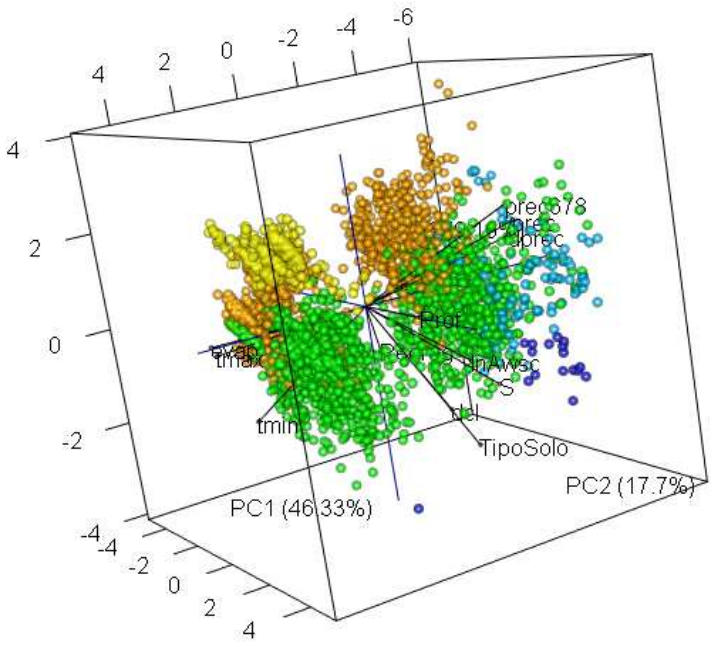
Universidade de Aveiro Departamento de Matemática
2012

Margarida Isabel Mendes Silva

CONTRIBUIÇÕES PARA O ESTUDO DO *Site Index* DA *Eucalyptus globulus*



Universidade de Aveiro Departamento de Matemática
2012

Margarida Isabel Mendes Silva

CONTRIBUIÇÕES PARA O ESTUDO DO *Site Index* DA *Eucalyptus globulus*

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Doutora Maria Manuela Souto de Miranda, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Dedico este trabalho aos meus pais, Paulo e Idalina, pelo incansável apoio e incentivo, e à minha filha, Catarina, cujo comportamento exemplar me permitiu dedicar muitas horas a este trabalho.

o júri

presidente

Doutora Isabel Maria Simões Pereira
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Doutor Gabriel Dehon Rezende
Director de Investigação Florestal do RAIZ, Instituto de Investigação da Floresta e Papel

Doutora Maria da Conceição Esperança Amado
Professora Auxiliar do Departamento de Matemática do Instituto Superior Técnico, Universidade Técnica de Lisboa

Doutora Maria Manuela Souto de Miranda
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

agradecimentos

À minha orientadora, Professora Doutora Maria Manuela Souto de Miranda, pela disponibilidade, incentivo, conhecimentos transmitidos e orientação científica.

Ao *grupo Portucel Soporcel* pela promoção deste trabalho e disponibilização de dados.

Ao Diretor Florestal do RAIZ Doutor Gabriel Dehon, pela oportunidade que criou e incentivo no desenvolvimento deste trabalho.

À Diretora de Finanças e Markting do RAIZ Engenheira Leonor Guedes, pelo incentivo e amizade demonstrados.

À minha colega de mestrado Sara Escudeiro Cruz, pela disponibilidade em ajudar-me sempre que surgiram dificuldades.

A todos os meus colegas do RAIZ pelo apoio, incentivo, amizade demonstrados e contributos para esta dissertação.

Agradeço em particular aos meus pais, Paulo e Idalina, por terem estado sempre atentos e disponíveis para ajudar quando foi necessário, aos meus irmãos, Claúdia e Zé e aos meus amigos, Maria João, Helena, Elisabete, António Paulo, Carlos e Susana, pelo incentivo, carinho e amizade.

palavras-chave

Site Index, Eucalyptus globulus, fatores edafo-climáticos*,* produtividade florestal, regressão em componentes principais, robustez*, outliers*.

resumo

O *Site Index* é um indicador da produtividade florestal que, no caso da espécie *Eucalyptus globulus*, corresponde à altura dominante dos povoamentos à idade de referência de 10 anos.

O presente trabalho consiste num estudo de predição do *Site Index* para essa espécie, em função de parâmetros edafo-climáticos, aplicando métodos da análise estatística multivariada a um conjunto de observações recolhidas em Portugal. Os métodos estatísticos usados convencionalmente não são robustos, no sentido em que são muito sensíveis a observações discordantes (*outliers*) e a outros afastamentos dos pressupostos dos modelos, pelo que podem conduzir a resultados distorcidos. Atualmente existem versões alternativas robustas, mas não estão muito divulgadas na investigação florestal.

No estudo aplicaram-se os métodos convencionais seguidos das respetivas versões robustas e exploraram-se diferenças nos resultados. De entre as variáveis ambientais relacionadas com o *Site Index*, identificaram-se as que mais contribuem para explicar a variabilidade das observações, aplicando Análise de Componentes Principais. Um fator caraterizado por variáveis climáticas (o *deficit-hídrico*), distinguiu claramente dois grupos de indivíduos que foram caraterizados pela Análise de Agrupamentos. O grupo com *deficit-hídrico* encontra-se em regiões cujas condições climáticas se traduzem em deficiente disponibilidade da água para potenciar o crescimento das plantas, enquanto o grupo sem *deficit-hídrico* surge em regiões onde a disponibilidade de água não é limitante para o crescimento. Para cada grupo, a dependência do *Site Index* relativamente às variáveis ambientais foi modelada por Regressão Linear em Componentes Principais, para superar as dificuldades decorrentes da correlação entre diversas dessas variáveis. Do ponto de vista da modelação, o trabalho não conduziu a avanços relevantes em relação às publicações existentes, mas ainda permitiu identificar as variáveis climáticas que, em cada grupo, mais contribuem para a distribuição do *Site Index* e evidenciar os benefícios da utilização de métodos robustos.

keywords

Site Index, Eucalyptus globulus, site factors, forest productivity, principal components regression, robustness, outliers.

abstract

The *Site Index* is an indicator of forest productivity. In the case of *Eucalyptus globulus* it corresponds to the height of dominant stand evaluated at the age of reference of 10 years.

The present document consists of a study of the prediction of the *Site Index* of that species as a function of site factors parameters. A data set sampled at Portugal was analyzed with multivariate statistical methods. Those methods are non robust when they are applied in a conventional way, in the sense that they are very sensitive to outliers and to deviations of the assumptions of the model, thus they might produce misleading results. Nowadays there exist alternative robust versions which give more safe results, but they are not very disseminated among forestry researchers.

The study was developed with both the conventional and the robust approaches and pointing out different conclusions whenever they are relevant. Among environmental variables, it was used Principal Component Analyze for selecting those with major contributions for explaining the variability in data. That method made possible to identify two groups of observations located in different geographical regions. A factor characterized by climate variables (*hidryc-deficit*) clearly distinguishes the two groups of individuals which were characterized by a Cluster Analyze. The group with water deficit corresponds to regions whose climatic characteristics traduce poor availability of water for enhancing the growth of the plants, while the group without water deficit is located in regions where the availability of water is not a limiting factor for plants' growth. For each group of observations the dependence of the Site Index on the environmental variables was modeled by Linear Regression in Principal Components, thus overcoming the problems caused by correlation between several environmental variables. The study did not produce a relevant improvement in modeling the *Site Index*; nevertheless it was possible to identify the variables which contribute more for its distribution and to highlight the benefits of using robust statistical methods.

Conteúdo

Lista de Símbolos e Siglas

dap Diâmetro à altura do peito (cm)

exp_{pond} transformada da Exposição

G Área basal (m^2)

h_{dom} Altura dominante (m)

k Número de grupos pretendidos na análise de agrupamentos

$lnAwsc$ logaritmo da quantidade de água no solo disponível para a planta

$prec_{678}$ Transformada da Precipitação média anual nos meses mais quentes(mm)

R^2 Coeficiente de Determinação

$tAlt$ Transformada da Altitude(m)

$tEvap$ Transformada da Evapotranspiração média anual(mm)

$tPrec$ Transformada da Precipitação média anual (mm)

AIC *Akaike Information Criterion* - Critério de Informação de Akaike

alt Altitude do local(m)

AMA Acréscimo Médio Anual em Volume ($m^3/ha/ano$)

awsc Quantidade de água no solo disponível para as plantas (mm)

BP *Breakdown point*(Ponto de rotura)

dcl Declive do solo (%)

dprec Número de dias com precipitação superior a 1mm

evap Evapotranspiração (mm)

exp Exposição solar (°)

IF *Influence Function*(Função de Influência)

IRWLS *Iteratively Reweighted Least Squares* - mínimos quadrados iterativamente pesados

KS Teste de ajustamento à distribuição *Normal* de Kolmogorov-Smirnov

lnAwsc logaritmo da quantidade de água no solo disponível para as plantas

MAD *Median Absolute Deviation* (Desvio absoluto mediano)

MBP Modelos de Base Processual

MCD *Minimum Covariance Determinant Estimator*

ME Modelos Empíricos

MH Modelos Híbridos

Pedreg Pedregosidade do solo (%)

prec Precipitação média anual (mm)

Prof Profundidade do solo (cm)

RCA *Robust Cluster Analysis* - Análise de Agrupamentos robusta

RMLR *Robust Multiple Linear Regression* - Regressão Linear Múltipla robusta

RPCA *Robust Principal Component Analysis* - Análise de Componentes Principais robusta

S Site index - Altura dominante à idade de referência de 10 anos (m)

SIG Sistema de Informação Geográfica

SSE Explained Sum of Squares

SSR Residual Sum of Squares

SST Total Sum of Squares

TIN Triangulated Irregular Network

tmax Temperatura média máxima anual ($^{\circ}C$)

tmin Temperatura média mínima anual ($^{\circ}C$)

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Na gestão florestal a utilização de modelos estatísticos é fundamental como suporte de decisões. São particularmente úteis os modelos com capacidade de predizer efeitos relacionados com as alterações climáticas. Existem fundamentalmente três tipos de modelos florestais: modelos de base processual (MBP); modelos empíricos (ME) e modelos híbridos (MH) .

Os MBP baseiam-se em processos fisiológicos que determinam o crescimento das plantas. Uma das limitações na aplicação destes modelos prende-se com o ainda insuficiente conhecimento sobre alguns processos importantes, ou com a inexistência de dados que permitam parametrizá-los para determinada espécie; outra limitação é necessitarem de muitos dados difíceis de recolher ou que não estão disponíveis.

Os modelos empíricos (ME) de suporte à gestão florestal, tipicamente baseiam-se em análises estatística da dependência de variáveis alvo, tais como volume de madeira, com base num número de variáveis independentes (regressoras), recolhidas a partir de inventários florestais e dados dos locais [14]. Estes modelos enquadram-se no pressuposto de condições locais fixas [37] e são frequentemente inadequados em condições de alteração ambiental. Contudo, recentes desenvolvimentos permitem a aplicação de ME em condições de alteração ambiental, nomeadamente, através do desenvolvimento de relações entre produtividade e ambiente, em geral, aplicando técnicas estatísticas tais como regressão linear [43], [15] e [18], modelos de regressão linear generalizados [36] entre outros [4] e [3].

Nestes estudos, a escala espacial e a espécie florestal sobre a qual foram aplicadas estas metodologias varia imenso, pelo que, as variáveis ambientais consideradas em cada caso e a *performance* obtida com os modelos produzidos também difere. Apesar da *performance* dos modelos obtidos não ser muito elevada, e apenas os tornar viáveis para serem

aplicados a determinadas escalas, o estabelecimento deste tipo de modelos está a ganhar novamente importância. As razões para este ganho de importância prendem-se com o fato de: os ME serem mais fáceis de aplicar em situações reais do que os MBP, uma vez que requerem muito menos informação, e ao desenvolvimento de novas metodologias estatísticas de modelação de dados multivariados, potenciadas pelo incremento das capacidade computacionais de processamento [14].

Finalmente, os MH surgem procurando contornar as limitações dos ME e MBP, integrando estes dois tipos de modelos e procurando fazer uso do que ambos têm de melhor.

O presente trabalho enquadra-se no âmbito dos ME, em particular, no desenvolvimento de relações entre produtividade e ambiente, com o objetivo de identificar quais as variáveis ambientais que condicionam o crescimento das árvores.

Neste capítulo faz-se uma pequena introdução a alguns conceitos fundamentais relacionados com produção florestal. São apresentadas a motivação e os objetivos gerais desta dissertação, bem como a organização da mesma.

1.1 Conceitos Florestais Fundamentais

O *Site index* é uma designação internacionalmente usada como indicador da capacidade produtiva de um povoamento florestal, que se baseia na altura das árvores dominantes de um povoamento a uma idade de referência [37].

O termo *site* refere-se a um local com determinada localização geográfica, considerado homogéneo em termos do seu ambiente físico e biológico, e capaz de sustentar o crescimento de determinado tipo de floresta. A qualidade de um *site* reflete as caraterísticas físicas e biológicas de uma localização geográfica. As propriedades que determinam a qualidade do local são-lhe genericamente inerentes: clima, solo, topografia e vegetação. Podem ser influenciadas pelas práticas de condução/gestão florestal (silvicultura) e pela espécie/genótipo utilizado, e podem ainda sofrer variações temporais. A produtividade florestal é uma estimativa quantitativa do potencial dum local para produção de biomassa lenhosa (madeira), sendo avaliada em termos de volume, em metros cúbicos (m^3) de madeira [37].

O volume de madeira é estimado por amostragem, em metros cúbicos por hectare (m^3/ha), sendo o hectare (ha) a unidade territorial de referência. São vários os métodos de amostragem utilizados na avaliação do volume, todos produzindo estimativas com elevada

precisão. Num estudo realizado por West [47], em floresta de eucalipto na Austrália, em 1979, aplicando o método considerado menos preciso, a diferença entre volume real e estimado era em média de 6% e frequentemente inferior a este valor. Para comparar a produção, entre dois *sites*, é frequente utilizar-se uma medida de acréscimo médio anual (AMA) em volume, expressa em metros cúbicos por hectare e por ano ($m^3/ha/ano$).

Entre outros métodos, o volume da madeira por hectare pode ser determinado tendo por base funções matemáticas que são desenvolvidas para tipos de floresta e regiões em particular. Estas funções permitem determinar o volume diretamente a partir da altura dominante (h_{dom}) e da área basal (G). São exemplos de aplicação destas funções a povoamentos de *Eucaliptus globulus* no território nacional, os modelos Globulus 2.1 e Globulus 3.0, desenvolvidos pela equipa de investigação da Professora Margarida Tomé [41].

A altura média das árvores de um povoamento é uma medida útil para avaliar a condição de crescimento de um povoamento. No entanto, é mais frequente a utilização da altura dominante (h_{dom}) para representar a altura de um povoamento florestal. A altura dominante é a média da altura das árvores com maior diâmetro num povoamento. Porque as árvores, individualmente, num povoamento, competem entre si por recursos do local (luz, água e nutrientes do solo), os seus tamanhos diferem. As mais competitivas tornam-se as mais grossas pela supressão das árvores mais pequenas, que eventualmente morrem. Assim, são as características das árvores mais competitivas e bem sucedidas que refletem a capacidade produtiva do local [47]. A altura dominante é utilizada como medida indireta da produtividade, estando esta fortemente correlacionada com o AMA em volume [37]. A área basal é o somatório das secções das árvores avaliadas a cerca de 1.3m do solo (diâmetro à altura do peito - *dap*) e é expressa em metros cúbicos por hectare (m^2/ha).

A produtividade de um local é desejavelmente quantificada através de índices. No presente contexto abordaremos apenas o *Site index* por ser o mais amplamente utilizado e reconhecido no âmbito da gestão florestal. Para utilizar a h_{dom} como índice de qualidade de um *site* existem normas adotadas internacionalmente. Especificamente, interessa considerar a h_{dom} a uma idade de referência, a qual é tomada para a espécie *Eucalyptus globulus* aos 10 anos. A variável que traduz a h_{dom} aos 10 anos de idade é designada por *Site index* ([37]) e será abreviada por S neste contexto.

A altura dominante é pouco influenciada pelas condições silviculturais, ao contrário da área basal. A biomassa de madeira depende muito da densidade (número de árvores por hectare). Se a densidade é baixa, podem não estar a utilizar-se na totalidade os

recursos disponíveis, e a biomassa resultante será igualmente baixa; no entanto, muitos investigadores demonstraram que a altura das árvores não é afetada significativamente. As alturas das árvores refletem a capacidade produtiva do local, mesmo quando a densidade é baixa e a biomassa produzida não a reflete [47].

Estas considerações conduziram a que os especialistas procurassem estudar a h_{dom} em função de indicadores/variáveis ambientais, tais como: a precipitação, a temperatura, as características de solo, etc. O estudo descrito neste trabalho procura dar algum contributo nesse sentido, para o caso particular da *Eucaliptus globulus* em Portugal.

Nas regiões mediterrânicas vários são os fatores que afetam o desempenho das plantações florestais com *Eucaliptus globulus*. Estes incluem o clima (precipitação, evaporação), o volume útil de solo (estimado através da profundidade e pedregosidade), a fertilidade do solo e a densidade de plantas por unidade de área. Todos estes fatores apontam para um fator crítico que é a disponibilidade de água para a planta, verificando-se que o crescimento aumenta com o aumento da precipitação e com o decréscimo de evaporação. Outros fatores, que se sabe serem influentes, são o declive do terreno e a densidade de plantas, os quais se tornam importantes quando a precipitação diminui e a evaporação aumenta. A *Eucaliptus globulus* responde positivamente a variáveis relacionadas com água, tais como: índice de umidade (com maior influência), precipitação média anual e número de dias com precipitação [6]. Muitos atributos importantes para o desempenho de plantações da *Eucaliptus globulus*, tais como, a profundidade do solo, a fertilidade, etc, não são recolhidos à escala e densidade adequada pelo que surgem pouco ou nada relacionados com o seu crescimento [20].

O termo *deficit hídrico* é frequentemente utilizado na gestão de povoamentos florestais para expressar condição de crescimento das plantas em situação de limitações da disponibilidade de água.

1.2 Motivação e Objetivos Gerais

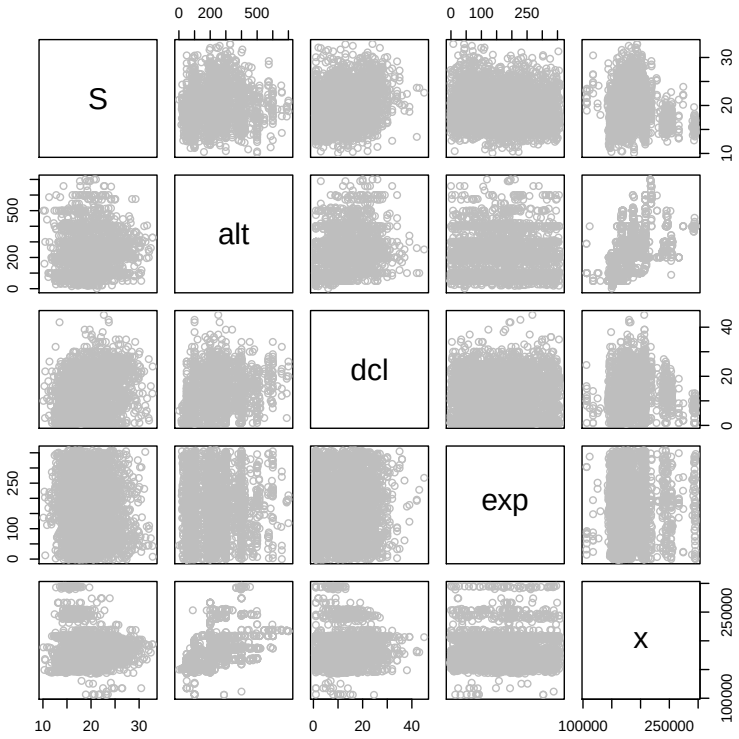
Para avaliar a adequação de um local em relação a um investimento florestal, usam-se estimativas de variáveis biométricas das árvores e dos povoamentos florestais, tais como o Acréscimo Médio Anual (AMA) em volume do local ou, com recurso a índices, tais como o *Site index*, calculadas a partir de observações, caso existam já dados desse local. Quando não existem dados, a avaliação da adequação de um local é efetuada com base

em estimativas conhecidas, obtidas em locais com características ambientais e físicas idênticas, ou seja, com qualidade semelhante. Por vezes não existem dados correspondentes a locais com indicadores de qualidade que assumam valores idênticos aos observados no local pretendido. Daí ser importante que se saiba como usar as variáveis ambientais para predizer uma variável biométrica que seja determinante no cálculo da produtividade. Ou seja, para avaliar a produtividade de um local, é de grande importância encontrar bons estimadores para predizer uma variável biométrica que seja determinante no volume de madeira, em função das características de qualidade do *site*.

