desempenho com o do algoritmo `nnetar` que facilita a utilização de redes neuronais para séries temporais. Contudo, o algoritmo não tem implementada uma ferramenta de escolha de ordem do modelo, e foi implementada a validação cruzada para séries temporais sazonais, para escolher a ordem do modelo. O MASE do modelo selecionado está apresentado na tabela 2 e teve um desempenho

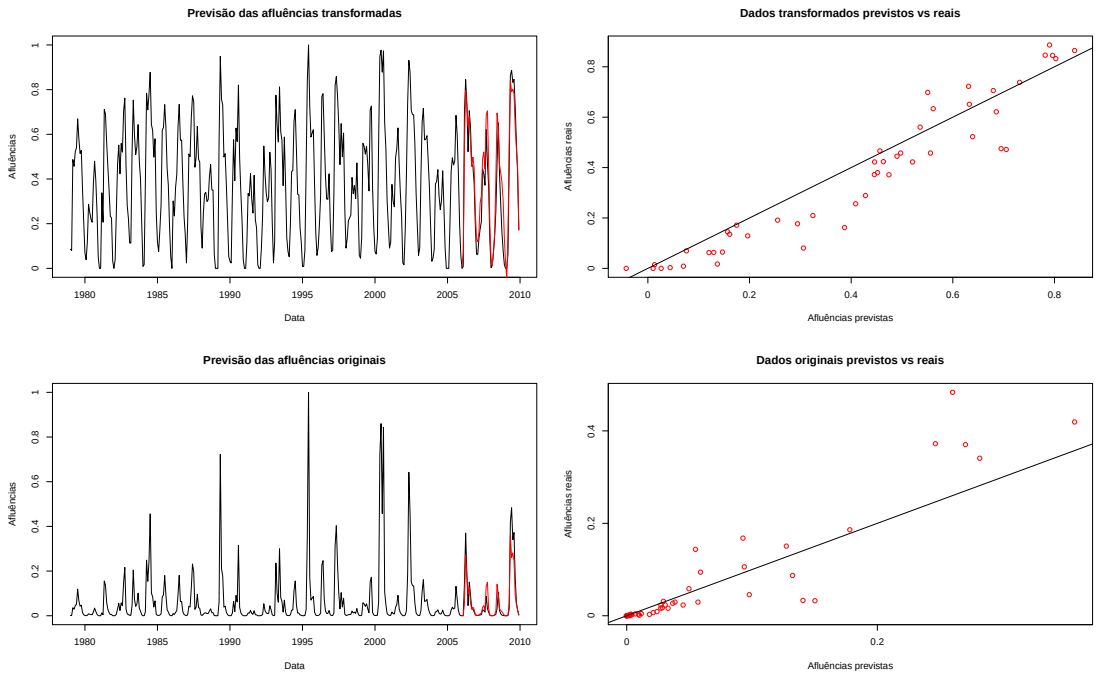


Fig. 9 – Ajuste da modelação de Quinta das Laranjeiras.
Cima: Esquerda, comparação da previsão com a série real dos dados transformados; Direita, gráfico da previsão contra os valores reais dos dados transformados.
Baixo: Esquerda, comparação da previsão com a série real dos dados originais; Direita, gráfico da previsão contra os valores reais dos dados originais.

inferior ao algoritmo `arima`.

Desta forma, o modelo escolhido para prever o conjunto de dados desde Agosto de 2010 até Julho de 2015 foi um NNSARX$(10, 5, 3, 3)_{12}$, estimado pelo algoritmo `nnet`. Os resíduos do modelo estimado para as afluências transformadas, na figura 8, apresentam correlações pouco significativas na sua ACF, normalidade no seu histograma e Q-Q plot, e estão dispersos em nuvem ao longo do tempo sem apresentar correlações ou tendências no gráfico de dispersão. Além disso, ao comparar a transformação inversa da previsão com o conjunto de teste, na figura 9, podemos ver que o modelo capta bem a variabilidade da série e, em geral, faz previsões inferiores aos valores reais. Uma característica importante para prevenir cheias nos AH.

Como as séries de Quinta das Laranjeiras têm um comportamento semelhante às sé-

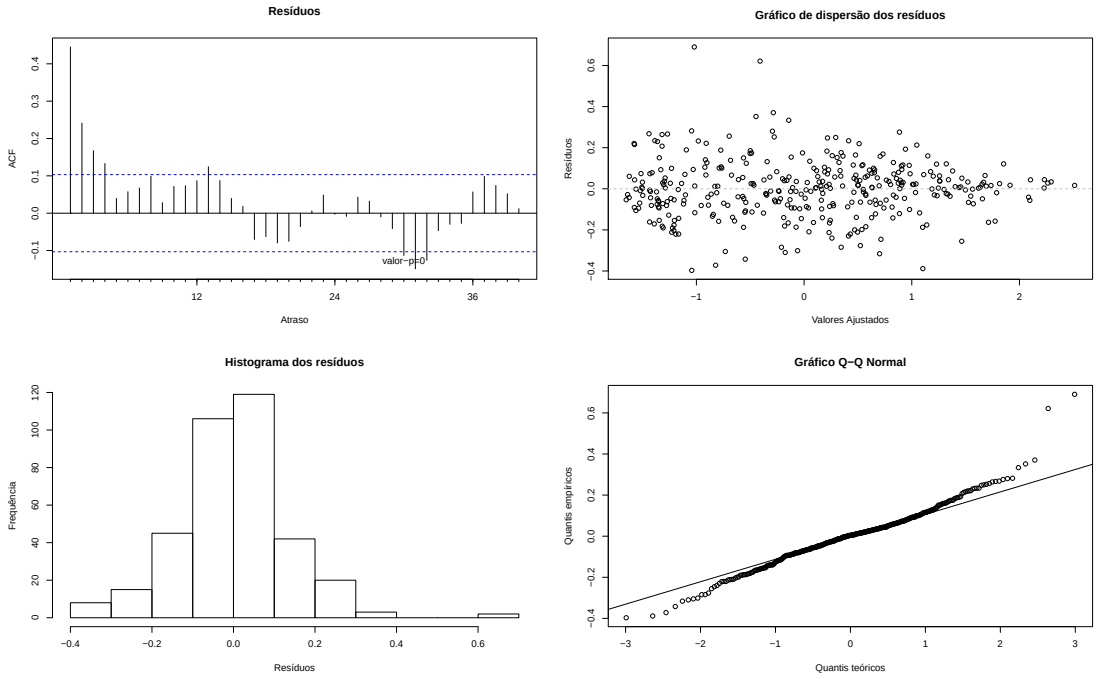


Fig. 10 – Resíduos da modelação de Daivões.
Cima: Esquerda, ACF dos resíduos; Direita, gráfico de dispersão.
Baixo: Esquerda, histograma de frequência absoluta; Direita, Q-Q plot contra distribuição normal.

ries de Vidago e Gouvães, devido à sua proximidade, e por serem também cabeceiras de uma cascata, vamos generalizar o modelo de Quinta das Laranjeiras para evitar aplicar estes passos demorados na modelação de Vidago e Gouvães, e vamos estimar o modelo NNSARX$(10, 5, 3, 3)_{12}$ com as duas séries de afluências para verificar se neste caso também obtemos um bom ajuste da previsão.

O MASE dos modelos indica um bom ajuste às séries e uma boa previsão contra o conjunto de treino. Assim, podemos utilizá-lo com Vidago e Gouvães para obter uma

| | SARIMA | Redes Neuronais | NNETAR |
|---|---|---|---|
| Daivões | 0.4362 | 0.8538 | 0.4104 |

Tabela 3 – Tabela do MASE obtido nos modelos selecionados em cada algoritmo aplicado às afluências de Daivões

boa previsão das suas séries.

No segundo grupo, estimamos vários modelos para as afluências de Daivões, só que desta vez utilizamos as afluências transformadas com variáveis exógenas, como determinado para Quinta das Laranjeiras, em que substituímos os dados de precipitação pelas afluências dos postos a montante, que neste caso são Vidago e Gouvães. Apenas usamos afluências dos postos a montante, do mesmo mês que estamos a modelar, ou seja, o modelo de Daivões tem somente duas variáveis exógenas. Selecionamos um modelo de cada algoritmo e representamos os seus MASE na tabela 3. No caso de Daivões, o algoritmo `nnet` foi o que teve o MASE mais elevado, sendo o `nnetar` aquele com o melhor desempenho com um modelo estimado NNSARX$(1, 2, 2)_{12}$.