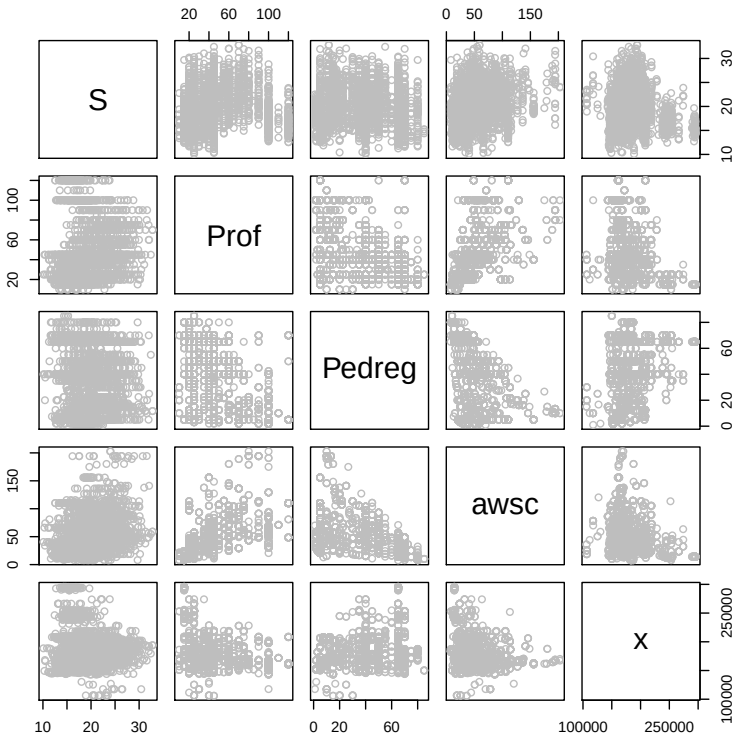
Uma vez que se pretende conhecer a distribuição de uma variável contínua, neste caso o *Site index*, como função de um conjunto de outras variáveis mensuráveis e contínuas, o relacionamento em causa poderá ser modelado por um modelo de regressão linear, modelação essa que tem sido referida preferencialmente na literatura e que é facilmente interpretado pelos utilizadores. A partir de conhecimentos florestais e empíricos é de esperar que o *Site index* se mostre claramente dependente de um conjunto de caraterísticas específicas. No entanto, e antecipando os resultados de uma análise preliminar de dados, a dependência de S em relação a cada um dos diversos regressores não mostrou indícios de linearidade. Tal constatação não confirma o conhecimento existente da dependência do S em relação a alguns regressores, pelo que, deve ser expressa por relações não lineares (polinomiais ou outras).

As figuras 1.1(a), 1.1(b) e 1.2(a), 1.2(b) apresentam matrizes de gráficos de dispersão que ilustram as afirmações anteriores. Na figura 1.1(a) apresenta-se a dispersão de S com variáveis regressoras relacionadas com a topografia e com a variável longitude. Na figura 1.1(b) apresenta-se a dispersão de S com variáveis regressoras relacionadas com o solo. Nas figuras 1.2(a) e 1.2(b) apresenta-se a dispersão de S com variáveis regressoras relacionadas com o clima e incluíram-se ainda as variáveis altitude e latitude por ser frequente o correlacionamento das variáveis climáticas com estas.

As considerações referidas levam a destacar alguns aspetos que foram determinantes na orientação do trabalho: há grupos de variáveis que parecem estar fortemente correlacionadas e que, pela sua natureza, apontam para o uso de Análise Fatorial, nomeadamente, para a definição de alguns fatores como clima, características de solo, etc. Não é evidente a dependência esperada do S em relação a certas variáveis, o que pode ser devido à forma da dependência. Por outro lado, entrando com conhecimentos específicos do problema, a

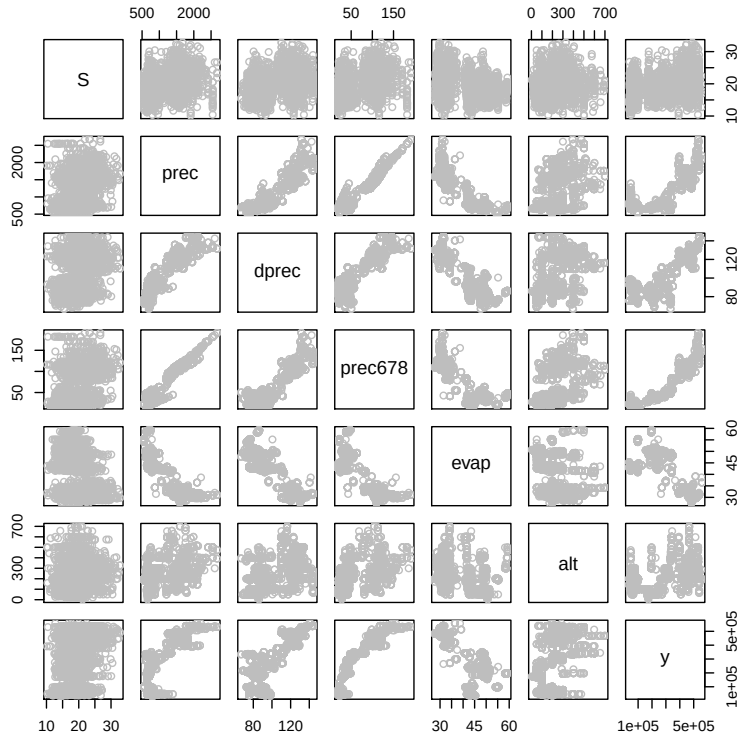


(a) Variáveis relacionadas com a topografia.

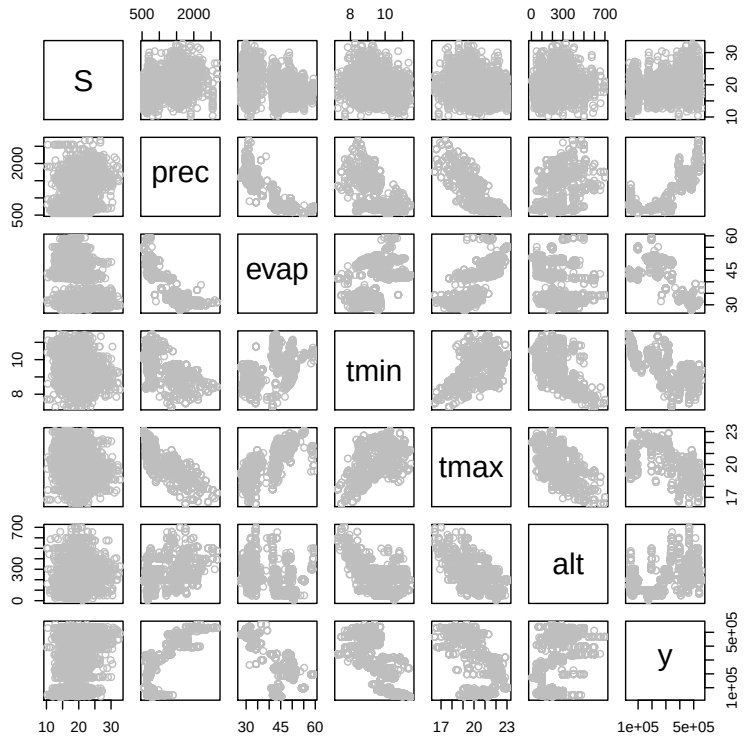


(b) Variáveis relacionadas com o solo.

Figura 1.1: Matrizes de diagramas de dispersão do *Site index* relativamente a algumas variáveis que descrevem em (a) a topografia e em (b) o solo.



(a) Variáveis relacionadas com o clima.



(b) Variáveis relacionadas com o clima.

Figura 1.2: Matrizes de diagramas de dispersão do *Site index* relativamente a algumas variáveis que descrevem em o clima (precipitação, temperaturas e variáveis com estas correlacionadas - altitude e latitude).

análise de alguns gráficos de dispersão aponta para a existência de dois grupos de observações. É o caso dos gráficos que descrevem a dependência de S relativamente à precipitação média anual dos meses mais quentes ($prec678$) ou relativamente ao número de dias anual com precipitação ($dprec$) na figura 1.2(a), que evidenciam a relação com variáveis com caraterísticas climáticas. Assim, decidiu-se continuar a investigação com a seguinte linha de trabalho:

1. Efectuar uma Análise de Componentes Principais (PCA) sobre o conjunto de todas as variáveis, de modo a investigar a possibilidade de redução do número de variáveis.

2. Considerar as observações agrupadas em 2 grupos e fazer o estudo separado de cada um dos grupos. Para agrupar as observações, utilizar-se-á Análise de Agrupamentos (CA).

3. Deixar de fora uma porção de observações, por ex., 10%, selecionadas ao acaso. A ideia é poder encontrar uma regra discriminante e, nessa altura, estimar o erro de classificação, o que tem que ser feito com uma subamostra não usada nos cálculos anteriores.

4. Fazer análise de regressão múltipla linear por grupo.

Os dados utilizados neste trabalho foram recolhidos em parcelas de inventário florestal de povoamentos comerciais puros e seminais da *Eucalyptus globulus*, com árvores da mesma idade e submetidos ao mesmo tipo de silvicultura. Localizam-se de norte a sul do país, particularmente nas regiões consideradas de aptidão para a produção comercial desta espécie. Do total das parcelas disponíveis, lançadas no período entre os anos 2000 e 2010, foram selecionadas 3358 parcelas representativas destes povoamentos florestais, bem adaptados às condições edafo-climáticas em presença e com idades compreendidas entre os 8.5 e os 13 anos.

Não existindo número suficiente de parcelas onde as medições se realizaram na idade de referência (aos 10 anos), utilizou-se um modelo de crescimento desenvolvido por M.Tomé [41] para estimar o *Site index* a partir da h_{dom}.

Na seleção das variáveis ambientais consideraram-se dois aspetos: por um lado, variáveis que contribuem para a disponibilidade de água às plantas; por outro, que a medição das variáveis seja de fácil acesso e recolha, de modo a que o modelo a desenvolver seja o mais operacional possível. Na seleção das variáveis relacionadas com o solo procuraram-se aquelas que poderão refletir a capacidade do solo para reter e disponibilizar água às

plantas. As variáveis relacionadas com o solo foram recolhidas a partir de processo de zonamento edáfico das Unidades de Gestão (RAIZ 2009, relatório). Foram ainda utilizadas variáveis que caraterizam a topografia do local e ainda, as variáveis longitude e latitude.

Todos os métodos estatísticos se baseiam no estabelecimento de pressupostos. Os mais utilizado são: o pressuposto de independência das observações e o de que as observações têm distribuição *Normal* (*Gaussian*). Estes pressupostos têm sido a base de trabalho para todos os métodos de regressão, análise de variância e análise multivariada clássicos. É no entanto frequente na prática, a observações não serem independentes e que assumir um modelo de distribuição *Normal* apenas descreve a maioria das observações, sendo que, algumas observações caem num diferente padrão ou não apresentam nenhum padrão. Uma pequena proporção destas observações, designadas por atípicas (*outliers*), mesmo apenas uma, podem distorcer os resultados dos métodos clássicos. A modelação estatística robusta procura obter métodos que produzam estimativas confiáveis dos parâmetros e associados intervalos de confiança, não apenas quando os dados seguem exatamente uma dada distribuição, mas também quando isto só acontece aproximadamente [29].

Em termos computacionais, a grande vantagem dos métodos clássicos é que requerem apenas processos que se baseiam em métodos de álgebra linear numérica bem estabelecidos, que são genericamente algoritmos muito rápidos. Por outro lado, processar estimativas robustas requer resolver problemas de otimização, frequentemente não lineares, que tipicamente envolvem um significativo aumento de complexidade computacional. A aplicação dos métodos robustos mais correntes seria impensável sem o atual poder dos computadores pessoais comuns, cada vez mais rápidos, com mais memória e mais baratos, o que é bom para a aplicação de estatística robusta [29].

Relativamente aos aspetos computacionais há ainda que considerar a escolha do software. Para a realização deste trabalho escolheu-se a linguagem e ambiente de computação estatística R [38]. O programa R está disponível gratuitamente na internet sobre uma *General Public License(GPL)*. É uma ferramenta de estatística poderosa e flexível, capaz de processar modelos complexos e suportados em enormes conjuntos de dados, permitindo ao utilizador seguir exatamente o que está ser calculado, possuindo ainda excelentes facilidades gráficas. É uma ferramenta em forte expansão, potenciada pela enorme comunidade de utilizadores em todo o mundo, estando a tornar-se a ferramenta mais utilizada por estatísticos, bioinformáticos e campos relacionados. Uma das suas caraterísticas mais

importante é o sistema de pacotes (*package system*) que permite aos utilizadores desenvolver software com aplicações específicas a determinado campo de atividade e partilhar com outros utilizadores suportado por manuais e exemplos de aplicação [42].

O estudo efetuado encontra-se organizado em mais três capítulos, para além da presente introdução. No Capítulo 2 sumarizam-se os métodos estatísticos que foram utilizados no estudo. Este capítulo é constituindo por cinco secções: as duas primeiras secções apresentam noções gerais de estatística multivariada, nomeadamente, de Análise de Componentes Principais (PCA) e de Análise de Agrupamentos. A metodologia da PCA foi útil para a redução do número de parâmetros edafo-climáticos, não só com interesse para a redução da dimensão do conjunto de possíveis regressores, mas também por proporcionar algum conhecimento adicional comprovar relações entre diversas características. A Análise de Agrupamentos (CA) foi usada para validar a existência de dois grupos de observações que posteriormente se verificou corresponderem a localizações geográficas já empiricamente identificadas com diferentes níveis de produtividade. A Análise de Agrupamentos serviu ainda para investigar eventuais vantagens na suposição de um maior número de grupos. As secções 2.4 e 2.5 são dedicadas à apresentação de Modelos de Regressão - a secção 2.4 em termos gerais do modelo e a secção 2.5 direcionada para a Regressão em Componentes Principais. A secção 2.6 foi dedicada à abordagem robusta dos métodos anteriormente descritos. O Capítulo 3 é destinado ao tratamento dos dados. Aplicam-se os métodos referidos no Capítulo 2 e analisam-se os resultados. Finalmente o Capítulo 4 é constituido por uma série de comentários conclusivos sobre o trabalho efetuado.

Capítulo 2

Revisão da Metodologia Estatística

No presente capítulo faz-se uma breve revisão das noções, métodos e resultados da análise estatística que serviram de suporte às técnicas utilizadas no seguimento do trabalho.
A exposição destina-se apenas a assegurar a compreensão dos métodos e dos resultados usados no capítulo seguinte, de modo a facilitar a leitura deste trabalho por utilizadores da estatística sem conhecimentos avançados na matéria.
Assim, o capítulo contém seis secções, onde se começa por relembrar alguns procedimentos de aplicação geral (como os testes de ajustamento e as transformações de Box-Cox), seguindo-se introduções à Análise em Componentes Principais e à Análise de Agrupamentos, uma revisão da Regressão Linear múltipla, a explicação das ideias básicas na Regressão em Componentes Principais e, finalmente, uma breve introdução à Robustez Estatística.

2.1 Metodologias de Pré-Processamento dos dados

Apresentam-se, nesta secção, um conjunto de metodologias utilizadas na verificação dos pressupostos de aplicação de técnicas estatísticas.

Testes de ajustamento à distribuição *Normal* univariada

Um dos pressupostos de aplicação de muitos métodos estatísticos é o de que os dados provêm de uma população com distribuição *Normal*. Para testar o pressuposto de normalidade dos dados podem ser efetuados vários testes formais de ajustamento. No trabalho foram executados dois tipos de testes de ajustamento, nomeadamente: o teste

de *Shapiro-Wilk* e o teste de *Lilliefors K-S*, este último baseado no teste de *Kolmogorov-Smirnov* (KS) . A escolha destes dois testes deveu-se a: por um lado, devido ao facto de serem referidos como dando bons resultados, especialmente o teste de *Shapiro-Wilk* [33] e por outro, por estarem bem documentados e programados em funções disponíveis no programa R, utilizado neste trabalho.

Em qualquer dos testes mencionados, pretende-se testar a hipótese inicial

$$H_0\text{: } \textit{a população segue uma distribuição Normal}$$

contra a hipótese alternativa

$$H_1\text{: } \textit{a população não segue uma distribuição Normal}$$

A hipótese inicial deve ser rejeitada para valores reduzidos do *p-value*, pois o *p-value* representa a probabilidade da estatística do teste tomar o valor que é observado na amostra, ou qualquer outro que seja ainda mais desfavorável a H_0, supondo que esta hipótese é verdadeira.
Se o teste for feito a um nível de significância α, $(0 < \alpha < 1)$, isso significa que

$$\alpha = P[\text{rejeitar}H_0|H_0\text{verdadeira}] = P[\text{região de rejeição}|H_0\text{verdadeira}].$$

Quando se decide o valor de α (em geral entre 1% e 10%), há motivos para rejeitar H_0 se *p-value* $\leq \alpha$, pois isso significa que o valor observado da estatística do teste (na amostra) pertence à região crítica.

O teste de ***Shapiro-Wilk*** foi proposto originalmente em 1965 e era recomendado para amostras de pequena dimensão (tamanho inferior a 50).
Dada uma amostra aleatória e as suas componentes ordenadas, $Y_1 < Y_2 < Y_3 < \cdots < Y_n$ a estatística do teste original foi definida como

$$W = \frac{(\sum_{i=1}^{n} a_i Y_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2.1}$$

onde Y_i é a *i-ésima* estatística de ordem, $\bar{Y}$ é a média amostral dos Y_i, $\mathbf{a}_i = (a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$, $\mathbf{m} = (m_1, \ldots, m_n)^T$ são os valores esperados da estatística de ordem quando a população segue uma distribuição *Normal* e $\mathbf{V}$ é a matriz de covariâncias da

estatística de ordem. O valor de W varia entre 0 e 1 e valores pequenos conduzem à rejeição da normalidade.

O teste de Shapiro-Wilk foi inicialmente modificado por Royston em 1982 [33], de modo a alargar a sua utilização a amostras de dimensões superiores. Em 1999 Royston forneceu um novo algoritmo que tem vindo a ser utilizado, com bons resultados, para amostras com dimensão $3 \leq n \leq 5000$.

O teste de ***Lilliefors K-S*** é uma modificação do de ***Kolmogorov-Smirnov* - KS**. Este último é apropriado quando os parâmetros da distribuição hipotética são totalmente conhecidos. Contudo, frequentemente os parâmetros da distribuição são desconhecidos. Neste caso, os parâmetros têm de ser estimados com base no próprio conjunto de dados [33] com que é efetuado o teste, o que obrigou a recalcular a distribuição da estatística do teste. Ao teste KS assim adaptado dá-se o nome de ***Lilliefors K-S***.

A estatística de *Kolmogorov-Smirnov* baseia-se na maior diferença vertical entre a distribuição hipotética e a empírica. Dados n pontos ordenados $x_1 < x_2 < x_3 < \cdots < x_n$, a estatística do teste é definida por

$$T = sup_x |F^*(x) - F_n(x)| \tag{2.2}$$

onde *sup* significa o supremo, F^* representa a função distribuição de probabilidades da distribuição hipotética e F_n é a função distribuição empírica, estimada com base na amostra a usar no teste. No teste de KS, $F^*(x)$ representa a função distribuição da distribuição *Normal*, com média μ e desvio padrão σ.
Rejeita-se H_0 quando a distribuição da amostra se afasta da distribuição hipotética, i.e., para valores elevados de $|F^*(x) - F_n(x)|$. Portanto, se o valor de T exceder o percentil $1 - \alpha$ da distribuição da estatística, então há motivos para rejeitar H_0 ao nível de significância α [33].