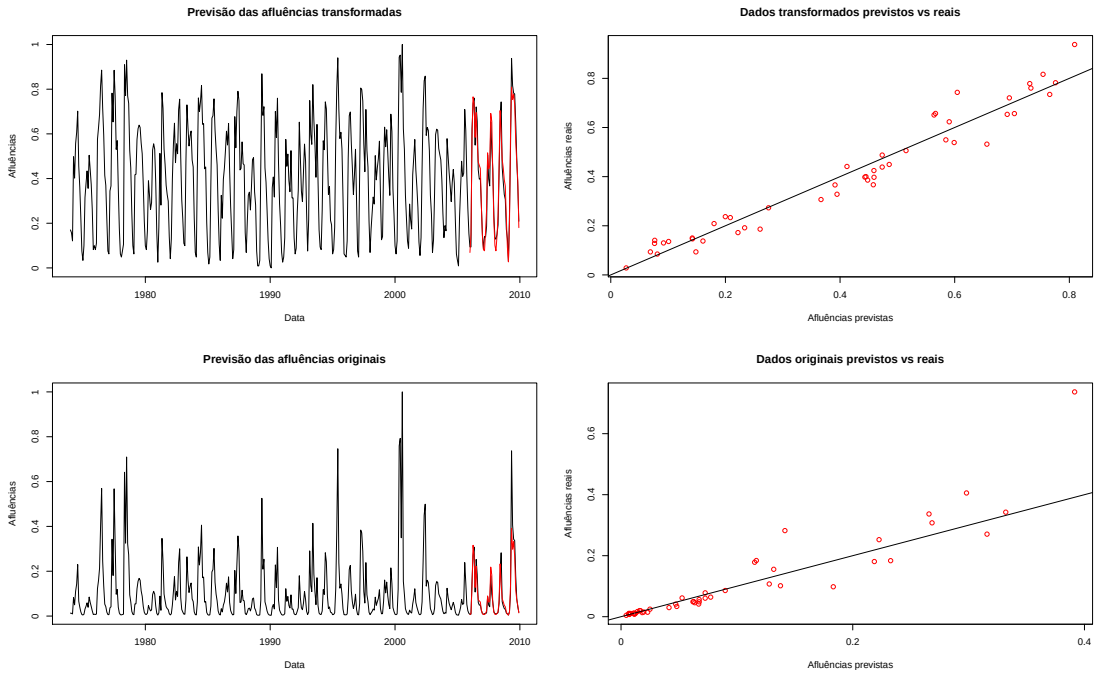


Fig. 11 – Ajuste da modelação de Daivões.
Cima: Esquerda, comparação da previsão com a série real dos dados transformados; Direita, gráfico da previsão contra os valores reais dos dados transformados.
Baixo: Esquerda, comparação da previsão com a série real dos dados originais; Direita, gráfico da previsão contra os valores reais dos dados originais.

Os resíduos do modelo, representados na figura 10, apresentam normalidade no his-

tograma e Q-Q plot, mas têm alguns resíduos mais elevados que os restantes na sua dispersão e vê-se na ACF que o modelo não captou todas as correlações dos dados. Isso explica porque é que o modelo não consegue captar um mês anormalmente húmido no ano de 2010, como é visível na figura 11.

## 3.3 Previsão das afluências mensais

A previsão de cada posto foi feita com os modelos utilizados na modelação do conjunto de treino. Foi estimado um modelo para cada série completa, e feita a sua previsão para 60 observações, cinco anos. A previsão foi feita de acordo com o que está disponível para cada algoritmo, à exceção do algoritmo `nnet`, para o qual foi implementada uma previsão recursiva. Para os postos do primeiro grupo utilizou-se o modelo NNSARX$(10, 5, 3, 3)_{12}$, estimado pelo algoritmo `nnet`. No segundo grupo também foi utilizado um modelo de redes neuronais, mas estimado pelo algoritmo `nnetar`, o modelo NNSARX$(1, 2, 2, 3)_{12}$. Em seguida, podem ser analisadas as previsões de cada posto de medição, bem como notar as características partilhadas entre as diferentes séries, tal como as variações entre regime húmido e seco.

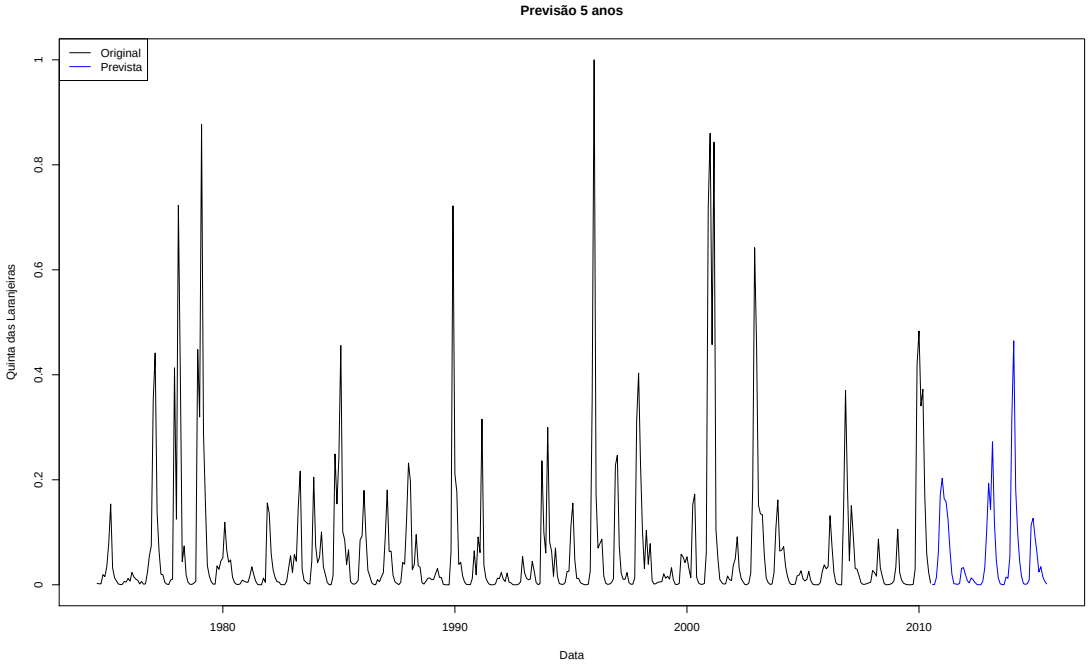


Fig. 12 – Previsão das afluências do posto de Quinta das Laranjeiras

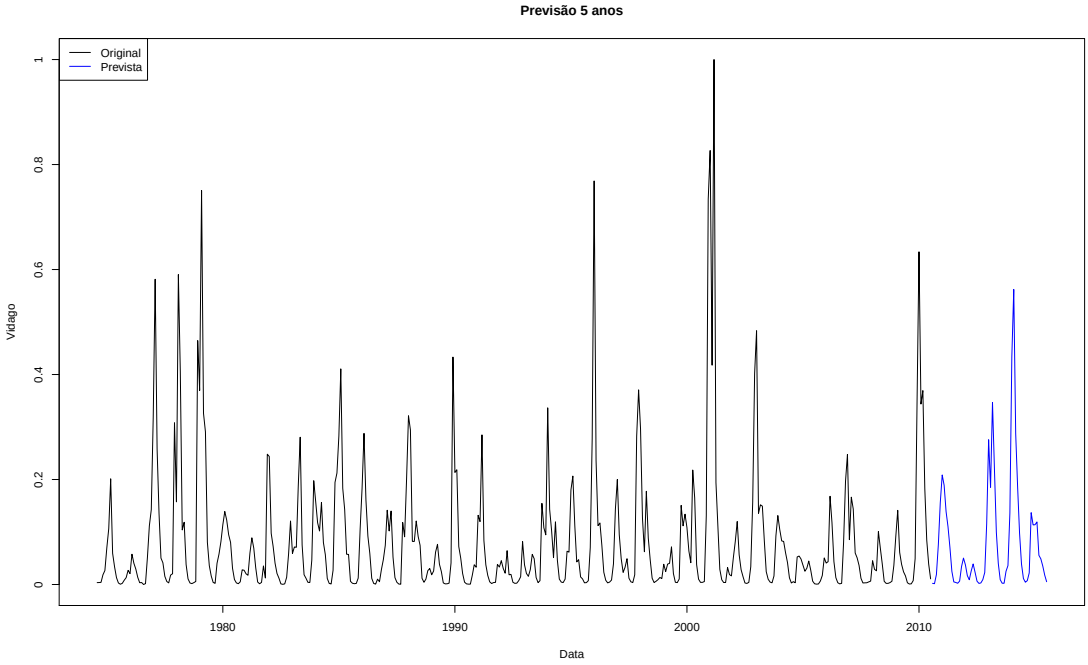


Fig. 13 – Previsão das afluências do posto de Vidago

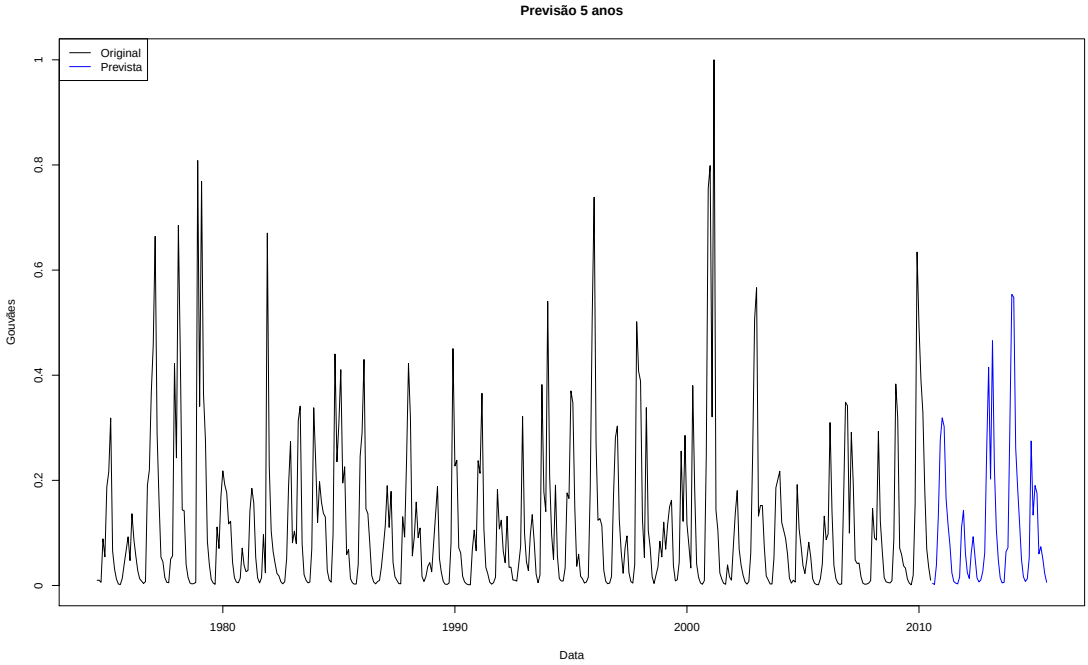


Fig. 14 – Previsão das afluências do posto de Gouvães

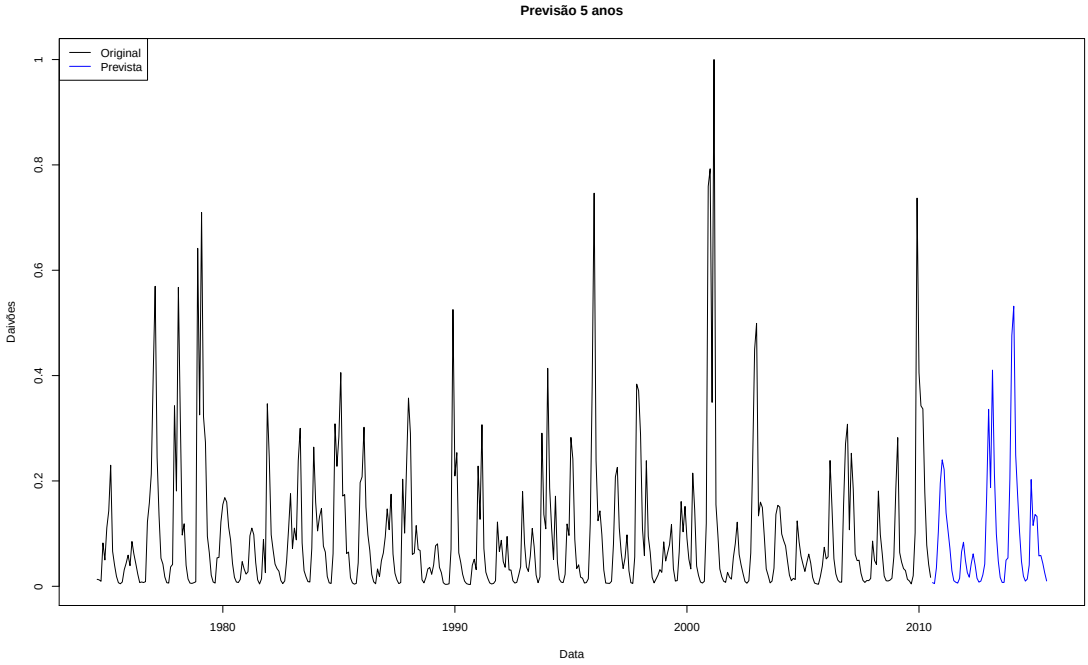


Fig. 15 – Previsão das afluências do posto de Daivões

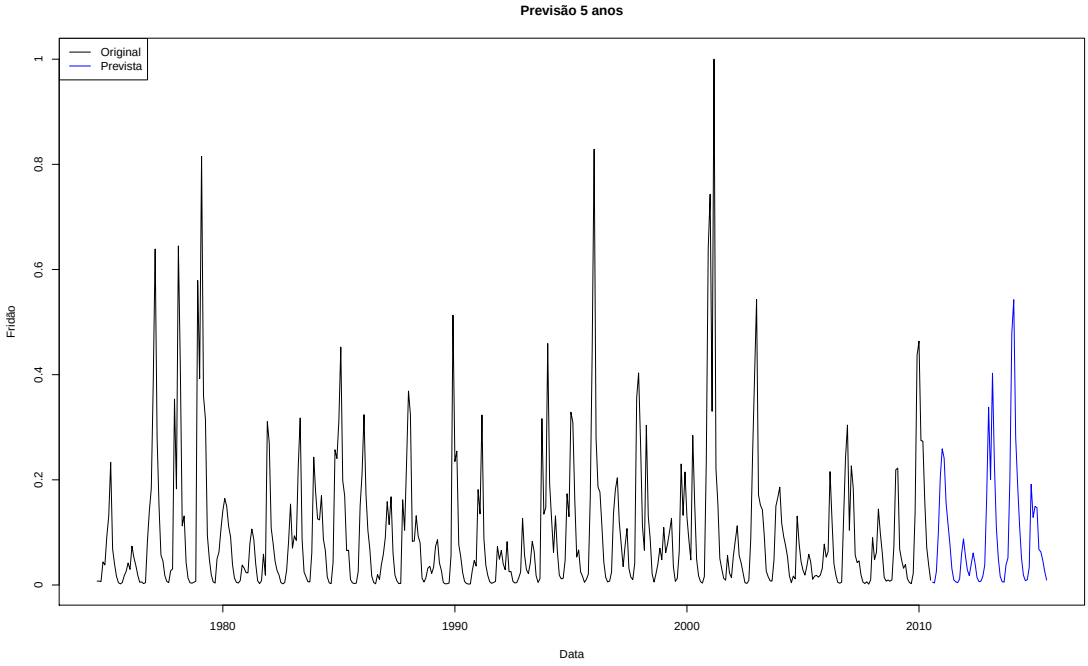


Fig. 16 – Previsão das afluências do posto de Fridão

# 4 Considerações finais

As afluências diárias apresentam um problema de modelação por terem sazonalidade não inteira, cada ano tem aproximadamente 365,25 dias. Além disso, os dados armazenados têm uma porção significativa de valores negativos e nulos que alteram a sua distribuição e tornam a previsão pouco fiável.

Por outro lado, utilizar as séries mensais com periodicidade anual acrescidas de dados de precipitação permite preencher as lacunas que impossibilitavam uma correta simulação da rede elétrica.