Transformação de Variáveis (*Box-Cox Transformation*)

Quando os dados não estão de acordo com o pressuposto de distribuição Normal, é possível tomar medidas adequadas para atenuar ou contornar o problema, transformando-

os de modo a que os dados transformados tenham uma distribuição próxima da *Normal*. Um de entre vários processos de transformação para a *Normal* consiste em aplicar a *transformação potência de Box-Cox*, desenvolvida pelos estatísticos George Box e David Cox em 1964: [44].

$$x_i^{\lambda} = \begin{cases} (x^{\lambda} - 1)/\lambda & \text{para} \quad \lambda \neq 0 \\ \log x & \text{para} \quad \lambda = 0 \end{cases} \tag{2.3}$$

O procedimento consiste em identificar um expoente λ mais apropriado para transformar a distribuição dos dados numa forma *Normal*. O valor de λ indica a potência à qual os dados de uma variável devem ser elevados. Mais especificamente, o algoritmo procura um valor λ entre -5 e 5 até o melhor valor ser encontrado. Representando por x os valores da variável, para $\lambda = 0$ a transformação é o logaritmo de x. As transformações mais frequentemente utilizadas são: $x^{-2} = 1/x^2$, $x^{-1} = 1/x$, $x^{-0.5} = 1/\sqrt{(x)}$, $x^0 = log(x)$, $x^{0.5} = \sqrt{(x)}$, $x^1 = x$, $x^2 = x^2$.
A transformação apenas se usa quando os dados são positivos e diferentes de zero. Isto contudo pode ser obtido adicionando um valor constante C a todos os dados, de forma a que todos se tornem positivos antes de serem transformados, ou seja, transformando-os de x em $x^* = (x + C)^{\lambda}$.

Estandardização de Variáveis

Frequentemente, as variáveis originais $X_1, X_2, \ldots, X_p$ são medidas em escalas ou unidades diferentes, o que conduz a grandes discrepâncias entre as variâncias. Deste modo, surge a necessidade de se estabelecer uma certa uniformização dos dados, o que se consegue através da estandardização (ou padronização) das variáveis.

$$Z_j = \frac{X_j - \mu_j}{\sqrt{\sigma_j}}, \quad j = 1, ..., p,$$

onde μ_j e σ_j representam, respetivamente, a média e a variância de X_j.

Podem, para o mesmo efeito, ser utilizadas outras medidas de tendência central e dispersão, nomeadamente, a mediana ($\tilde{x}$) e o desvio absoluto mediano (MAD) .

$$Z_j = \frac{X_j - \tilde{x}_j}{\sqrt{MAD_j}}, \quad j = 1, ..., p,$$

As variáveis $Z_j, \quad (j = 1, 2, \ldots, p)$ têm valor médio nulo e variância unitária.

A matriz de covariâncias das variáveis Z_j é igual à matriz de correlações das variáveis X_j, isto é,

$$Cov(Z_i, Z_j) = \frac{Cov(X_i, X_j)}{\sqrt{\sigma_i}\sqrt{\sigma_j}} = Corr(X_i, X_j). \tag{2.4}$$

Em termos geométricos, a estandardização dos dados equivale a uma translação do centro de gravidade da núvem (o ponto constituído pelo valores médios das variáveis) para a origem do referencial, e cada eixo, de acordo com o valor do desvio padrão da variável correspondente, será estendido (se $\sqrt{\sigma}_j < 1$) ou contraído (se $\sqrt{\sigma}_j > 1$), com fatores de alteração das escalas diferenciados para cada eixo.

Em termos de amostras multivariadas, a estandardização processa-se de acordo com o mesmo princípio, mas agora subtraindo a cada *i-ésima* observação da *j-ésima* variável, a média amostral $\bar{x_j}$ das observações dessa variável e dividindo a diferença pelo desvio padrão das observações da mesma variável, ou seja, as novas observações estandardizadas são da forma

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j}, \quad i = 1, \ldots, n; \; j = 1, \ldots, p.$$

Com a estandardização dos dados elimina-se o problema da escala de medida das variáveis [25].

2.2 Análise de Componentes Principais

A análise de componentes principais (PCA) é uma técnica estatística de análise multivariada, utilizada frequentemente para revelar relações entre variáveis que uma análise preliminar de dados não permite fazer. O objetivo da PCA é encontrar uma transformação ortogonal das variáveis originais, que defina um novo conjunto de variáveis, não correlacionadas entre si e que possam explicar a maior parte possível da variabilidade dos dados. Do ponto de vista da aplicação do método, usa-se a PCA para encontrar um novo conjunto de variáveis não correlacionadas, preferencialmente, de dimensão inferior ao conjunto de variáveis originais e que, ainda assim, expliquem a maior proporção da variabilidade. Às variáveis ortogonais dá-se o nome de Componentes Principais (*Principal Components* - PC). Na sua determinação, as PC são ordenadas por ordem decrescente de variância, de forma a que a primeira PC explica a maior fração da variância total dos dados. Para além da sua utilização na análise exploratória de dados, a PCA é também uma ferramenta poderosa, em combinação com outros métodos, nomeadamente, de regressão linear, na construção de modelos preditivos (assunto a abordar na secção 2.4 deste capítulo). Note-se que a metodologia tem vantagens quando as variáveis originais são correlacionadas, mas não acrescentaria nada de novo se as variáveis originais já não fossem correlacionadas. A técnica da PCA foi inicialmente descrita por *Karl Pearson* em 1901 e posteriormente consolidada por *Hotteling* em 1931.

As Componentes Principais (*Principal Components* - PC)

Seja $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ o vetor das variáveis originais que foram observadas na amostra de dimensão n, seja $\mathbf{\Sigma}$ a sua matriz de covariâncias (com $det(\mathbf{\Sigma}) > 0$) e sejam $\lambda_1 > \ldots, \lambda_p$ os valores próprios de $\mathbf{\Sigma}$, ordenados por ordem decrescente. As componentes principais são combinações lineares dessas p variáveis, definidas por

$$PC_j = e_{1j}X_1 + e_{2j}X_2 + \cdots + e_{pj}X_p = \mathbf{e}_j^T\mathbf{X} \tag{2.5}$$

onde $j = 1, 2, \ldots, p$ e $\mathbf{e}_j^T = (e_{1j}, e_{2j}, \ldots, e_{pj})$ são vetores de constantes de norma unitária, ou seja, tais que $e_j^T e_j = 1$, para todo o j e que $Corr(PC_i, PC_j) = 0$, para qualquer par (i, j), com $i \neq j, i, j = 1, 2, \ldots, p$.

As componentes principais PC verificam as seguintes propriedades:

1. A variância da *j-ésima* PC é igual ao *j-ésimo* valor próprio da matriz de covariâncias $\mathbf{\Sigma}$, i.e.,
$$Var(PC_j) = \mathbf{e}_j^T \mathbf{\Sigma} \mathbf{e}_j = \lambda_j \ , 1 \leq j \leq p.$$

2. A soma das variâncias da PCs é igual à soma das variâncias das variáveis originais e igual à soma dos valores próprios de $\mathbf{\Sigma}$, i.e.,
$$\sum_{j=1}^{p} Var(PC_j) = \sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \lambda_j = tr(\mathbf{\Sigma}).$$

3. O coeficiente de correlação (de Pearson) entre a variável original X_k e a j-ésima PC é igual
$$\rho_{k,j} = e_{kj} \frac{\sqrt{\lambda_j}}{\sqrt{Var(X_k)}}, \quad k, j = 1, 2, \ldots, p \quad j \neq k.$$

4. $e_j^T e_j = 1, \quad j = 1, 2, \ldots, p$, isto é, o vector e_j^T tem norma unitária.

A decisão sobre o número de componentes principais a considerar depende da percentagem de explicação pelas primeiras k componentes principais e é uma questão subjetiva. Existem, no entanto, critérios práticos empíricos para esse efeito, tais como:

1. Decidir com base na representação gráfica, por ordem decrescente, da percentagem de variação total explicada por cada componente (gráfico do cotovelo);
2. Incluir o número mínimo de componentes que expliquem 85% da variância total.
3. Reter somente aquelas componentes cujas variâncias são maiores do que um.
4. Tomar como última componente aquela cujo valor próprio é igual ou superior à média dos restantes.

Interessa lembrar que as PCs não são invariantes. Quando as variáveis são estandardizadas, i.e., se a PCA não é conduzida com as variáveis originais, mas sim a partir de variáveis
$$Z_j = (X_j - \mu_j)/\sqrt{V(X_j)},$$
(onde $\mu_j = E(X_j)$), a matriz de covariâncias dos Z_j coincide com a matriz de correlações dos X_j. As componentes principais passam a ser
$$PC_j^* = \mathbf{e}_j^T \mathbf{Z} = \mathbf{e}_j^T (\mathbf{\Sigma}_{diag}^{1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu}).$$

A soma das variâncias das PCs mantém-se, mas a proporção da variância total devida à *i-ésima* componente passa a ser calculada por λ_j/p. Nestas condições, o coeficiente de correlação entre a variável estandardizada Z_j e a componente principal PC_j^* é $\rho_{k,j} = e_{kj}\sqrt{\lambda_j}$.

Em termos amostrais, as PCs são estimadas usando as médias, variâncias e correlações amostrais (corrigidas), consoante o caso.

Finalmente, chama-se a atenção para uma questão de terminologia inglesa, que é usada nos programas de cálculo estatístico: os coeficientes lineares que definem as PCs são designados por *loadings* e os coeficientes de cada indivíduo numa PC são designados por *scores*.

2.3 Análise de Agrupamentos

A Análise de Agrupamentos (*Cluster Analysis* - CA), constitui uma tarefa importante da análise exploratória de dados, que consiste em agrupar um conjunto de observações, de forma a que as observações num mesmo grupo sejam mais similares entre elas (grupos homogéneos) do que com as outras noutros grupos (heterogeneidade entre grupos).
O critério em que assenta a decisão de similaridade ou dissimilaridade entre dois indivíduos baseia-se numa medida de proximidade. A medida de proximidade será uma distância, pelo que a similaridade entre dois indivíduos será tanto maior quanto menor for a distância entre eles. Então, a similaridade (ou dissimilaridade) entre dois indivíduos pode ser expressa como uma função de distância entre dois pontos no espaço p-dimensional que os representa. Com base nessa distância, é então calculada a distância de cada ponto a todos os outros pontos, constituindo-se deste modo uma matriz de distâncias, D, designada por *matriz de proximidade*. Esta matriz é uma matriz quadrada de ordem n, simétrica e com todos os elementos da diagonal principal nulos, tendo como termo genérico $d_{i,j}$ - a distância entre o i-ésimo e o j-ésimo indivíduos.
A distância satisfaz as seguintes propriedades:

1. $d_{ij} \geq 0, \quad \forall i, j = 1, \ldots, n;$
2. $d_{ii} = 0, \quad \forall i = 1, \ldots, n;$
3. $d_{ij} = d_{ji}, \quad \forall i, j = 1, \ldots, n;$
4. $d_{ij} \leq d_{ik} + d_{jk}, \quad \forall i, j, k = 1, \ldots, n$

Para avaliar a similaridade usam-se distâncias definidas de um modo mais geral, em que a última propriedade pode não se verificar. A medida de distância mais frequentemente utilizada é a **distância euclidiana** definida por

$$d_{i,j} = \| \mathbf{x}_i - \mathbf{x}_j \| = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{\sum_{k=1}^{p} (\mathbf{x}_i - \mathbf{x}_j)^2}, \tag{2.6}$$

onde $\mathbf{x}_i$ representa o i-ésimo indivíduo.
No entanto, existem outras medidas de distância, como por exemplo, a distância Euclideana Generalizada, que é da forma:

$$d_{i,j} = \| \mathbf{x}_i - \mathbf{x}_j \|_{\mathbf{W}} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} (\mathbf{x}_i - \mathbf{x}_j)},$$

onde W é uma matriz definida positiva.

No caso particular em que $\mathbf{W}$ é a inversa da matriz de covariâncias das variáveis, e que se pretende medir a distância de um indivíduo à média das observações, a distância generalizada dá origem à ***distância de Mahalanobis***. A distância de Mahalanobis do *i-ésimo* indivíduo (à média $\boldsymbol{\mu}$) é definida por

$$d_i = \| \mathbf{x}_i - \boldsymbol{\mu} \|_{\mathbf{M}} = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}.$$

A distância de Mahalanobis é invariante a mudanças de escala nas variáveis.

Em Análise de Agrupamentos existem dois tipos de métodos - os métodos hierárquicos, em que o agrupamento dos indivíduos se desenvolve por etapas - e os métodos não-hierárquicos - em que o número de grupos é pré-fixado. O método que será usado no âmbito deste trabalho é um método não-hierárquico, designado por *K-means*.
De acordo com este método, o processo de formação dos grupos é iterativo, estabelecendo-se à partida o número de grupos (k) pretendidos e uma classificação inicial de todos os indivíduos nos k grupos ou, para cada um dos grupos, um centroide inicial (uma semente). Os centroides iniciais podem ser definidos pelo utilizador ou determinados aleatoriamente. Em cada uma das iterações, e com base numa medida de proximidade, é associado um novo agrupamento de indivíduos a cada um dos centroides determinados na iteração anterior. De seguida, procede-se à atualização desses centroides, com transferências dos indivíduos entre grupos, de modo a que, para cada grupo, se atinja o menor afastamento interno entre os indivíduos que o compõem, e os centroides respetivos. Designa-se por erro interno ou afastamento interno a soma dos quadrados das distâncias dentro dos grupos, somada para todos os grupos

$$E = \sum_{i=1}^{k} \sum_{x} ||c_i - x||^2 = \sum_{i=1}^{k} \sum_{x} d^2_{c_i,x}, \tag{2.7}$$

onde c_i representa o centroide do i-ésimo grupo C_i. O ciclo de iterações termina quando nenhum dos indivíduos muda de grupo, ou seja, quando deixam de ocorrer variações nos centroides. Esta situação corresponde a um mínimo local de E, mas não necessariamente a um mínimo global. Isto acontece porque o processo não vai incidir sobre todos os k agrupamentos possíveis, mas sim apenas sobre aqueles que correspondem aos centroides inicialmente especificados.

Por outro lado, interessa também que os grupos sejam dissemelhantes entre si; ou seja, interessa definir critérios para medir o afastamento entre grupos. Existem diversos critérios, entre os quais:

- O Método do vizinho mais próximo: o afastamento entre dois grupos é medido pela menor distância entre um elemento de um dos grupos e um elemento do outro.

- O Método do vizinho mais distante: o afastamento entre dois grupos é medido pela maior distância entre um elemento de um dos grupos e um elemento do outro.

- O Método das distâncias médias: o afastamento entre dois grupos é medido pela média de todas as distâncias entre um elemento de um dos grupos e um elemento do outro.

- O Método dos centroides: o afastamento entre dois grupos é medido pela distância entre os respetivos centroides.

Para cada definição de dissemelhança vai corresponder um agrupamento diferente dos mesmos dados. Em termos práticos, pode ser conveniente experimentar diferentes distâncias, para ver até que ponto é que o agrupamento é resistente ao critério escolhido.

Análise Discriminante Quadrática (QDA)

Na validação de métodos e de modelos, frequentemente, é reservado um conjunto de dados para avaliar o desempenho das técnicas quando são aplicados a novos dados. Quando, no conjunto de dados, são identificados agrupamentos e, para cada grupo, é produzido um modelo diferente, é desejável a existência de um critério de classificação de novos indivíduos, nomeadamente, das observações reservadas para validação. Para usar o modelo adequado, as novas observações precisam de ser classificadas nos grupos antes estabelecidos.
A Análise Discriminante é uma técnica estatística multivariada que se utiliza para identificar as caraterísticas que distinguem os membros de um grupo, de modo a que, conhecidas as caraterísticas de um novo indivíduo, se possa prever a que grupo pertence.

Uma Análise Discriminante é altamente sensível à presença de observações atípicas, que podem ter um largo impacto nas médias dos grupos e também aumentar as variâncias.

Na Análise Discriminante admite-se que temos duas ou mais populações e interessa conhecer regras que possam separar os indivíduos dessas populações. O critério mais simples e de maior divulgação é o estabelecido pelo discriminante linear de Fisher, para separar dois grupos. Este critério não exige que os grupos sigam distribuições normais, mas pressupõe que ambos os grupos pertençam a populações com a mesma matriz de covariâncias. Nesse caso, a classificação de novos indivíduos é definida com base numa combinação linear das componentes, que resulta na seguinte regra: uma observação $\mathbf{x}_0$ é classificada no Grupo 1 se

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_{pool}^{-1} \mathbf{x}_0 \geq (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_{pool}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2),$$

onde $\bar{\mathbf{x}}_1$ e $\bar{\mathbf{x}}_2$ denotam, respetivamente, as médias do Grupo 1 e do Grupo 2 e S_{pool} é uma estimativa da matriz de covariâncias comum, obtida a partir das observações dos dois grupos; caso contrário, o indivíduo $\mathbf{x}_0$ é classificado no Grupo 2.

Como se referiu, este critério só se aplica quando a matriz de covariâncias é a mesma nas duas populações de onde provêm os grupos. Quando os grupos seguem distribuições com diferentes matrizes de covariâncias, existe um processo alternativo, que estabelece a separação através de uma regra discriminante quadrática, desde que ambas as populações sigam distribuições normais. Nessas condições, sendo $\mathbf{y}_0 = -\frac{1}{2}\mathbf{x}_0^T(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})^T\mathbf{x}_0 + (\bar{\mathbf{x}}_1^T\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2^T\mathbf{S}_2^{-1})\mathbf{x}_0$, o indivíduo $\mathbf{x}_0$ é classificado no Grupo 1, se

$$\bar{\mathbf{y}}_0 \geq \ln(\frac{c_{12}}{c_{21}}\frac{p_2}{p_1}) + k,$$

onde $\mathbf{S}_1$ e $\mathbf{S}_2$ são as estimativas (de máxima verosimilhança) de cada uma das matrizes de covariâncias,

$$k = \frac{1}{2}\ln(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|} + \frac{1}{2}(\bar{\mathbf{x}}_1^T\mathbf{S}_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^T\mathbf{S}_2^{-1}\bar{\mathbf{x}}_2),$$

c_{12} e c_{21} são custo de má classificação (geralmente supostos unitários), e $p_i,\ i = 1, 2$ é a probabilidade da observação pertencer a cada um dos respetivos grupos (geralmente supostas iguais ou estimadas pelas proporções das dimensões dos grupos); caso contrário, o indivíduo é classificado no Grupo 2. Note-se que nas expressões anteriores, as médias amostrais e as matrizes de covariâncias representam os estimadores convencionais das respetivas médias e matrizes de covariâncias das populações.

2.4 Regressão linear

O modelo de regressão tem como objetivo estudar a dependência, em média, de uma ou mais variável(eis) resposta (dependente) em relação a um conjunto de outras variáveis, designadas por regressoras (independentes). Especificamente, a equação de regressão num modelo de regressão linear com uma única variável resposta toma a seguinte forma

$$E[Y] = \beta_0 + \beta_1 x_1 + ... + \beta_r x_p. \tag{2.8}$$

A variável resposta Y é uma variável aleatória contínua, enquanto que os regressores $x_1, x_2, ..., x_p$ são variáveis fixas. O termo *linear* deve-se ao facto da média ser uma função linear dos parâmetros desconhecidos $\beta_0, \beta_1, ..., \beta_p$.

Para o modelo traduzir a relação de dependência em termos das observações individuais, é adicionada à equação uma nova variável aleatória ϵ, designada por *erro* do modelo e que contabiliza erros de medição e efeitos de outras variáveis não explicitamente consideradas. Deste modo, as observações verificam a equação

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_r x_{ir} + \epsilon_i,\ i = 1, \ldots, n,$$

onde é assumido os termos do erro terem as seguintes propriedades:

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2(constante)$
- $Cov(\epsilon_i, \epsilon_k) = 0, i \neq k.$

Em geral, um quarto pressuposto é assumido, para que seja possível realizar testes de hipóteses e declarações sobre a incerteza associada, que é o pressuposto de normalidade na distribuição dos termos do erro ϵ.

Note-se que num modelo da forma referida, é possível incluir termos que representem potências dos regressores, exprimindo-os como novos regressores. Por outro lado, embora geralmente os regressores representem variáveis numéricas, também podem permitir a inclusão de variáveis categóricas. Nesses casos, a presença ou ausência de alguma categoria, que se espera influenciar uma resposta, é considerada definido uma variável que assume apenas os valores 0 e 1: o valor 0 faz com que o coeficiente desapareça da equação de regressão e o valor 1 faz com que o coeficiente atue como um termo constante adicional num

modelo de regressão, representando a presença da categoria pretendida. Essas variáveis tomam o nome de variáveis *mudas* (*dummy variables*).