O algoritmo `arima` foi o mais rápido a estimar os parâmetros, com as redes neuronais, do algoritmo `nnetar`, a apresentarem bons resultados com uma pequena perda de velocidade. O algoritmo `nnet`, apesar de mais lento, é o mais personalizável e permite utilizar como *inputs* observações não consecutivas, ou sazonalmente separadas. No geral, o algoritmo mais prático foi o `nnetar` pela facilidade de utilização, bons resultados na previsão e tempo computacional moderado.

A modelação dos postos de Daivões e Fridão deixou espaço para melhorias, por não ter conseguido captar na totalidade a correlação com observações anteriores. Contudo, os postos restantes apresentaram boas previsões para o conjunto de treino, com a vantagem de utilizar o mesmo modelo, treinado para apenas um dos postos.

## 4.1 Trabalho futuro

Seria interessante utilizar dados de afluências diárias limpos de valores negativos para aplicar o método dos fragmentos sugerido por Svanidze (Svanidze, 1980) ao agregar as observações em matrizes $31 \times 12$ preenchendo cada coluna com as observações dos 12 meses do ano. Este método permite preservar as correlações entre dados do mesmo mês presentes na mesma coluna, analisar as correlações entre os meses consecutivos presentes em colunas consecutivas e ainda observar as correlações anuais entre observações de diferentes matrizes (anos) com o mesmo índice.

Quanto aos dados mensais podem-se validar as previsões com dados reais e melhorá-las a partir de novos dados importantes para a afluência como a permeabilidade dos solos, desvios agrícolas, evaporação nas albufeiras e observações de afluentes anteriores sem relevância energética.

O algoritmo de `nnet`, em que se utiliza uma matriz de observações em vez de uma série temporal, poderia facilitar a sua utilização se tivesse uma função que permitisse utilizar uma série temporal e escolher as observações a utilizar. De forma, a escolher um qualquer vetor de observações, além daquelas utilizadas na estrutura dos modelos autorregressivos. Além disso, implementar redes neuronais para modelos ARMA, explicaria, em geral, com menos parâmetros os conjuntos de dados, o que permite modelações mais precisas e rápidas.

Na modelação dos postos de Daivões e Fridão foram utilizadas variáveis exógenas do(s) posto(s) anterior(es), para o mesmo mês que se pretende prever. Se utilizarmos mais observações, de meses anteriores, devemos conseguir captar as correlações que os modelos usados não conseguiram.

# Bibliografia

[1] Hastie, T., Tibshirani, R. e Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[2] Hipel, K. e McLeod, A. *Time Series Modelling of Water Resources and Environmental Systems*. Vol. 45. Developments in Water Science. Elsevier, 1994. URL: http://www.sciencedirect.com/science/article/pii/S016756480870652X.

[3] Hyndman, R. J. "Another Look at Forecast-Accuracy Metrics for Intermittent Demand". Em: *Foresight, International Journal of Applied Forecasting* (2006), pp. 43–46.

[4] Hyndman, R. J. *Transforming data with zeros*. [Online; último acesso 20 de Janeiro de 2016]. 2010. URL: http://robjhyndman.com/hyndsight/transformations.

[5] Hyndman, R. J. e Athanasopoulos, G. *Forecasting: principles and practice*. OTexts, 2014. URL: https://www.otexts.org/book/fpp.

[6] Kottegoda, N. *Stochastic Water Resources Technology*. Macmillan, 1980.

[7] Murteira, B. J. F., Müller, D. A. e Turkman, K. F. *Análise de sucessões cronológicas*. Lisboa: McGraw-Hill, 1993.

[8] Ochoa-Rivera, J. C., García-Bartual, R. e Andreu, J. "Multivariate synthetic streamflow generation using a hybrid model based on artificial neural networks". Em: *Hydrology and Earth System Sciences* 6.4 (2002), pp. 641–654. URL: http://www.hydrol-earth-syst-sci.net/6/641/2002/.

[9] O'Hara, R. B. e Kotze, D. J. "Do not log-transform count data". Em: *Methods in Ecology and Evolution* 1.2 (2010), pp. 118–122. URL: http://dx.doi.org/10.1111/j.2041-210X.2010.00021.x.

[10] Raman, H. e Sunilkumar, N. "Multivariate modelling of water resources time series using artificial neural networks". Em: *Hydrolog. Sci. J* (1995), pp. 145–163.

[11] Ripley, B. D. e Hjort, N. L. *Pattern Recognition and Neural Networks*. New York, NY, USA: Cambridge University Press, 1995.

[12] Salas, J. *Applied Modeling of Hydrologic Time Series*. Water Resources publication. Water Resources Publications, 1980. URL: http://www.engr.colostate.edu/ce/facultystaff/salas/publications.shtml.

[13] Shumway, R. e Stoffer, D. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer New York, 2010. URL: http://www.stat.pitt.edu/stoffer/tsa3/.

[14] Svanidze, G. *Mathematical Modeling of Hydrologic Series for Hydroelectric and Water Resources Computations*. Mathematical modeling of hydrologic series. Water Resources Publications, 1980.

# Apêndice

O trabalho computacional aplicado sobre os dados foi feito num computador com

- Processador: Intel(R) Core(TM)2 Duo CPU E8400 @3.00GHz 3.00GHz
- RAM: 4,00GB
- Sistema Operativo: Windows 7 Enterprise (64 bits) Service Pack 1

e recurso ao software R (www.r-project.org)

```
> sessionInfo()
R version 3.3.0 (2016-05-03)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1 (build 7601)

locale:
[1] LC_COLLATE=Portuguese_Portugal.1252  LC_CTYPE=Portuguese_Portugal.1252
[3] LC_MONETARY=Portuguese_Portugal.1252 LC_NUMERIC=C
[5] LC_TIME=Portuguese_Portugal.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] nnet_7.3-12           lubridate_1.5.6       caschrono_1.4
 [4] timeSeries_3022.101.2 its_1.1.8             Hmisc_3.17-4
 [7] ggplot2_2.1.0         Formula_1.2-1         survival_2.39-2
[10] lattice_0.20-33       e1071_1.6-7           forecast_7.1
[13] timeDate_3012.100     zoo_1.7-13            tseries_0.10-35
```

```
loaded via a namespace (and not attached):
 [1] Rcpp_0.12.5         magrittr_1.5        cluster_2.0.4
 [4] splines_3.3.0       munsell_0.4.3       colorspace_1.2-6
 [7] quadprog_1.5-5      stringr_1.0.0       plyr_1.8.3
[10] tools_3.3.0         parallel_3.3.0      grid_3.3.0
[13] data.table_1.9.6    gtable_0.2.0        latticeExtra_0.6-28
[16] class_7.3-14        Matrix_1.2-6        gridExtra_2.2.1
[19] RColorBrewer_1.1-2  acepack_1.3-3.3     rpart_4.1-10
[22] fracdiff_1.4-2      stringi_1.0-1       scales_0.4.0
[25] chron_2.3-47        foreign_0.8-66
```