O modelo pode ser apresentado em notação matricial pela equação

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.9}$$

onde $\mathbf{Y}$ é o vetor resposta e $\mathbf{X}$ é a matriz dos valores que as variáveis regressoras assumem. Os pressupostos anteriores sobre os erros exprimem-se por $E(\boldsymbol{\epsilon}) = \mathbf{0}$ e $Cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

O método mais comum para estimar os coeficientes de regressão é o método dos mínimos quadrados. Como o nome indica, este método seleciona o estimador que minimiza o quadrado das distâncias entre os valores observados da variável resposta e aqueles que são preditos pelo modelo; ou seja, minimiza

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2. \tag{2.10}$$

Admitindo que a matriz $\mathbf{X}^T\mathbf{X}$ tem determinante não nulo, o estimador de $\boldsymbol{\beta}$ obtido pelo método dos mínimos quadrados é da forma

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

onde $\mathbf{X}^T$ indica a matriz transposta de $\mathbf{X}$. As diferenças

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_p x_{ip},\ i = 1, 2, ..., p, \tag{2.11}$$

designam-se por resíduos. O vetor dos resíduos $\hat{\epsilon} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ contém informação relativa ao parâmetro desconhecido σ^2, uma vez que $Var(\epsilon_i) = \sigma^2, \forall i = 1, \ldots, n$. Assim, σ^2 pode ser estimado pela variância amostral dos resíduos, i.e., por

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2.$$

Os resíduos assumem um papel fundamental na análise da regressão, uma vez que se usam na validação dos pressupostos do modelo. Para além disso, também ajudam a

interpretar a qualidade do ajustamento do modelo, através da decomposição básica da soma dos quadrados em torno da média:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2 \tag{2.12}$$

A igualdade anterior estabelece que a variabilidade total tem duas componentes - uma que é devida ao modelo com os regressores considerados (SSR) e outra que é devida à soma dos quadrados dos resíduos (aos erros do modelo) (SSE). Deste modo, quando o ajustamento do modelo é bom, a parte da variabilidade total (SST) que é explicada por SSR deve ser grande em relação à que se deve aos erros do modelo (SSE). Tomando essa comparação, a qualidade do ajustamento do modelo pode ser avaliada pelo *coeficiente de determinação*

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Este coeficiente varia entre 0 e 100% e quanto mais próximo estiver de 1 melhor, pois significa que os resíduos são pequenos e que a variância das observações está a ser bem explicada pelo modelo. Por isso, o coeficiente de determinação é um dos critérios mais usados na escolha dos modelos.
Importa referir que, quando a um modelo com um determinado conjunto de regressores se adicionam novos regressores, o valor de R^2 pode vir aumentado simplesmente devido à presença dos novos regressores, sem que isso signifique que o ajustamento é melhor. Por esse motivo, ao comparar dois modelos nessas condições, é preferível usar o coeficiente de determinação ajustado, que é definido a partir do anterior por

$$R^2_{adj} = 1 - \frac{SSE}{SST}\frac{n-1}{n-p-1},$$

mas que já contém uma correção relativa ao número de regressores no modelo.

A questão da opção por modelos com mais ou menos regressores, escolhidos dentro de um conjunto de p regressores possíveis, diz respeito à seleção de variáveis em modelos encaixados. Interessa escolher modelos de regressão que contenham o menor número possível de regressores, mas que contenham todos aqueles que são estatisticamente significativos; e que, por outro lado, não contenham regressores que estejam correlacionados entre si.

Um dos métodos mais vulgares para fazer a seleção de regressores é o da regressão passo-a-passo (*stepwise*): vai-se testando a sucessiva adição e remoção de variáveis através de um critério de comparação de modelos. O modo de procura da melhor regressão pode ser por seleção *forward*, por eliminação *backward* ou por eliminação bidirecional (*both*). Na seleção *forward* começa o processo sem variáveis e testa a adição sucessiva das variáveis; na eliminação *backward* o processo começa com todas as variáveis e vai testando a eliminação sucessiva das variáveis uma a uma; na eliminação bidirecional combina os dois procedimentos anteriores, testando em cada passo a inclusão ou eliminação de cada variável no modelo.

Um dos critérios usados na seleção de variáveis é o do erro quadrático médio dos resíduos. Um outro critério que também é frequentemente usado, e que não é baseado na decomposição da variabilidade, é o Critério de Informação de Akaike (AIC) (*Akaike Information Criterion*), referido, por exemplo em [48] , como apresentando vantagens relativamente a outros critérios frequentemente utilizados. Para cada modelo do conjunto de modelos possíveis, calcula-se

$$AIC = 2k - 2ln(L),$$

onde k é o número de parâmetros do modelo e L é o máximo valor da função de verosimilhança da amostra para o modelo estimado. De acordo com este critério, é preferível optar pelo modelo que apresentar o menor AIC.

2.5 Regressão em Componentes Principais

Quando se colocam situações em que, à partida, existem muitos regressores, uma das maiores dificuldades na implementação de modelos de regressão linear é a dos regressores apresentarem correlação linear muito elevada entre si. Existem diversas propostas de metodologia para ultrapassar essa dificuldade, sendo a mais tradicional a seleção de variáveis, já referido na secção anterior.

Ora a decomposição da variabilidade das observações, que leva à inclusão ou à exclusão de um certo conjunto de regressores, é decidida num cenário ideal de não correlacionamento entre regressores. Existindo correlacionamento linear considerável entre os regressores, a situação aproxima-se da multicolinearidade entre as variáveis e o determinante da matriz $\mathbf{X}^T\mathbf{X}$ passa a ser muito próximo de zero; qualquer pequena alteração nas observações pode ser ampliada pela correlação e conduzir a instabilidade numérica e a diferentes soluções na seleção dos regressores. De facto, as técnicas de seleção das variáveis só estão teoricamente garantidas se os regressores forem não correlacionados entre si.

Ao escolher o conjunto de regressores através de uma análise prévia de componentes principais, esse problema é ultrapassado, ficando asseguradas as condições ideais para ser analisada a decomposição da variabilidade. Assim, uma forma de contornar este problema é a aplicação da Regressão Linear em Componentes Principais (*Principal Component Regression PCR*).

De um modo geral, a metodologia consiste em aplicar a Análise de Componentes Principais (PCA) para determinar as Componentes Principais (PC) a partir de um conjunto de variáveis originais; seguida da aplicação da regressão linear múltipla, tendo por regressores, não os regressores iniciais, mas sim as PCs determinadas. Estas PCs são ortogonais entre si e como tal, são não correlacionadas.

Assim, seja $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ o vetor dos regressores originais. As componentes principais são combinações lineares desses p regressores, que, eventualmente, eram correlacionados entre si. Ao considerar as primeiras k componentes principais como regressores, fica desde logo assegurado que essas variáveis explicam a maior parte da variabilidade na amostra; e, assim, os regressores originais continuam a contribuir para explicar a variabilidade da variável resposta, mas só indiretamente (não explicitamente na equação de regressão).

Deste modo, decide-se o número k de PCs a reter e, de seguida, passa a aplicar-se a análise da regressão linear múltipla a um modelo da forma

$$y_i = \beta_0 + \beta_1 \times PC_{i1} + \beta_2 \times PC_{i2} + \cdots + \beta_k \times PC_{ik} + \epsilon_i \tag{2.13}$$

onde a *i-ésima* observação da *j-ésima* componente principal ($j = 1, 2, \ldots, k$) depende das observações originais por

$$PC_{ij} = e_{1j}x_{i1} + e_{2j}x_{i2} + \cdots + e_{pj}x_{ip}. \tag{2.14}$$

Logo, usando regressão em componentes principais, o modelo de regressão pode ser muito simplificado, reduzindo fortemente o número de regressores, entrando indiretamente com todas as variáveis originais significativas e assegurando, simultaneamente o não correlacionamento desejável entre regressores.

2.6 Abordagem robusta

2.6.1 Conceitos importantes

Os modelos e os métodos estatísticos baseiam-se, usualmente, na definição dum conjunto de pressupostos, tais como a distribuição normal dos dados, ou a independência das observações. Contudo, na prática, estes pressupostos podem não se verificar. É o que acontece, por exemplo, se os dados são hipoteticamente provenientes de populações normais e na realidade apresentam distribuições com caudas mais pesadas; ou quando os dados incluem observações atípicas, geralmente designadas por *outliers*, o que é muito frequente ocorrer, por vezes apenas devido a erros na introdução dos dados.
Nalguns casos, basta uma única observação atípica para que as estatísticas calculadas assumam valores completamente diferentes do esperado e até disparatados. A situação mais simples para exemplificar o prejuízo causado por um único *outlier*, é imaginar o efeito que ele pode ter sobre o valor da média amostral ou da variância amostral. Por exemplo, numa amostra constituída pelos valores (1, 2, 1, 2, 1), a mediana é 1, a média é 1.4 e o desvio padrão é 0.55; se, involuntariamente (p.ex., devido ao teclado de introdução dos dados), tiver sido registada a amostra (1, 2, 1, 2, 111) apenas uma das componentes mudou, no entanto o valor das estatísticas anteriores passaram a ser, respetivamente, de 2 para a mediana, 23.4 para média e de 48.98 para o desvio padrão! Ou seja, bastou alterar um só valor na amostra, para que os valores de estatísticas como a média ou o desvio padrão deixem de fazer sentido.

De facto, os métodos estatísticos mais divulgados têm boas propriedades quando se aplicam nas condições previstas (ao longo trabalho, a aplicação de cada método nesse cenário ideal será designado por método convencional ou na versão convencional); mas podem ser desastrosos quando há pequenos desvios dos pressupostos - no exemplo anterior, foi o caso da média ou do desvio padrão, mas não da mediana. Por isso, é importante utilizar técnicas que sejam robustas, no sentido em que, se existirem pequenos afastamentos das condições do modelo, esses métodos não produzam resultados catastróficos.
Este conceito de robustez inclui a noção de resistência - quando o que está em causa é a sensibilidade a alterações no valor numérico das estatísticas (como no exemplo dado), mas é mais amplo, de forma a abranger outro tipo de dificuldades (como no exemplo referido de distribuições com caudas mais pesadas).

Os problemas associados à falta de robustez das estatísticas agravam-se muito quando

os dados são multi-dimensionais; nesses casos, é difícil ou impossível detetar observações atípicas com técnicas exploratórias de dados ou através da verificação dos pressupostos, e os prejuízos podem ser enormes.

A análise estatística robusta fornece soluções para contornar os problemas acima referidos, atenuando os efeitos que possam ser devidos a afastamentos dos pressupostos, incluindo a presença de *outliers*. O objetivo dos métodos estatísticos robustos é dar resultados confiáveis mesmo quando os pressupostos assumidos para a análise estatística tradicional não são preenchidos. Os métodos robustos também podem ser utilizados na própria deteção e confirmação de *outliers*.

No seguimento do trabalho, aplicaram-se sempre os métodos de análise estatística no cenário pressuposto convencional e procuraram-se métodos alternativos robustos (o que se designou por versão robusta dos métodos).

As alternativas robustas têm vindo a ser teoricamente desenvolvidas nas últimas décadas e estão já inseridas em muitos dos programas de computação estatística. As principais desvantagens na sua utilização são a interpretação, por vezes pouco intuitiva, e o custo computacional, uma vez que é frequente os métodos robustos recorrerem a cálculos que apenas são viáveis com o auxílio do computador. Mas esta principal desvantagem está atualmente a ser ultrapassada, graças à evolução das capacidades de cálculo e à vulgarização crescente de programas estatísticos.

Assim, existem uma série de noções que permitem caraterizar a robustez dum estimador estatístico ou de um método estatístico. Deste modo, passam a apresentar-se algumas conceitos importantes neste contexto, como é o caso do ponto de rotura (*breakdown point*) ou da função influência. Estas noções são mais facilmente transmitidas quando aplicadas a estimadores (ou seja, a estatísticas destinadas a produzir estimativas) de parâmetros unidimensionais e de populações univariadas, mas facilmente se generalizam aos casos multivariados.

Ponto de rotura - (Breakdown point)

O *ponto de rotura* traduz a mais pequena proporção de contaminação (na amostra) que o estimador consegue suportar, sem produzir valores arbitrários nas estimativas. O *ponto de rotura* amostral foi definido em [32] da seguinte forma: numa dada amostra

$(x_1, x_2, \ldots, x_n)$ substituam-se m observações (quaisquer) $x_{i1}, x_{i2}, \ldots, x_{im}$ por valores arbitrários $y_1, y_2, \ldots, y_m$ e designe-se a nova amostra por $(z_1, z_2, \ldots, z_n)$. O ponto de rotura em dimensão finita de T_n é dado por

$$\varepsilon^*(T_n) = \frac{1}{n} max \left\{ m : \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |T_n(z_1, \ldots, z_n)| < \infty \right\}.$$

O *ponto de rotura* assintótico do estimador, em geral, corresponde a $\varepsilon^*(T) = \lim_{n\to\infty} \varepsilon^*(T_n)$. Os estimadores robustos têm *ponto de rotura* positivo, significando isso que uma certa parte dos dados pode corresponder a *outliers*, e mesmo assim gerar resultados úteis. Portanto, é uma medida de robustez mais associada à noção de resistência. Por outro lado, o valor do *ponto de rotura* não depende de cada amostra concreta. Por exemplo, para a média amostral $\varepsilon^*(\bar{x}) = 0$, qualquer que seja a amostra em questão.

Função de Influência

Uma outra noção fundamental em robustez é a de *função de influência*. Esta noção envolve conceitos matemáticos mais avançados, como se pode ver pela definição que se transcreve de seguida.

Seja T um funcional estatístico e F uma distribuição pertencente ao domínio de T. Chama-se *função de influência* de T na distribuição F à função

$$IF(x; T, F) = \lim_{\varepsilon\to 0^+} \frac{T((1-\varepsilon)F + \varepsilon\Delta_x) - T(F)}{\varepsilon},$$

definida pontualmente nos pontos $x \in \Omega \subset \mathbb{R}$ para os quais o limite existe.

Para facilitar a sua compreensão, admita-se que, à partida, a população segue uma distribuição hipotética identificada por F; mas que a verdadeira distribuição é tal, que podem ocorrer valores de acordo com F, com uma probabilidade grande $(1-\varepsilon)$, ou pode ocorrer a observação x com uma probabilidade pequena (ε); então, a mistura $F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$ representa uma distribuição contaminada. Designe-se agora por $T(F)$ o parâmetro estimado por T na distribuição F. Então, ao comparar $T(F)$ com $T(F_{\varepsilon,x})$, a função de influência avalia a variação do estimador perante cada observação discordante x. Deste modo, o estimador é robusto quando a sua função de influência é limitada.

Uma das principais medidas de robustez é a sensibilidade a grandes erros, definida por $\gamma^* = \sup_x |IF(x; T, F)|$. Usando esta medida, o estimador é robusto quando $\gamma^* < \infty$.
Estatísticas tradicionalmente muito usadas em populações normais, como a média ou a variância amostrais têm sensibilidade a grandes erros infinita e, por isso, também não são robustas de acordo com esta medida.
O mesmo acontece quando se generaliza ao caso de populações normais multivariadas: por motivos idênticos aos anteriores, as covariâncias e as correlações amostrais também não são robustas. Portanto, ao usar o vetor das médias amostrais para estimar a tendência central da população ou a matriz de covariâncias amostrais par estimar a escala, está-se a utilizar procedimentos que comprovadamente não são robustos.

A procura de alternativas que sejam robustas passou por estatísticas como a mediana - como medida de tendência central, ou como o desvio absoluto mediano - como medida de dispersão (abreviado por MAD neste trabalho e que se define como a mediana dos desvios absolutos entre as observações e a sua mediana). Mas, estas estatísticas não têm as melhores propriedades quando as populações seguem distribuições normais.

Concretamente, o problema anterior conduz a uma outra noção muito importante no desenvolvimento de métodos robustos, que é a de *eficiência assintótica*.

Eficiência

A eficiência é uma propriedade dos estimadores que indica até que ponto é que um estimador é ótimo, em termos de dispersão das estimativas. A noção está relacionada com a variância do estimador, no sentido em que um estimador eficiente tem a menor variância possível e que, dados dois estimadores, é mais eficiente aquele que tiver menor variância.

Transcreve-se de seguida a definição formal, sem a aprofundar, uma vez que requere conhecimentos de inferência estatística que não serão desenvolvidos neste trabalho, como é o caso da Quantidade de Informação de Fisher da amostra (a qual traduz a quantidade de informação que a amostra contém sobre o parâmetro a estimar).
A definição de eficiência aplica-se a estimadores que sejam centrados, ou seja, tais que $E[\theta] = \theta$, onde $\theta \in \Theta$ é o parâmetro desconhecido a estimar, e calcula-se em distribuições em que exista a Quantidade de Informação de Fisher (como é o caso da distribuição

Normal). Nessas condições, a eficiência de um estimador T (unidimensional) define-se por

$$eff(T) = \frac{1/I(\theta)}{Var[T]},$$

onde Var[T] designa a variância do estimador e $I(\theta)$ a quantidade de informação de Fisher da amostra. Assim, $eff(T)$ é a variância mínima possível para um estimador não enviesado, a dividir pela sua atual variância. A eficiência varia entre 0 e 100% e um estimador diz-se eficiente (numa dada distribuição) quando atinge 100% de eficiência nessa distribuição.

O estimador diz-se assintoticamente eficiente se é eficiente na distribuição assintótica.

Para estimadores que sejam enviesados, com viés $b(\theta) = E[T] - \theta$, a noção de eficiência traduz-se através do erro quadrático médio, definido por $E[(T-\theta)^2]$, uma vez que $E[(T-\theta)^2] = Var[T] + b(\theta)^2$.
Um estimador é tão mais eficiente quanto menor for o seu erro quadrático médio. Note-se que, se o estimador for centrado, o viés $b(\theta)$ é nulo e o erro quadrático médio coincide com a variância do estimador. Logo, dados dois estimadores do mesmo parâmetro, é mais eficiente aquele que tiver menor erro quadrático médio.

O erro quadrático médio estabelece um importante critério para efetuar comparações entre estimadores.

2.6.2 Métodos robustos

Nesta secção descrevem-se, em termos muito gerais, as metodologias robustas que foram utilizadas no capítulo seguinte.

Análise de Componentes Principais Robusta (RPCA)

A abordagem convencional da PCA baseia-se essencialmente no calculo dos valores e dos vetores próprios da matriz de covariâncias ou das correlações amostrais. Contudo os resultados podem ser extremamente sensíveis à presença de observações atípicas nos

dados. Estas observações podem artificialmente aumentar a variância numa direção diferente e esta direção ser erradamente reconhecida como uma componente principal. Estas discrepâncias tem consequências na análise e nas representações gráficas relacionadas com as componentes principais. O método mais intuitivo e direto de obter componentes principais robustas é substituir as estimativas clássicas de localização e de escala pelas suas análogas robustas. Os estimadores-M de localização e de dispersão foram utilizados com este propósito, mas, estes estimadores têm ponto de rotura reduzido em conjuntos com elevado número de parâmetros. Para contornar este problema utilizam-se entre outros, o estimador MCD [40].

MCD (Minimum covariance determinant estimator)

Em termos muito gerais, o estimador MCD original é um estimador da matriz de covariâncias definido por um algoritmo computacional complexo e que usa as seguintes estatísticas de localização e de escala:

1. $\hat{\boldsymbol{\mu}}_0$ - média das h observações para as quais o determinante da matriz de covariância amostral é mínimo, onde $[(n+p+1)/2] \leq h \leq n$ (p é o número de parâmetros e n é a dimensão da amostra).

2. $\hat{\boldsymbol{\Sigma}}_0$ - a correspondente matriz de covariância, multiplicada por um fator de consistência c_0.

O estimador MCD apenas pode ser calculado quando $h > p$, pois de outra forma, a matriz de covariância de qualquer subconjunto das h observações é singular. Uma vez que $h \geq [(n+2)/2]$, esta condição é satisfeita quando $n \geq 2p$.

O estimador MCD é mais robusto quando se toma $h = [n+p+1]/2$. Mas, infelizmente, o MCD é pouco eficiente para o modelo *Normal*. De modo a aumentar a eficiência mantendo a robustez, podem aplicar-se os estimadores ponderados, com

$$\hat{\boldsymbol{\mu}}_{MCD} = \frac{\sum_{i=1}^{n} W(d_i^2)\mathbf{x}_i}{\sum_{i=1}^{n} W(d_i^2)} \tag{2.15}$$

$$\hat{\boldsymbol{\Sigma}}_{MCD} = c_1 \frac{1}{n} \sum_{i=1}^{n} W(d_i^2)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MCD})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MCD})^T \tag{2.16}$$

com $d_i = \sqrt{(\mathbf{x} - \hat{\boldsymbol{\mu}}_0)^T \hat{\boldsymbol{\Sigma}}_0^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_0)}$ e W é uma função adequada de ponderação.

Com base na matriz de covariância MCD pode ser construída uma matriz de correlações robustas. Para todo o $1 \leq i, j \leq p,\ i \neq j$, a correlação robusta entre variáveis X_i e X_j pode ser estimada através de

$$r_{ij}^{rob} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \tag{2.17}$$

com s_{ij} sendo a *i,j-ésimo* elemento da estimativa MCD da covariância.
Como a distância robusta é pouco sensível a observações atípicas, pode ser utilizada para identificá-las. Isto é extremamente útil para conjuntos de dados com mais do que duas dimensões, o que torna difícil ou impossível a sua visualização. Um modo de detetar observações atípicas é representar graficamente a distância robusta de Mahalanobis (com base no estimador MCD) *versus* a distância convencional de Mahalanobis. O valor de corte é determinado por um percentil da distribuição do χ^2 (por exemplo, $\sqrt{\chi^2_{p,0.975}}$) e baseia-se na distribuição assintótica da distância robusta. Em [22], para além da definição do estimador MCD, apresentam-se as suas propriedades (ponto de rotura, função influência) que comprovam as boas características de robustez deste estimador.
No capítulo seguinte, os cálculos e resultados foram determinados com o método MCD, mas numa versão computacionalmente mais refinada, que é designada por *Fast MCD*.

Análise de Agrupamentos Robusta (RCA)

Tal como acontece com outros procedimentos estatísticos não robustos, os métodos de agrupamento podem ser fortemente influenciados até por pequenas frações de observações atípicas. Por exemplo, devido a observações atípicas, dois ou mais grupos podem ser artificialmente agrupados, ou um grupo não informativo pode ser formado apenas com escassas observações atípicas. A aplicação de métodos robustos neste contexto é deste modo desejável. Por outro lado, as técnicas robustas por vezes agrupam as observações atípicas num único grupo, permitindo análises estatísticas diferenciadas.

Descreve-se genericamente o método robusto que se baseia no Agrupamento Aparado (*Trimmed cluster*) das observações, em particular, o método *TCLUST* de acordo com [17]. Contrariamente a outras abordagens, em que se tenta acomodar no modelo as observações atípicas, no agrupamento aparado procura-se excluí-las completamente do modelo.

A primeira tentativa de remover o efeito negativo das observações atípicas na análise de agrupamentos, consistiu em aplicar métodos de agrupamento aparado no método convencional *k-means*. Neste método, o agrupamento aparado baseia-se na remoção de uma fração α das observações mais atípicas, deste modo evitando a influência dessas obser-

vações. Este processo também serve para identificar observações atípicas com potencial interesse. O procedimento considera todo o conjunto de dados e cada observação é considerada *outlier* ou incluída num grupo, sendo que a totalidade das observações atípicas é também agrupada num grupo à parte.

A aplicação do método impõe a definição de certas restrições sobre as matrizes de covariâncias dos grupos. No presente trabalho utilizou-se o método conforme está programado na package *tclust* do programa R, definindo uma restrição com base nos valores próprios das matrizes de covariância dos grupos. Representando por $\lambda_l(\mathbf{\Sigma_j})$ os valores próprios das matrizes de dispersão dos k grupos e

$$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l(\mathbf{\Sigma_j}) \quad e \quad m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l(\mathbf{\Sigma_j}) \tag{2.18}$$

os máximos e os mínimos valores próprios, a razão M_n/m_n, que representa o número de condição, terá de ser inferior ou igual que um determinado valor da restrição (que por sua vez é maior ou igual que 1). Este tipo de restrição confina os limites dos eixos de densidade constante da distribuição, através da matriz Σ_j estimada, assumindo a normalidade. Deste modo, controlam-se simultaneamente o tamanho relativo dos grupos e também o desvio da forma esférica em cada grupo.

2.6.3 Regressão linear múltipla robusta

Regressão linear múltipla robusta (RMLR)

O método dos mínimos quadrados, geralmente utilizado na estimação dos coeficientes do modelo de regressão, na presença de observações atípicas pode conduzir a resultados distorcidos. Os estimadores dos mínimos quadrados têm ponto de rotura igual a zero e uma função de influência ilimitada. O objetivo da regressão robusta consiste em reduzir a influência das observações atípica, mantendo-as no processo de estimação, em vez da alternativa radical de, pura e simplesmente, as eliminar. Em termos genéricos, a regressão robusta atua de modo a atribuir *pesos* às observações, consoante a influência que exercem no processo de estimação. A observações a que correspondem resíduos de grande magnitude será atribuído um menor *peso*. Ao limitar a influência das observações atípicas, a regressão robusta realiza um ajustamento que reflete principalmente a contribuição da

maioria dos dados.

Foram desenvolvidos diversos métodos robustos de estimação dos parâmetros do modelo de regressão. Aqui referem-se os estimadores-MM e as suas propriedades. Os estimadores-MM são estimadores robusto, com elevado ponto de rotura e que podem atingir alta eficiência [32] em populações normais.

Os estimadores-MM da regressão são estimadores-M calculados a partir de estimativas iniciais convenientes. No caso de um só regressor, o estimador-M de regressão é definido implicitamente por $\hat{\beta}$ tal que:

$$\hat{\beta}\ minimiza\ \sum_{i=1}^{n} \rho(\frac{y_i - x_i\hat{\beta}}{\hat{\sigma}}), \tag{2.19}$$

para funções ρ convenientes, onde $\hat{\sigma}$ é um estimador de escala da variância dos erros do modelo, sendo que nas propostas iniciais se recomendava o uso do MAD para esse efeito. No caso dos estimadores-MM, a estimativa de escala $\hat{\sigma}$ é também obtida com um estimador-M de escala (ver, por exemplo [29]).

Análise Discriminante Quadrática Robusta (QDA)

A versão robusta da análise discriminante quadrática não introduz conceitos teóricos adicionais em relação aos já focados nos tópicos anteriores. O método segue a explicação apresentada na secção 2.3, substituindo todas as estimativas convencionais das matrizes de covariâncias por estimativas robustas, em particular, as produzidas pelo MCD (na versão computacional mais refinada, *Fast-MCD*).

Capítulo 3

O *Site index* em função de parâmetros edafo-climáticos

3.1 Descrição dos Dados

Os dados utilizados neste trabalho provêm de povoamentos comerciais puros da *Eucalyptus globulus*, com árvores da mesma idade e submetidos ao mesmo tipo de silvicultura. Localizam-se de norte a sul do país, particularmente nas regiões consideradas de aptidão para a produção comercial desta espécie.

Os dados foram recolhidos em parcelas do inventário florestal anual, sendo as parcelas circulares com 400m^2 de área. A intensidade de amostragem é em média de uma parcela por cada 4 ha.

Recorda-se que, como se referiu no Capítulo 1, é usual e vantajosa a utilização da variável altura dominante (h_{dom}) para representar a altura do povoamento florestal, definida como sendo a altura média das 100 árvores com maior diâmetro por hectare. Nas parcelas consideradas, a altura dominante (h_{dom}) é avaliada com base na medição da altura das 4 árvores com maior diâmetro à altura do peito. A altura destas árvores é recolhida com o equipamento hipsómetro *Vertex* [1]. Por outro lado a variável *Site index* (S) (que é o principal objetivo deste estudo), em povoamentos da *Eucalyptus globulus*, corresponde à altura dominante (h_{dom}) a uma idade de referência de 10 anos e expressa a qualidade da estação.

Do total das parcelas disponíveis, lançadas no período entre os anos 2000 e 2010, foram selecionadas 3358 parcelas representativas de povoamentos florestais comerciais seminais, bem adaptados às condições edafo-climáticas em presença e isentos de eventos tais como

ocorrência de problemas silviculturais e ocorrência de pragas ou doenças.

Não existindo número suficiente de parcelas onde as medições se realizaram na idade de referência (aos 10 anos), utilizou-se um modelo de crescimento desenvolvido por M.Tomé [41] para estimar o *Site index* a partir da h_{dom}. No entanto, para evitar erros, associados a este método, reduziram-se as parcelas apenas às medidas entre os 8.5 e 13 anos, uma vez que quanto maior amplitude de projeção em idade, maior é o erro da estimativa. Deste modo, os dados a usar foram recolhidos em conformidade com as normas internacionais referidas na secção 1.1 do capítulo introdutório e são constituídos por um conjunto de 3358 observações disponíveis.

O presente trabalho foi desenvolvido apenas com um subconjunto destas observações. De facto, foram selecionadas, aleatoriamente 3022 observações, isto é, 90%, permitindo deste modo reservar uma subamostra para validação e comparação dos resultados obtidos pelos diferentes modelos e métodos de estimação. No seguimento, a subamostra usada é mais frequente designada por *dados de calibração*. A subamostra complementar é constituída por 336 observações é neste estudo mais frequentemente designada por *dados de validação*. A figura 3.1 apresenta a localização geográfica, no território nacional, onde foram recolhidos os dados utilizados na análise.

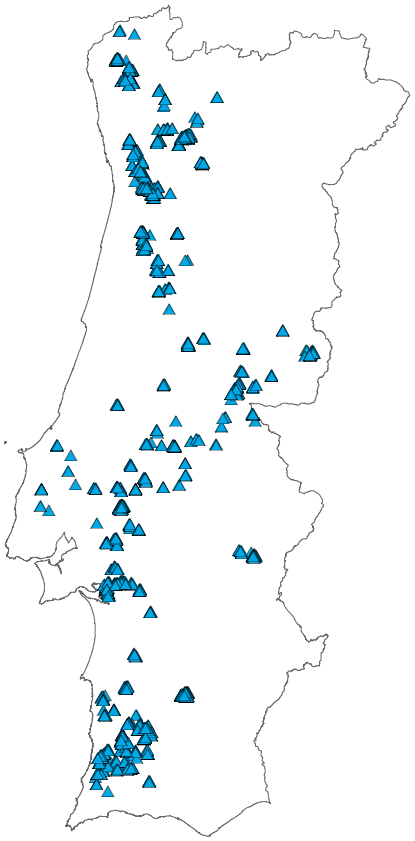


Figura 3.1: Localização das parcelas onde foram recolhidas as medições do crescimento das árvores e das variáveis ambientais.

Na seleção das variáveis ambientais consideraram-se dois aspetos: por um lado, variáveis que contribuem para a disponibilidade de água às plantas; por outro, que a medição das variáveis seja de fácil acesso e recolha, de modo a que o modelo a desenvolver seja o mais operacional possível. Assim, a seleção de variáveis climáticas recaiu sobre as seguintes:

- a precipitação média anual (mm);
- o número médio anual de dias com precipitação superior a 1 mm ;
- a precipitação média anual nos meses mais quentes (mm);
- a temperatura média máxima anual ($^{\circ}C$);
- a temperatura média mínima anual ($^{\circ}C$);
- a evapotranspiração média anual (mm).

Na seleção das variáveis relacionadas com o solo procuraram-se aquelas que poderão refletir a capacidade do solo para reter e disponibilizar água às plantas. Assim, as variáveis selecionadas foram:

- a profundidade do solo (cm);
- a pedregosidade do solo (%);
- a quantidade de água no solo disponível para as plantas (mm) .

Ainda, de modo a caraterizar a topografia do local, selecionaram-se mais três variáveis, nomeadamente:

- a altitude (m);
- a exposição ($^{\circ}$);
- o declive (%).

Foram ainda consideradas as variáveis latitude e longitude do local, medidas no centro da parcela de inventário, uma vez que a produtividade é um indicador com variação espacial. Na tabela 3.1 estão descritas as variáveis anteriormente referidas.

Associadas às coordenadas, latitude e longitude, de cada local, foram recolhidas as variáveis topográficas, climáticas e de solo.

As variáveis topográficas foram determinadas com base na produção dum Modelo Digital de Terreno: *Triangulated Irregular Network* (TIN), recorrendo à aplicação *ArcGIS for Desktop Basic 10* da *ESRI* [10] e à informação de curvas de nível em formato digital à escala 1:25000. As variáveis climáticas foram recolhidas a partir de cartografia digital nacional.

As variáveis de solo foram recolhidas a partir de cartografia de zonamento edáfico, produzida ao nível das unidades de gestão (RAIZ, 2009, relatório). Neste processo de cartografia de solos, são recolhidas informações de profundidade do solo, pedregosidade do solo, entre outras, ao nível do perfil do solo. A delimitação de uma zona homogénea

envolve a associação desta a, no máximo, dois perfis de solo, que caraterizam dois tipos distintos de solo. Nas situações em que uma parcela de inventário se encontrava numa zona caraterizada por dois tipos de solo, os critérios utilizados na associação a um tipo de solo foram os seguintes: tipo de solo caraterizado pelo ou pelos perfis mais próximos da localização da parcela de inventário; produtividade avaliada na parcela de inventário; tipo de solo mais representativo.

Para além das variáveis já referidas, foi considerada uma variável que descreve a capacidade do solo para armazenar a água e disponibilizá-la para a planta (*available water storage capacity* - *awsc*). Esta variável, neste estudo, foi estimada utilizando os dados recolhidos de textura, profundidade e pedregosidade do solo e relações estabelecidas entre estes fatores conforme, por exemplo, em ([7]).

Foi ainda testada a transformação da variável exposição, de acordo com metodologia descrita em [5], e ainda, de modo a refletir o conhecido comportamento distinto em regiões de elevado e reduzido *deficit hídrico*. Assim, a exposição foi classificada em níveis, variando de 2 (para exposições mais favoráveis) até ao nível zero (nas exposições menos favoráveis). Na região sul do país, onde o *deficit hídrico* é mais intenso, a classificação de 2 foi atribuída às exposições (Norte - Este) e a classificação de 0 às exposições (Sul - Oeste). Na região norte, onde o *deficit* hídrico é menos significativo ou inexistente, a classificação de 2 foi atribuída às exposições (Sul - Oeste) e a classificação de 0 às exposições (Norte - Oeste). Assim, após a codificação referida, a exposição solar passou a ser estudada através da nova variável, designada no seguimento por exp_{pond}..

Na tabela 3.2 estão descritas as variáveis e respetivos métodos de recolha.

Estudos desenvolvidos por diversos autores, nomeadamente, em [15] e [18], com objetivos idênticos ao do presente trabalho, referem como vantajosas as seguintes transformações de variáveis:

(a) Para o *Site index*: considerar a transformação de S em $S^{1/2}$;
(b) Para a capacidade do solo de armazenar e disponibilizar água às plantas: considerar a transformação de awsc em log(awsc);
(c) Para a altitude: considerar a transformação de alt em $\sqrt{alt}$.

De modo a avaliar se essas transformações continuariam a ser adequadas, na análise deste conjunto de dados foi utilizada a transformação de *Box-Cox* para determinar, para cada variável, o expoente λ que aplicado à variável a sua distribuição se aproxime mais da

Normal. Analisaram-se os expoentes e o resultado da transformação variável a variável. No caso da variável *Site index* a transformação com base no expoente sugerido, $\lambda = 0.057$, aproxima-a mais da distribuição *Normal*, ver Figura 3.2(b) na Secção 3.2, mas não o suficiente para que aplicados testes de ajustamento à distribuição *Normal*, nomeadamente o teste de *Lilliefors K-S*, a confirmem.

Na transformação, sugerida para a variável quantidade de água no solo disponível para às plantas *awsc*, $\lambda = 0.052$, o comportamento assemelha-se ao produzido pela função logarítmica sobre esta variável, pelo que se optou pela transformação logarítmica, ver Figura 3.3(b) da Secção 3.2. As transformações aplicadas às variáveis: Precipitação média anual *prec*, $\lambda = -0.44$; Evapotranspiração média anual *evap*, $\lambda = 0.178$ e Altitude *alt*, $\lambda = 0.386$, também parecem ser benéficas, ver Figura 3.4 da Secção 3.2.

Apesar de em todas as situações anteriormente referidas a distribuição da variável após a transformação se aproximar mais da distribuição *Normal*, em nenhum caso é confirmada via utilização de testes de hipóteses estatísticos com este objetivo, ver tabela B.9 no Apêndice B.

Assim, estas variáveis foram tratadas com e sem as transformações referidas. Os resultados posteriormente obtidos mostraram que a transformação referida em (b), tal como as que resultaram da aplicação da transformação de *Box-Cox* acima referidas, se traduziram em vantagens para a modelação, com exceção da variável *Site index*. Logo, o trabalho prosseguiu com os valores do *Site index* sem qualquer transformação, enquanto que para a quantidade de água no solo disponível para as plantas se utilizou a transformação logaritmica e para as variáveis precipitação média anual altitude do local e evapotranspiração média anual se utilizou as transformações sugeridas pelo método *Box-Cox*.

Foram ainda consideradas algumas variáveis categóricas adicionais de modo a procurar identificar o maior conjunto possível de fontes de variabilidade nos dados e tentar identificar grupos naturais de observações. Estas variáveis foram as seguintes: tipo de solo (Arenossolo, Leptossolo, Regossolo, Umbrissolo e Cambissolo), de acordo com o sistema internacional de classificação dos solos [11]; grupo genérico litológico (Areias, Granito e Xisto) de acordo com cartografia nacional incluída no Atlas do Ambiente [9]; carácter *húmico* na classificação do tipo de solo, refletindo a quantidade de matéria orgânica no solo; carácter *hidromórfico* na classificação do tipo de solo, refletindo a ocorrência de encharcamento dos solos; estrutura dos solos natural ou modificado por intervenção humana (solos antrópicos ou não antrópicos). Estas variáveis foram tratadas como variáveis *mudas* (*Dummy variables*).

Variável	Descrição da Variável	Unidades
	Numéricas	
S	*Site index* - Altura dominante (h_{dom}) projetada para idade de referência de 10 anos	m
x	Latitude do local	m
y	Longitude do local	m
alt	Altitude do local	m
dcl	Declive do local	%
exp	Exposição do local	graus
Prof	Profundidade do Solo	cm
Pedreg	Pedregosidade do Solo	%
prec	Precipitação média anual	mm
dprec	Número de dias com precipitação superior a 1 mm	
prec678	Precipitação média dos meses mais quentes	mm
evap	Evapotranspiração real média anual	mm
tmin	Temperatura mínima média anual	$^{\circ}C$
tmax	Temperatura máxima média anual	$^{\circ}C$
awsc	available water storage capacity (quantidade de água no solo disponível para as plantas)	mm
	Categóricas	
Litologia	Areias(1), Xisto(2), Granito(3), Outros(0)	
Tipo de Solo	Arenossolos (Areno(1)), Leptossolos (Lepto(2)), Regossolos (Rego(3)), Umbrissolos (Umbri(4)), Cambissolos (Cambi(5))	
MO	Carácter húmico do solo (presente - 1 ou ausente - 0)	
HidroFAO	Carácter hidrófito do solo (presente - 1 ou ausente - 0)	
Antropico	Carácter antrópico do solo (presente - 1 ou ausente - 0)	
Profundidade	Classes de Profundidade do solo (Prof1: <=30 e >=100; Prof2: 30 a 65; Prof3: 65 a 100)	cm
Pedregosidade	Classes de Pedregosidade do Solo (Pedreg1: <=30; Pedreg2: 30 a 65; Pedreg3: >65)	%
deficitHidro	Deficit hídrico - assumido para precipitação média anual inferior a 1000 mm (Com 1; sem 0)	

Tabela 3.1: Variáveis e respetivas unidades.