Com base nas versões utilizadas, os resultados das modelações podem ser reproduzidas pois foi fixada a semente do gerador de números aleatórios antes de cada modelo.

```
set.seed(1, kind = "Mersenne-Twister", normal.kind = "Inversion")
```

**Exemplo de modelação de um posto de medição**

*### Modelação com SARDX do posto Q. Laranjeiras com precipitação*
*#*
*# André Moutinho <– andre.moutinho@ren.pt*
*# Última edição: 20160415*

source(”sc20160420_import_nasa.R”) *#Variáveis exógenas NASA*

x<–postos$qlar_t[224:660] *#Observações*
x.datas<–postos[224:660,1] *#Datas das observações*

source(”TransfWH.R”) *#Transformação de Wilson–Hilferty*

```
## Transformação dos dados da NASA

chuva1.WH<−transfWH(precip.qlar.mes.ne[1:384])
chuva1<−chuva1.WH$Z
Ns<−length(chuva1)

chuva2.WH<−transfWH(precip.qlar.mes.nw[1:384])
chuva2<−chuva2.WH$Z
# length(chuva2)

chuva3.WH<−transfWH(precip.qlar.mes.se[1:384])
chuva3<−chuva3.WH$Z
# length(chuva3)

chuva4.WH<−transfWH(precip.qlar.mes.sw[1:384])
chuva4<−chuva4.WH$Z
# length(chuva4)


## Criação da matriz de variáveis exógenas

lags.s<−7
exog.tot1<−matrix(rep(chuva1,length.out=(Ns+1)*(lags.s+1)),
nrow=Ns+1,ncol=lags.s+1)
# View(exog.tot)
exog1<−exog.tot1[1:372,(lags.s+1):1]
# View(exog)
```

```
exog.tot2<−matrix(rep(chuva2,length.out=(Ns+1)*(lags.s+1)),
nrow=Ns+1,ncol=lags.s+1)
exog2<−exog.tot2[1:372,(lags.s+1):1]

exog.tot3<−matrix(rep(chuva3,length.out=(Ns+1)*(lags.s+1)),
nrow=Ns+1,ncol=lags.s+1)
exog3<−exog.tot3[1:372,(lags.s+1):1]

exog.tot4<−matrix(rep(chuva4,length.out=(Ns+1)*(lags.s+1)),
nrow=Ns+1,ncol=lags.s+1)
exog4<−exog.tot4[1:372,(lags.s+1):1]

exog<−rbind(matrix(NA,nrow=60,ncol=4*NCOL(exog1)),
cbind(exog1,exog2,exog3,exog4))


## Transformação das observações
x.WH<−transfWH(x)

z<−x.WH$Z[1:432]
Nx<−length(z)

## Conjunto de teino e teste
matriz.nnet<−cbind(z,exog)

corte<−384

treino<−matriz.nnet[1:corte,]
teste<−matriz.nnet[(corte+1):dim(matriz.nnet)[1],]
```

```
## Escolha do número de meses anteriores de variáveis exógenas
best.aic<−rep(NA,8)
for (ll in 1:8){
        lags<−c(1+1:ll,9+1:ll,17+1:ll,25+1:ll)
        best.aic[ll]<−arima(matriz.nnet[,1],order=c(0,0,0),n.cond=60,
                xreg=matriz.nnet[,lags])$aic
}
which(best.aic==min(best.aic,na.rm=TRUE))# ll=8, aic=4325.084

## Inicialização de variáveis
pp<−11;PP<−5;#lags<−c(1:pp,12*(1:PP))

mod.qlar<−vector(”list”,12*6)
prev.qlar<−vector(”list”,12*6)



burnin<−60
fim<−floor((corte−burnin)/12)

ii<−0;jj<−0;ll<−8;lags<−c(1+1:ll,9+1:ll,17+1:ll,25+1:ll)

## Ciclo para calcular os vários modelos
for (jj in 0:PP){
        for(ii in 0:pp){
                try({
                set.seed(1, kind = ”Mersenne−Twister”, normal.kind = ”Inversion”)
                mod.qlar[[(pp+1)*jj+ii+1]]<−arima(treino[,1],order=c(ii,0,0),
                seasonal=list(order=c(jj,0,0),period=12),n.cond=60,
```

```
                xreg=treino[,lags])

                prev.qlar[[(pp+1)*jj+ii+1]]<−predict(
                mod.qlar[[(pp+1)*jj+ii+1]],n.ahead=48,
                newxreg=teste[,lags])
                })
                print(c(ii,jj))

}}
```