Variável	Método de Recolha	Fonte
	Numéricas	
S	Altura dominante aos 10 anos com equação do mod. Globulus 3.0 (M.Tomé et. al., 2003)	Aliança Florestal - grupo Portucel Soporcel
x	GPS (WGS84)	Aliança Florestal - grupo Portucel Soporcel
y	GPS (WGS84)	Aliança Florestal - grupo Portucel Soporcel
alt	MDT com base em curvas de nível e pontos cotados à escala 1:25000	CM IGeoE Série M888 à escala 1:25000
dcl	MDT com base em curvas de nível e pontos cotados à escala 1:25000	CM IGeoE Série M888 à escala 1:25000
exp	MDT com base em curvas de nível e pontos cotados à escala 1:25000	CM IGeoE Série M888 à escala 1:25000
Prof	Avaliada com base na abertura de perfis de solo	RAIZ, Caraterização edafo-climática
Pedreg	Avaliada com base na abertura de perfis de solo	RAIZ, Caraterização edafo-climática
prec	Cartografia Digital	Instituto Meteorológico, 1971-2000
dprec	Cartografia Digital	Instituto Meteorológico, 1971-2000
prec678	Cartografia Digital	Instituto Meteorológico, 1971-2000
evap	Cartografia Digital	Instituto Meteorológico, 1971-2000
tmin	Cartografia Digital	Instituto Meteorológico, 1971-2000
tmax	Cartografia Digital	Instituto Meteorológico, 1971-2000
awsc	função da textura, profundidade, pedregosidade.	RAIZ, Caraterização edafo-climática
	Categóricas	
Litologia	Cartografia Digital	Atlas do Ambiente
Tipo de Solo	Sistema de Classificação de Solos da FAO	RAIZ, Caraterização edafo-climática
MO	Sistema de Classificação de Solos da FAO	RAIZ, Caraterização edafo-climática
HidroFAO	Sistema de Classificação de Solos da FAO	RAIZ, Caraterização edafo-climática
Antropico	Sistema de Classificação de Solos da FAO	RAIZ, Caraterização edafo-climática
Profundidade	Avaliada com base na abertura de perfis de solo	RAIZ, Caraterização edafo-climática
Pedregosidade	Avaliada com base na abertura de perfis de solo	RAIZ, Caraterização edafo-climática
deficitHidro	Precipitação média anual	Instituto Meteorológico, 1971-2000

Tabela 3.2: Método de recolha e fonte de informação sobre as variáveis.

3.2 Análise exploratória de dados

3.2.1 Análise preliminar

O tratamento das observações começou por uma análise exploratória dos dados, com um conjunto de 3022 observações que formam a *amostra de calibração*. As medidas descritivas e as representações gráficas sugerem distribuições com características diferentes consoante as variáveis. Nas figuras 3.2, 3.3 e 3.4 apresentam-se representações gráficas univariadas da distribuição do *Site index*, nomeadamente: o diagrama de dispersão (relativamente ao índice das observações); o gráfico de *quantis* relativo à distribuição *Normal*; o histograma e o diagrama de bigodes, bem como de algumas outras variáveis que se pareça virem a ser relevantes, nomeadamente: Precipitação média anual (*prec*), evapotranspiração (*evap*) e quantidade de água que o solo disponível para as plantas (*awsc*).

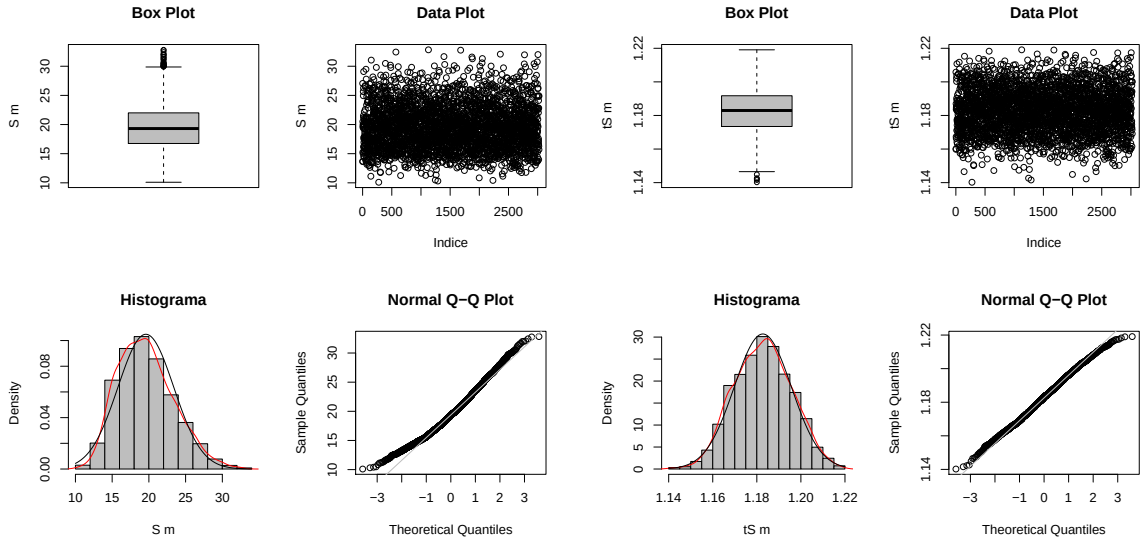
A variável *Site index* apresenta uma distribuição aproximadamente normal com ligeira assimetria positiva, refletindo frequências mais elevadas em valores mais reduzidos da amostra. A caixa de bigodes evidencia um grupo de observações atípicas no seguimento do bigode superior. Quanto às correspondentes representações gráficas para as outras variáveis, destacam-se desde já as seguintes apreciações:

(a) tanto a precipitação como a evapotranspiração têm caixas de bigodes assimétricas (embora em sentidos opostos); ambas parecem afastar-se consideravelmente da distribuição *Normal*; sobretudo, nota-se a eventual existência de dois grupos (pelo menos), o que é sugerido pelos histogramas bimodais e também pelos diagramas de dispersão, embora com menos ênfase.

    Estas considerações, juntamente com outras que serão referidas, contribuíram para que se viesse a utilizar a análise de agrupamentos para investigar/confirmar a possibilidade do *Site index* dever ser estimado de modo diferente para diferentes grupos de observações, talvez associados a distintas localizações geográficas.

(b) A distribuição do logaritmo da quantidade de água no solo disponível para às plantas apresenta uma caixa de bigodes e um histograma que sugerem simetria da distribuição, enquanto que no gráfico de *quantis* se observa que a distribuição *Normal* parece modelar bem a parte central da distribuição, mas não as caudas.

Aplicados testes de hipóteses estatísticos de ajustamento para validar a distribuição *Normal* univariada dos dados, para todas as variáveis, incluindo o *Site index*, esta hipótese



(a) Dados Originais. (b) Dados após a transformação de *Box-cox*.

Figura 3.2: Representação gráfica univariada da variável *Site index* (m).

é rejeitada. O *p-value* do teste de ajustamento à distribuição Normal do *Site index* foi de $p_1 = 2.28 \times 10^{-17}$ e $p_2 = 6.27 \times 10^{-10}$, respetivamente com o teste de *Shapiro-Wilk* e de *Lilliefors (K-S)*. Os *p-value* verificados para as restantes variáveis podem ser consultados na Tabela B.3 do apêndice B.

Nas tabelas 3.3(a) e 3.3(b) apresentam-se, respetivamente, medidas de tendência central (média e mediana) e medidas de dispersão (desvio padrão e MAD) para o conjunto de dados utilizados na calibração e na validação. Mais estatísticas sumárias para cada variável podem ser consultadas nas Tabelas B.1 e B.2 no Apêndice B.

Após uma análise univariada prosseguiu-se com a análise preliminar bivariada. Começou por se produzir matrizes de diagramas de dispersão, já apresentados na secção 1.2 do capítulo introdutório. Recorda-se que nessas representações gráficas se observou: a possibilidade de existirem grupos de observações; a eventual necessidade de transformação de variáveis; e o relacionamento linear existente entre algumas variáveis.

Para além da análise gráfica acrescenta-se agora informação sobre correlações amostrais. Tendo em conta as preocupações do trabalho do ponto de vista da robustez estatística, calcularam-se as matrizes de correlação utilizando o método convencional baseado no coeficiente de correlação de *Pearson* e método robusto - o *Fast MCD (Minimum Covariance Determinant Estimator)*. Os resultados encontram-se nas Tabelas 3.4(a) e 3.4(b).

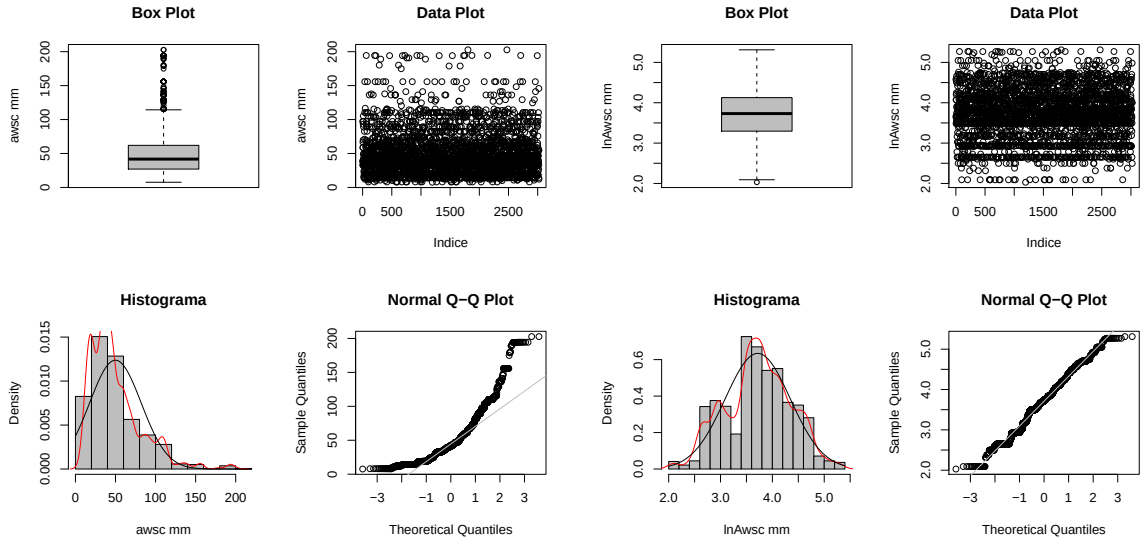
(a) Amostra de Calibração (com 3022 observações)

	Media	Mediana	Desvio Padrão	MAD
S	19.59	19.31	3.80	3.87
alt	212.91	200.00	126.93	148.26
dcl	12.11	12.00	8.65	10.38
exp	170.57	168.69	93.82	108.24
tmin	9.46	9.46	0.82	0.95
tmax	20.45	20.48	1.49	1.90
prec	1126.54	868.10	477.80	421.06
dprec	102.54	96.00	20.59	26.69
prec678	68.10	48.10	41.93	39.31
evap	40.90	42.40	8.26	12.01
Prof	43.67	37.50	26.59	18.53
Pedreg	39.51	40.00	24.11	37.06
awsc	49.88	41.74	32.23	26.19
x	184131.30	177303.50	29523.41	23146.33
y	297583.65	306739.84	164140.41	224555.80

(b) Amostra de Validação (com 336 observações).

	Media	Mediana	Desvio Padrão	MAD
S	19.34	19.01	3.82	3.63
alt	196.35	200.00	120.08	148.26
dcl	12.77	13.00	9.16	10.38
exp	183.89	186.34	94.51	103.34
tmin	9.57	9.54	0.80	0.96
tmax	20.59	20.69	1.45	1.85
prec	1080.74	812.55	478.17	257.75
dprec	100.71	92.95	20.22	21.05
prec678	63.80	41.60	41.99	29.31
evap	41.68	42.70	8.24	11.51
Prof	43.04	40.00	25.55	22.24
Pedreg	39.15	40.00	24.07	37.06
awsc	52.43	42.52	33.46	27.02
x	182615.42	176635.50	29607.90	22797.82
y	280293.67	270431.19	164565.90	256447.04

Tabela 3.3: Medidas de tendência central e de dispersão.

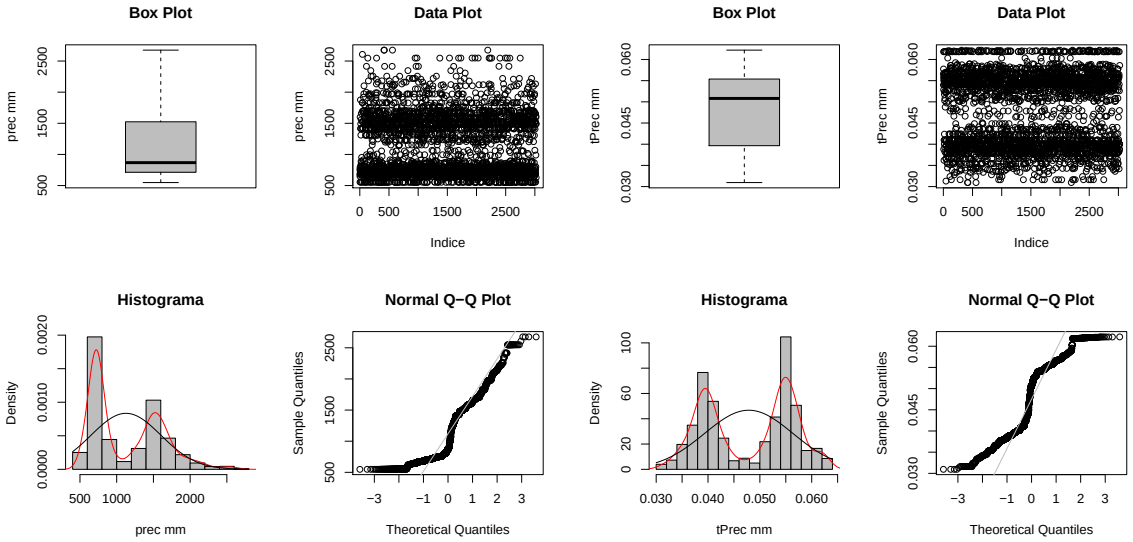


(a) Dados Originais. (b) Dados após transformação logarítmica.

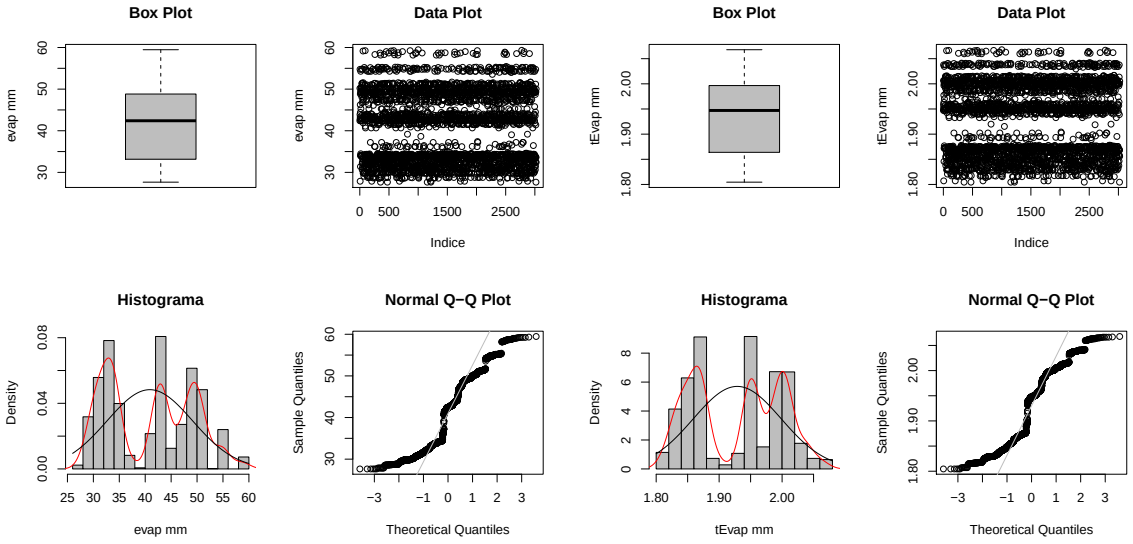
Figura 3.3: Representação gráfica univariada da variável quantidade de água no solo disponível para às plantas (mm).

Contrariamente ao que seria de esperar, a correlação das diversas variáveis com o *Site index* é sempre baixa ou mesmo muito reduzida. Na versão convencional, o maior valor ocorre relativamente à variável evapotranspiração média anual e é de apenas 39%. Na versão robusta, o maior valor ocorre nas variáveis Precipitação média anual; Precipitação média anual dos meses mais quentes e Evapotranspiração média anual, que traduzem a disponibilidade ou indisponibilidade de água para as plantas ao longo do ano e em particular nos meses secos (menor precipitação e maior evapotranspiração), e é apenas de 37%. É ainda possível verificar que as correlações entre variáveis em estudo e o *Site index* é ligeiramente superior quando utilizado método robusto. A análise das matrizes evidencia também um forte correlacionamento entre as variáveis climáticas e destas com a altitude e latitude do local. A variável exposição, na sua forma transformada apresenta muito baixa correlação com qualquer das outras variáveis.

Em face dos comentários anteriores, nomeadamente da forte correlação entre as diversas variáveis explicativas, optou-se por fazer uma análise de componentes principais. Também nesta análise se recorreu ao método convencional e a um método robusto.



(a) Precipitação (mm)



(b) Evapotranspiração (mm)

Figura 3.4: Representação gráfica univariada de variáveis. À esquerda variável original e à direita variável transformada.

(a) Coeficiente de Correlação de *Pearson*

	S	alt	exp	dcl	tmin	tmax	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	0.11	-0.10	0.26	-0.25	-0.29	0.35	0.36	0.34	-0.39	0.16	-0.10	0.34	-0.10	0.33
alt	0.11	1.00	-0.02	0.35	**-0.56**	**-0.65**	**0.52**	0.40	**0.50**	-0.39	-0.24	0.43	-0.07	**0.52**	0.41
exp	-0.10	-0.02	1.00	0.03	0.01	0.02	0.00	-0.01	-0.01	0.04	-0.00	0.10	-0.09	-0.01	-0.03
dcl	0.26	0.35	0.03	1.00	-0.24	-0.48	0.37	0.41	0.33	-0.47	-0.13	**0.54**	-0.02	0.10	0.27
tmin	-0.25	-0.56	0.01	-0.24	1.00	**0.57**	**-0.63**	**-0.60**	**-0.70**	**0.53**	-0.04	-0.16	-0.05	-0.36	**-0.74**
tmax	-0.29	-0.65	0.02	-0.48	0.57	1.00	**-0.86**	**-0.81**	**-0.79**	**0.83**	0.11	-0.41	-0.10	0.03	**-0.65**
prec	0.35	0.52	0.00	0.37	-0.63	-0.86	1.00	**0.93**	**0.97**	**-0.88**	-0.07	0.27	0.15	-0.02	**0.85**
dprec	0.36	0.40	-0.01	0.41	-0.60	-0.81	0.93	1.00	**0.93**	**-0.90**	-0.09	0.34	0.09	-0.01	**0.87**
prec678	0.34	0.50	-0.01	0.33	-0.70	-0.79	0.97	0.93	1.00	**-0.85**	-0.09	0.26	0.10	0.12	**0.94**
evap	-0.39	-0.39	0.04	-0.47	0.53	0.83	-0.88	-0.90	-0.85	1.00	0.13	-0.40	-0.07	0.11	**-0.76**
Prof	0.16	-0.24	-0.00	-0.13	-0.04	0.11	-0.07	-0.09	-0.09	0.13	1.00	-0.44	**0.54**	-0.34	-0.07
Pedreg	-0.10	0.43	0.10	0.54	-0.16	-0.41	0.27	0.34	0.26	-0.40	-0.44	1.00	-0.47	0.35	0.22
awsc	0.34	-0.07	-0.09	-0.02	-0.05	-0.10	0.15	0.09	0.10	-0.07	0.54	-0.47	1.00	-0.29	0.04
x	-0.10	0.52	-0.01	0.10	-0.36	0.03	-0.02	-0.01	0.12	0.11	-0.34	0.35	-0.29	1.00	0.26
y	0.33	0.41	-0.03	0.27	-0.74	-0.65	0.85	0.87	0.94	-0.76	-0.07	0.22	0.04	0.26	1.00

(b) *Fast MCD - Minimum Covariance Determinant*

	S	alt	exp	dcl	tmin	tmax	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	0.25	-0.10	0.29	-0.30	-0.31	0.36	0.34	0.36	-0.36	0.05	-0.01	0.26	0.15	0.35
alt	0.25	1.00	-0.05	0.44	**-0.56**	**-0.68**	**0.67**	**0.59**	**0.62**	**-0.63**	-0.40	**0.51**	-0.13	**0.58**	0.49
exp	-0.10	-0.05	1.00	0.04	0.07	0.04	-0.05	-0.05	-0.06	0.06	0.02	0.09	-0.10	-0.05	-0.07
dcl	0.29	0.44	0.04	1.00	-0.30	**-0.58**	0.49	**0.51**	0.43	**-0.60**	-0.21	**0.62**	-0.05	0.19	0.34
tmin	-0.30	-0.56	0.07	-0.30	1.00	**0.53**	**-0.71**	**-0.68**	**-0.77**	**0.58**	0.06	-0.21	0.05	**-0.65**	**-0.81**
tmax	-0.31	-0.68	0.04	-0.58	0.53	1.00	**-0.88**	**-0.84**	**-0.79**	**0.89**	0.29	**-0.58**	0.06	-0.15	**-0.67**
prec	0.36	0.67	-0.05	0.49	-0.71	-0.88	1.00	**0.95**	**0.98**	**-0.92**	-0.28	0.47	-0.10	0.40	**0.92**
dprec	0.34	0.59	-0.05	0.51	-0.68	-0.84	0.95	1.00	**0.94**	**-0.92**	-0.27	**0.51**	-0.10	0.43	**0.91**
prec678	0.36	0.62	-0.06	0.43	-0.77	-0.79	0.98	0.94	1.00	**-0.88**	-0.25	0.39	-0.09	0.49	**0.97**
evap	-0.36	-0.63	0.06	-0.60	0.58	0.89	-0.92	-0.92	-0.88	1.00	0.32	**-0.60**	0.10	-0.29	**-0.80**
Prof	0.05	-0.40	0.02	-0.21	0.06	0.29	-0.28	-0.27	-0.25	0.32	1.00	-0.44	**0.53**	-0.31	-0.14
Pedreg	-0.01	0.51	0.09	0.62	-0.21	-0.58	0.47	0.51	0.39	-0.60	-0.44	1.00	-0.44	0.25	0.27
awsc	0.26	-0.13	-0.10	-0.05	0.05	0.06	-0.10	-0.10	-0.09	0.10	0.53	-0.44	1.00	-0.18	-0.07
x	0.15	0.58	-0.05	0.19	-0.65	-0.15	0.40	0.43	0.49	-0.29	-0.31	0.25	-0.18	1.00	0.49
y	0.35	0.49	-0.07	0.34	-0.81	-0.67	0.92	0.91	0.97	-0.80	-0.14	0.27	-0.07	0.49	1.00

Tabela 3.4: Matriz de Correlações (com 3022 observações).

(a) PCA convencional.

	Valores próprios	%de Variância	% de variância acumulada
PC 1	5.8	**53.1**	53.1
PC 2	2.1	**19.0**	72.1
PC 3	0.9	**8.3**	80.4
PC 4	0.7	**6.5**	**87.0**
PC 5	0.5	4.9	91.8
PC 6	0.4	3.4	95.2
PC 7	0.2	1.9	97.1
PC 8	0.2	1.4	98.5
PC 9	0.1	0.8	99.3
PC 10	0.0	0.4	99.8
PC 11	0.0	0.2	100.0

(b) PCA Robusta.

	Valores próprios	%de Variância	% de variância acumulada
PC 1	2.1	**46.8**	46.8
PC 2	1.6	**28.3**	75.1
PC 3	0.8	**6.8**	81.9
PC 4	0.8	**6.3**	**88.2**
PC 5	0.7	5.6	93.9
PC 6	0.5	2.8	96.7
PC 7	0.3	1.2	97.9
PC 8	0.3	1.1	99.0
PC 9	0.2	0.5	99.5
PC 10	0.2	0.3	99.8
PC 11	0.1	0.1	100.0

Tabela 3.5: PCA. Valores próprios, percentagem de variação total e acumulada por PC.

3.2.2 Análise de Componentes Principais convencional e robusta

Na Análise de Componentes Principais (PCA) foram utilizadas 3022 observações, que constituem os *dados de calibração*, e o conjunto das variáveis descritas na secção 3.1, o qual incluiu: as variáveis originais ou as suas transformadas (nos casos da precipitação, da evapotranspiração, da altitude e da capacidade do solo disponibilizar água às plantas).

Adicionalmente utilizaram-se ainda as variáveis *mudas* (*Dummy Variables*) como variáveis suplementares, que não entram na definição das componentes.

Os dados foram previamente estandardizados. Na aplicação de métodos convencionais utilizou-se o método de estandardização convencional, isto é: os dados foram centrados e reduzidos com base nas respetivas médias e desvio padrão amostrais. No caso dos métodos robustos, utilizou-se para o mesmo efeito a mediana e o desvio absoluto mediano (MAD), seguindo a sugestão de [29].

Nas Tabelas 3.5 e 3.6 apresentam-se os resultados obtidos.

A PCA confirmou a elevada correlação entre algumas das variáveis e a existência de,

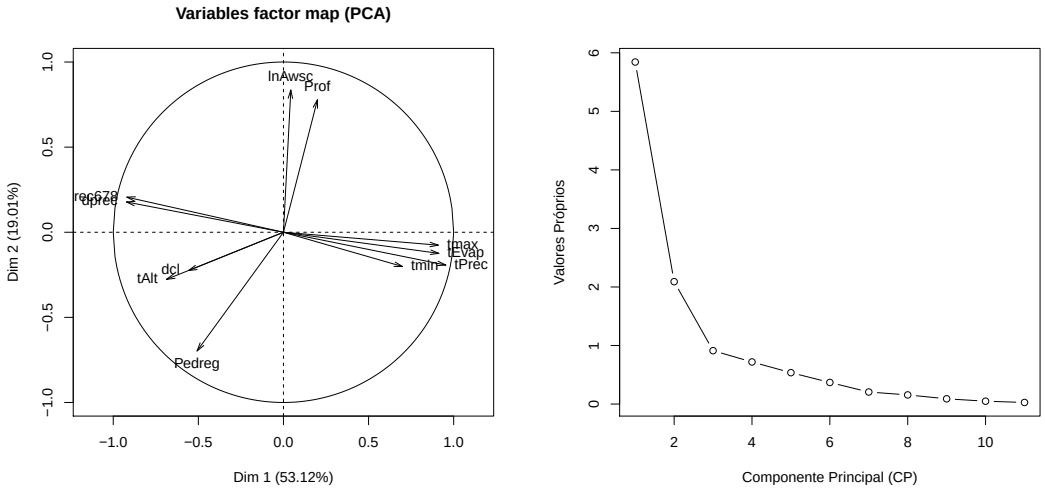
(a) PCA convencional.

	PC1	PC2	PC3	PC4	PC5	PC6
tAlt	-0.28	-0.19	0.04	**0.67**	0.37	0.26
dcl	-0.23	-0.16	**0.75**	-0.03	-0.10	**-0.53**
tmin	0.29	-0.14	0.26	**-0.55**	**0.42**	0.42
tmax	**0.38**	-0.05	-0.07	-0.03	-0.20	-0.28
tPrec	**0.39**	-0.13	0.08	0.16	0.01	-0.03
dprec	**-0.38**	0.12	-0.10	-0.28	-0.04	0.00
prec678	**-0.38**	0.14	-0.23	-0.10	-0.04	-0.06
tEvap	**0.38**	-0.09	0.02	0.33	-0.05	-0.05
Prof	0.08	**0.54**	**0.30**	0.15	**-0.56**	0.45
Pedreg	-0.21	**-0.48**	0.29	-0.02	-0.26	0.43
lnAwsc	0.02	**0.58**	0.33	0.09	0.50	-0.04

(b) PCA Robusta.

	PC1	PC2	PC3	PC4	PC5	PC6
tAlt	**-0.34**	-0.07	0.22	-0.31	**0.67**	0.38
dcl	-0.22	0.01	**0.72**	-0.18	-0.27	-0.49
tmin	0.27	-0.16	0.14	-0.16	**-0.53**	**0.65**
tmax	**0.33**	-0.12	-0.07	0.08	0.04	-0.27
tPrec	**0.35**	-0.16	0.10	-0.09	0.15	-0.04
dprec	**-0.33**	0.15	-0.11	0.09	-0.24	0.06
prec678	**-0.45**	0.21	-0.29	0.15	-0.10	-0.00
tEvap	0.30	-0.12	0.03	-0.03	0.26	-0.12
Prof	0.29	**0.76**	0.33	0.39	0.16	0.19
Pedreg	-0.18	-0.17	**0.43**	0.14	-0.05	0.25
lnAwsc	0.09	**0.48**	-0.12	**-0.79**	-0.10	-0.05

Tabela 3.6: PCA. Contributo das variáveis por PC, com realce dos maiores contributos por PC.



(a) Gráfico de correlações.

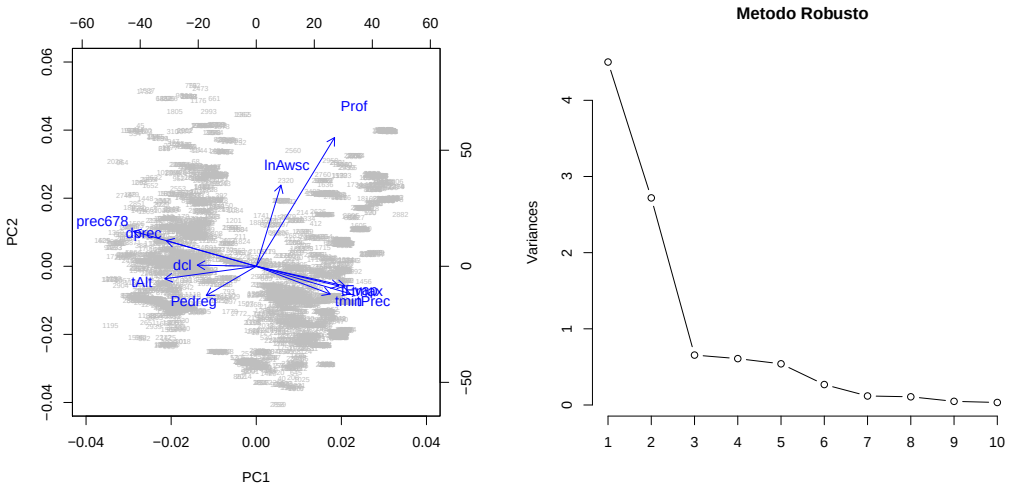
(b) Valores próprios por PC.

Figura 3.5: PCA convencional. Gráficos de análise de Componentes Principais.

pelo menos, dois grupos de dados, como se nota nos gráficos da Figura 3.5(a) e 3.6(a).

Globalmente, como se lê na Tabela 3.6, as três primeiras componentes traduzem mais de 80% da variância acumulada, quer pela abordagem convencional, quer pela abordagem robusta. Usando ainda a 4ª componente, a variância acumulada é superior a 85%.

Na primeira Componente Principal (1ª PC), a qual explica a maior parte da variabilidade nos dados, fica explicada cerca de 50% dessa variabilidade (mais concretamente, 53.1% usando a PCA convencional e 46.8% usando a PCA robusta. As variáveis com maior peso nesta componente são as climáticas, particularmente, a transformada da Precipitação média anual ($tPrec$) usando a PCA convencional e a precipitação média anual dos meses mais quentes ($prec_{678}$) usando a PCA robusta. Outras variáveis climáticas importantes, quer na PCA convencional como na robusta são: o número de dias com precipitação média superior a 1 mm ($dprec$), a Temperatura média máxima anual ($tmax$) e a Evapotranspiração média anual ($tEvap$). As variáveis $tmax$, $tPrec$ e $tEvap$ apresentam-se correlacionadas entre si e com correlação positiva. As variáveis $dprec$ e $prec_{678}$ apresentam-se correlacionadas entre si e com correlação negativa. Na 1ª PC, a diferença mais notória ao usar a PCA robusta é o facto da variável transformada da altitude ($tAlt$) ser uma das que mais contribui, o que não acontece na versão convencional. Ainda na 1ª PC o peso relativo das variáveis climáticas parece ser melhor caracterizado na versão robusta do que na versão convencional, destacando-se claramente a precipitação média anual dos meses mais quentes ($prec_{678}$) do conjunto das variáveis relevantes.



(a) Representação simultânea de variáveis e indivíduos.

(b) Valores próprios por PC.

Figura 3.6: PCA robusta. Gráficos de análise de Componentes Principais.

Admitindo a existência de 2 grupos de observações, a distinção entre os dois grupos é determinada pelo clima (ver resultados em 3.6. Assim, um dos grupos carateriza-se por apresentar elevada precipitação média anual, elevada evapotranspiração média anual, precipitações médias dos meses mais quentes mais elevadas e temperatura máxima média mais baixa. No outro grupo as mesmas variáveis apresentam comportamento inverso. A 1ªPC traduz, sobretudo, o comportamento das variáveis climáticas. Relacionando as variáveis que mais contribuem para a 1ª PC, com o *Site index*, verifica-se que a 1ª PC pode traduzir uma condição limitante ao crescimento das árvores, vulgarmente designado por *deficit hídrico*. Em termos geográficos, os dois grupos, correspondem a condições climáticas distintas entre a região norte/litoral e as regiões sul e interior, a primeira *sem deficit hídrico* e a segunda *com deficit hídrico*.

Passando à 2ª componente, que explica a maior parte da variabilidade remanescente, as variáveis com maior peso são as relacionadas com caraterísticas do solo, em particular, o logaritmo da quantidade de água no solo disponível para as plantas ($lnAwsc$), na PCA convencional, e a profundidade do solo ($Prof$) na PCA robusta. Na 2ª PC a diferença mais notória entre PCA convencional e robusta é que, enquanto na PCA convencional o peso de cada uma das variáveis de solo é idêntico, na PCA robusta destaca-se claramente a profundidade do solo ($Prof$), seguida da quantidade de água no solo disponível para a planta ($lnAwsc$), sendo que a pedregosidade do solo ($Pedreg$) é apenas relevante na PCA convencional.

Pelo que se referiu, a PCA confirma que as variáveis climáticas são variáveis relevantes no conjunto das observações, indicando mesmo a existência de grupos. Não acontece o mesmo com as variáveis relativas ao solo. No sentido de melhorar os descritores relacionados com o solo, decidiu-se considerar uma nova PCA incluindo variáveis qualitativas. Os resultados não foram suficientemente interessantes para serem incluídos no trabalho, mas ajudaram a caraterizar melhor as observações em termos de características do solo. A Figura 3.7 foi retida para ilustrar essa tentativa.

Na representação gráfica das observações, em função das variáveis categóricas, beneficiou-se de uma particularidade do função utilizada na extração das componentes principais, que é a de possibilitar a inclusão de variáveis designadas por *suplementares*, que não entram na definição das componentes, mas que é possível representá-las nos diagramas de dispersão. Foi a seguinte, a informação adicional recolhida durante com esta abordagem:

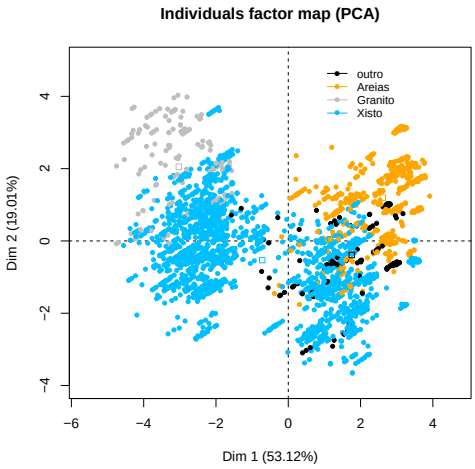
- O tipo litológico dominante são os xistos e distribuem-se de igual modo pelos dois grupos. Os grupos separam dois tipos litológicos distintos: Areias e Granitos, estando os primeiros no grupo que apresenta *deficit hídrico* e os segundos no grupo sem *deficit hídrico* 3.7(a);

- O *deficit hídrico* traduz-se por precipitações inferiores a 1000 mm separando claramente os indivíduos em dois grupos 3.7(c);

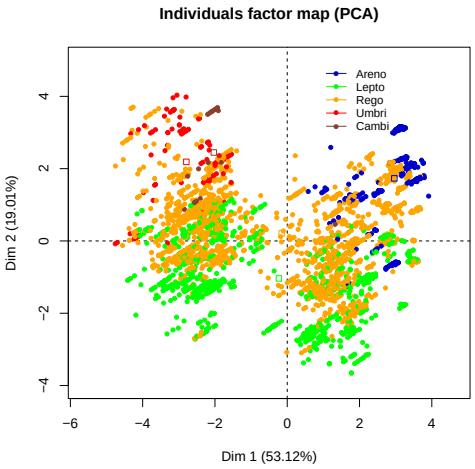
- Os tipos de solo mais frequentes são os Leptossolos e os Regossolos e distribuem-se de igual modo pelos dois grupos. Os dois grupos opõem os Arenossos dos Umbrissolos e Cambissolos, estando os primeiros no grupo com *deficit hídrico* e os segundos no grupo sem *deficit hídrico* 3.7(b);

- O carácter húmico dos solos (quantidade de matéria orgânica no solo) está predominantemente associado ao grupo sem *deficit hídrico* 3.7(d);

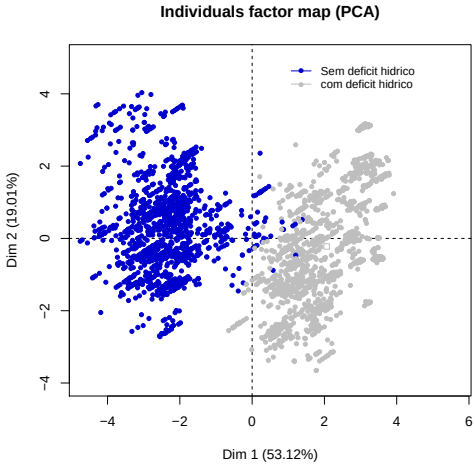
- Em termos de Pedregosidade de solo podemos verificar que, dentro de cada grupo, indivíduos com classe de pedregosidade inferior a 30% opõem-se a individuos com classes de pedregosidade maior que 65% 3.7(f).



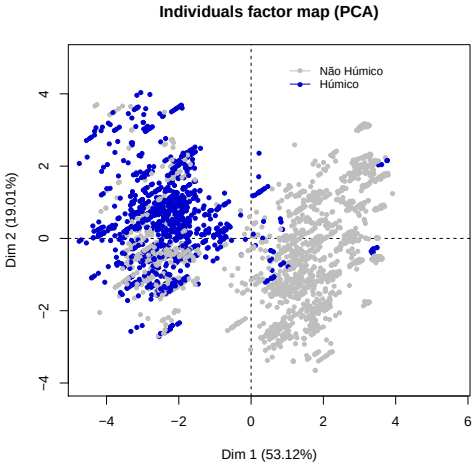
(a) Tipo de Litologia.



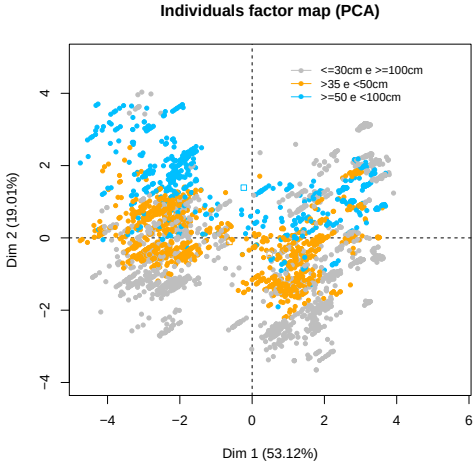
(b) Tipo de Solo.



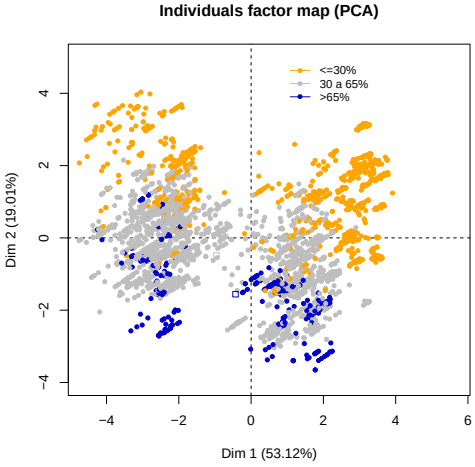
(c) Deficit hídrico.



(d) Carácter húmico do solo.



(e) Profundidade do Solo.



(f) Pedregosidade do solo.

Figura 3.7: PCA convencional. Dispersão de observações no gráfico de Componentes Principais 1 e 2, coloridos em função de diversas variáveis categóricas.

A variável exposição, seja na forma inicial (exp) ou na forma transformada exp_{pond}, foi excluída do estudo, porque o seu contributo para a explicação da variância é praticamente nulo. Tal dever-se-á à escala/método utilizado na recolha desta variável, uma vez que é amplamente reconhecido e evidente o seu efeito sobre a produtividade, pelo menos na espécie em estudo. Também se optou por excluir as variáveis longitude (x) e latitude (y), porque se verificou que a sua inclusão confunde a expressão tanto de variáveis climáticas como edáficas.

As variáveis transformada da altitude ($tAlt$) e declive (dcl) surgem nas 3ª e 4ª PCs contribuindo em menor grau para a explicação da variância total. No entanto, optou-se por mantê-las na análise, uma vez que se pretende mais à frente testar a aplicação de modelo de regressão em PCs e que nesse caso, a sua inclusão poderá ser importante. Porém, alguns autores, nomeadamente, como [14], recomendam que neste tipo de estudos não sejam utilizadas variáveis como a altitude e o declive, por, frequentemente, influenciarem os resultados de forma incorreta, precisamente devido à elevada correlação com outros fatores ambientais de clima e solo. Em muitos outros estudos relevantes, tais como em [3], as variáveis altitude e longitude são apresentadas como altamente significativas na modelação do *Site index*. Pelo que, tendo em conta os trabalhos referidos e os resultados do presente estudo, depreende-se que a relevância ou não das variáveis acimas referidas esteja relacionada com a escala espacial e espécie em estudo, tal como, descrito em [2].

Os resultados obtidos com a aplicação da PCA robusta são ligeiramente diferentes dos obtidos com a metodologia convencional. Com as mesmas variáveis, a quantidade de variância explicada até à 3ª PC aumenta ligeiramente. A variável precipitação média dos meses mais quentes ($prec_{678}$) surge neste caso, em termos de contributo para explicar a variância da 1ªPC, destacada das restantes variáveis climáticas. Também na 2ªPC a variável profundidade do solo ($Prof$) destaca-se de entre as variáveis relacionadas com o solo. A pedregosidade surge agora com maior expressão na 3ªPC onde a variável declive (dcl) é a mais expressiva. A variável quantidade de água no solo disponível para a planta ($lnAwsc$) passa da 2ªPC para a 4ªPC, e a variável transformada da altitude ($tAlt$) surge associada à temperatura média mínima na 4ªPC.

Concluindo, de um modo geral, usando a abordagem robusta, a variância parece, melhor explicada e melhor distribuída pelo conjunto das componentes que maior quantidade de variabilidade explicam.

Para além das contribuições das variáveis, procurou identificar-se observações atípicas (*outliers*). Para tal, usaram-se representações gráficas como as da Figura 3.9 que apresenta distâncias de *Mahalanobis* convencional e robusta. Foram identificadas 119 observações atípicas, o que corresponde a cerca de 4% dos dados. Essas observações mereceram atenção especial o que confirmou algumas condições particulares. A localização geográfica destas observações atípicas pode ser vista na Figura 3.8. No sul do país, estas observações correspondem a condições particulares de melhoria das condições climáticas, determinadas pela altitude do local, nomeadamente, na Serra de Monchique (no Algarve) e na Serra do Açor (no Interior Alentejano) ou a condições de solo particularmente limitantes (com solos muito pouco profundos). Na região centro, as observações atípicas localizam-se ou na região Oeste ou no Interior de Castelo-Branco e Portalegre; no primeiro caso, as observações caraterizam-se por apresentar reduzida precipitação média anual, mas bem distribuída ao longo do ano, com maior número médio de dias com precipitação superior a 1 mm; no segundo caso, correspondem a condições extremas de reduzida disponibilidade de água, associada a tipos de solo com muito pouca capacidade de reter e disponibilizar água às plantas. Na Região Norte, as observações atípicas estão também associadas a duas condições distintas, uma de temperatura média mínima muito baixa associada a elevada altitude e outra a precipitações médias anuais particularmente elevadas, frequentemente superiores a 2000 mm e com elevado número médio de dias com precipitação superior a 1 mm.

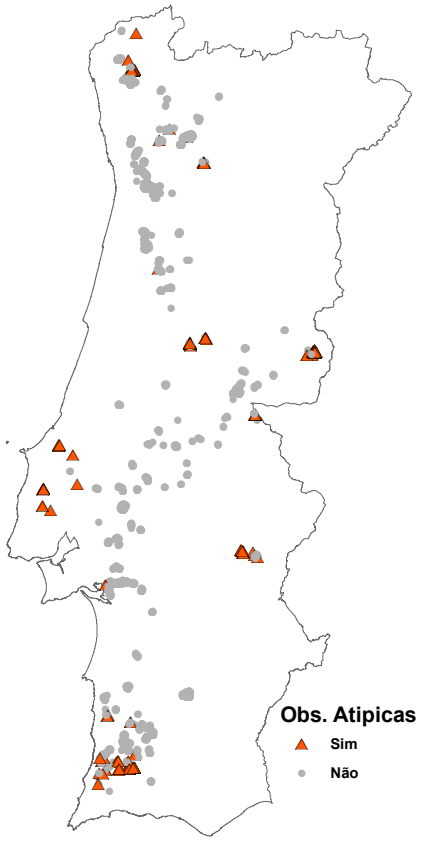


Figura 3.8: PCA robusta. Localização geográfica das observações atípicas.

A análise das observações consideradas como mais atípicas, sugeriu que essas observações representam situações reais que interessa incluir na modelação. No entanto, há que considerar que a sua exclusão poderá beneficiar a modelação do *Site index* nas restantes

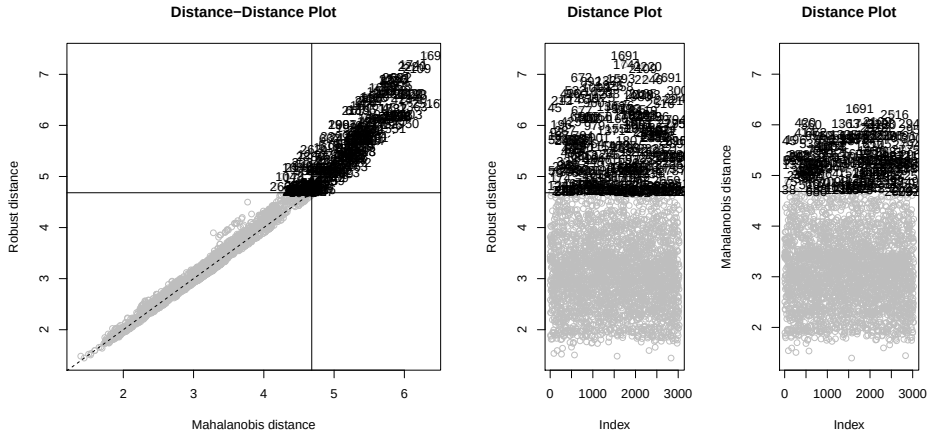


Figura 3.9: PCA. Distâncias de *Mahalanobis* convencionais e robustas.

condições ambientais.

De qualquer modo, é importante notar que a metodologia de identificação de observações atípicas funcionou bem, na medida em que foi possível confirmar a existência de condições reais específicas relativas a essas observações.

Note-se ainda que a versão robusta da distância de *Mahalanobis* coloca as observações atípicas como estando mais afastadas do restante grupo do que a versão convencional (ver a imagem à direita na Figura 3.9).

De um modo geral, a solução obtida com métodos robustos parece descrever melhor a importância das variáveis na distribuição das observações, permitindo com maior segurança a seleção de um conjunto de variáveis mais relevantes, nomeadamente: precipitação média dos meses mais quentes ($prec_{678}$); profundidade do solo ($Prof$); declive do solo (dcl); pedregosidade do solo ($Pedreg$); logaritmo da quantidade de água no solo disponível para a planta ($lnAwsc$); transformada da altitude ($tAlt$) e temperatura média mínima anual ($tmin$).

Em função dos resultados, a análise prosseguiu com uma Análise de Agrupamentos (CA).

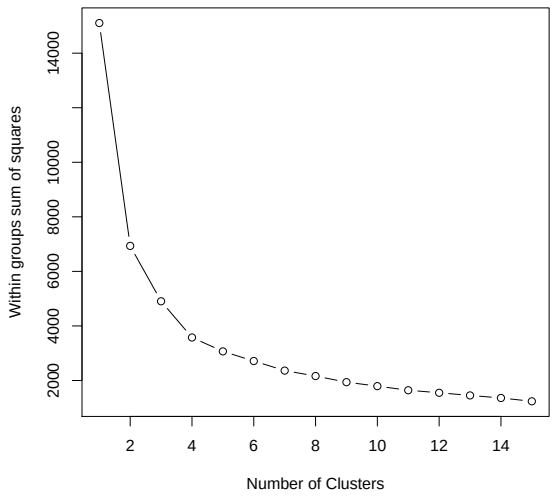


Figura 3.10: CA convencional. Método *k-means*. Número de grupos *versus* soma dos quadrados dos desvios na seleção do número de grupos.

3.2.3 Análise de Agrupamentos convencional e robusta

Uma vez que a Análise em Componentes Principais sugeriu a existência de dois grupos de observações e que esse agrupamento faz sentido do ponto de vista do conhecimento empírico, o estudo prosseguiu com uma análise de agrupamentos, de modo a confirmar a existência dos dois grupos, ou até a considerar um maior número de grupos.

Nesta secção apresentam-se os resultados obtidos considerando ambas as abordagens convencional e robusta. As variáveis que entraram na determinação dos grupos foram as anteriormente selecionadas via Análise em Componentes Principais (PCA). Usaram-se os dados estandardizados como anteriormente, uma vez que foi nessas condições que surgiu a conveniência do eventual agrupamento.

Na análise de agrupamentos (CA) convencional utilizou-se o método não hierárquico *k-means*. Este método requer a especificação prévia do número de grupos a formar (k). Apesar de haver uma predisposição para considerar dois grupos, seguiu-se a metodologia disponível para a determinação de k, produziu-se o gráfico do número de grupos *versus* a soma dos quadrados dos desvios entre grupos, visível na Figura 3.10. A soma dos quadrados dos desvios entre grupos estabiliza muito quando se consideram 4 grupos; no entanto, os resultados anteriores evidenciavam a possibilidade de agrupar os dados em apenas 2 grupos e, por outro lado, é para $k = 2$ que o decréscimo é mais significativo. Deste modo, produziram-se resultados para 2 a 4 grupos.

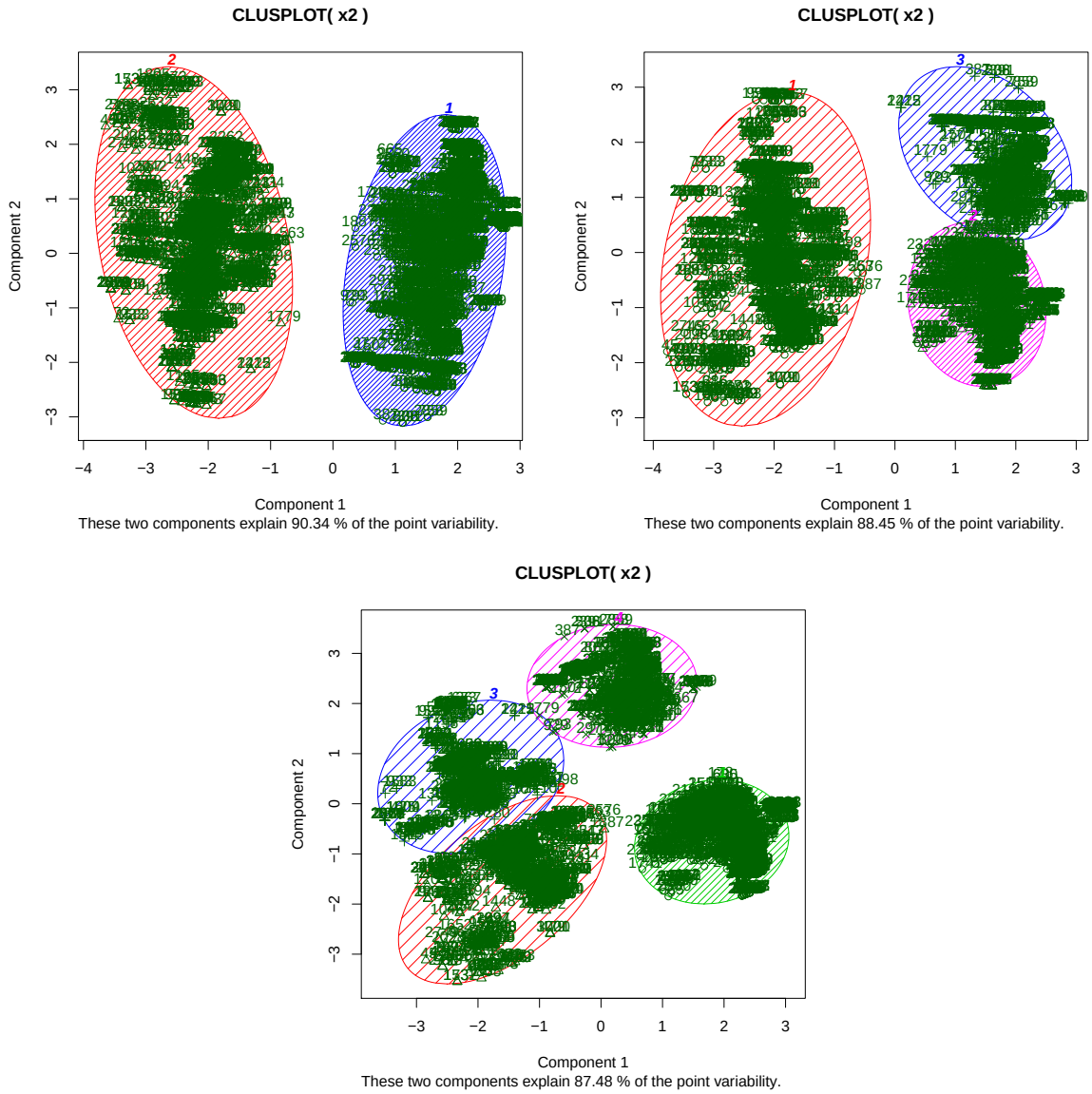


Figura 3.11: CA convencional, método *k-means*. Solução de agrupamento com $k = 2$, $k = 3$ e $k = 4$ grupos.

Os resultados obtidos com $k = 2$, $k = 3$ e $k = 4$ grupos, encontram-se na Tabela B.13 do Apêndice B. Graficamente, a Figura 3.11 ilustra o agrupamento das observações para cada caso. Os valores médios que as variáveis assumem por grupo, refletem a coerência de todos os agrupamentos. Por outro lado na figura 3.11 observa-se aumento da sobreposição entre grupos e a diminuição da variabilidade explicada com o aumento de k.

Na análise de agrupamentos robusta, como foi referido na secção 2.6, usou-se o método *Trimmed cluster*. A escolha deste método foi motivada pela existência da *package* do R *tclust* destinada a efetuar Análise de Agrupamentos robusta, bem documentada e

suportada no artigo [17]. Neste método, a alternativa robusta é desenvolvida através da identificação e inclusão, num agrupamento à parte, duma percentagem das observações mais atípicas. Deste modo, os grupos ficam melhor definidos sem observações atípicas e estas são também devidamente caraterizadas.

Este método permite efetuar diferentes agrupamentos, dependendo das estruturas de covariâncias dos grupos e do número de condição das matrizes de covariância.

Considerando o tipo de dados de cada grupo e os objetivos pretendidos utilizou-se um método que se baseia nos valores próprios da matriz de covariância (*eigen*). O fator de restrição é definido pela razão entre o máximo e mínimo valores próprios (M_n/m_n). Para além do método de agrupamento e do fator de intensidade, para usar o mesmo, é ainda necessário determinar a percentagem de observações atípicas a excluir (α) e finalmente o número de grupos a formar (k).

Assim, o primeiro passo passou por determinar um valor inicial para o fator de restrição (M_n/m_n). A escolha da solução inicial pode ser feita com base no conhecimento empírico do problema de aplicação. Não havendo tais valores disponíveis, calcularam-se os valores próprios das matrizes de covariância de cada um dos grupos formado na análise de agrupamentos convencional. Os resultados constam da Tabela B.14 no Apêndice B. Os valores máximos encontrados para o fator de restrição nos diversos agrupamentos foram de aproximadamente 43 com $k = 2$ e $k = 3$ e aproximadamente 65 com $k = 4$.

No passo seguinte produziram-se curvas CTL (Curvas de classificação da verossimilhança aparada)CTL*Classification Trimmed Likelihood Curves*, Figura 3.12. Estas curvas ajudam a determinar o número ótimo de grupos k e percentagem das observações mais atípicas a excluir (α), para determinado método de agrupamento e fator de restrição escolhido.

A aplicação do método levanta dificuldades significativas, por não estarem disponíveis valores de referência para o número de condição. Apesar em [17] o autor referir como valores razoáveis para a maioria das situações entre 5 e 10, no entanto, estes valores para o presente estudo não nos pareceram aceitáveis. Assim, verificou-se quanto menor o fator de restrição aplicado menor é o número de grupos proposto. A existência de dois grupos surge sempre como solução. Já a inclusão de 3, 4 ou até 5 grupos depende do valor do fator de restrição.

Por outro lado, de acordo com as recomendações associadas ao método, em nenhuma situação foi clara a seleção da percentagem de observações mais atípicas a excluir. De facto, as curvas CTL surgiram sempre muito paralelas entre si não se verificando a con-

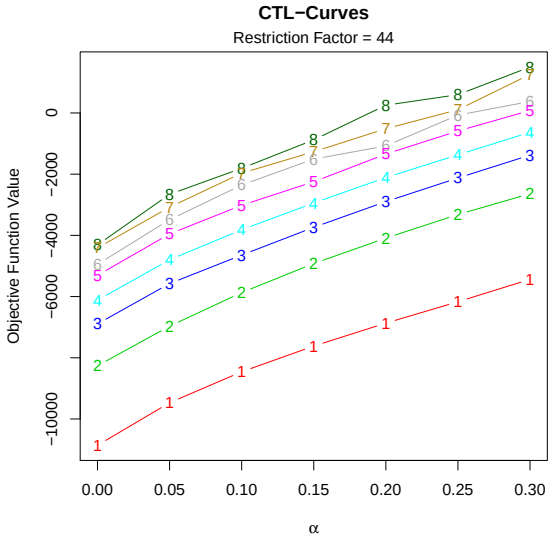


Figura 3.12: CA robusta. Método *Trimmed cluster*. Curvas CTL na seleção do número de grupos.

vergência destas para nenhuma solução de agrupamento entre 2 e 5 grupos. Como na Análise de Componentes Principais robusta (RPCA) foi possível identificar cerca de 4% de observações atíticas, decidiu-se por utilizar, também neste caso, um valor dessa ordem, pelo que se efetuou a restante análise considerando o parâmetro que representa a percentagem das observações mais atípicas como sendo 5%.

De seguida, testaram-se as soluções de agrupamento para 2, 3 e 4 grupos. Destacam-se os seguintes comentários:

1. Tanto com métodos convencionais como robustos, as soluções de agrupamento apresentadas estão de acordo com o conhecimento empírico sobre distribuição e comportamento das variáveis ambientais.

2. As soluções de agregação são muito idênticas quando se consideram 2 grupos. Já com 3 e 4 grupos, os indivíduos são agrupados de modo diferente, particularmente, no centro e sul do país, ver Figura 3.8;

3. A aplicação do método robusto revelou-se mais complicada devido à necessidade de escolher um fator de intensidade de aplicação do método de agregação (valor menor ou igual que a razão entre máximo e mínimo valor próprio M_n/m_n por grupo);

4. Tomando o parâmetro α fixo nos 5%, a distribuição das observações atípicas variou ligeiramente, insidindo principalmente nas parcelas mais a norte do território Continental, na Serra de Monchique no Algarve, na Serra D´Oça no Interior Alen-

(a) $k = 2$

	det.	VP max.	VP min.	Mn/mn
Grupo1	0.28	2.05	0.25	8.37
Grupo2	0.02	2.70	0.06	44.38

(b) $k = 3$

	det.	VP max.	VP min.	Mn/mn
Grupo1	0.03	2.53	0.07	35.39
Grupo2	0.20	2.04	0.18	11.44
Grupo3	0.09	2.39	0.06	38.20

(c) $k = 4$

	det.	VP max.	VP min.	Mn/mn
Grupo1	0.20	1.96	0.18	10.79
Grupo2	0.02	2.80	0.04	75.32
Grupo3	0.09	2.40	0.06	38.33
Grupo4	0.13	2.05	0.13	15.25

Tabela 3.7: CA Robusta, método *Trimmed Cluster*. Determinante, valores próprios máximos e mínimos e fator de restrição máximo.

tejano e um pouco por todo o Vale do Tejo desde o interior de Castelo-Branco até à Região Oeste 3.8. No Norte, essas observações, correspondem a combinações de clima e solo particularmente raras e de elevadíssima aptidão produtiva; no Vale do Tejo dever-se-á a extrema diversidade de tipos de solo de origem sedimentar (areias, entre outros) com características frequentemente peculiares. Já nas Serras Alentejanas, o efeito da Altitude será a razão principal para surgirem classificadas como observações atípicas.

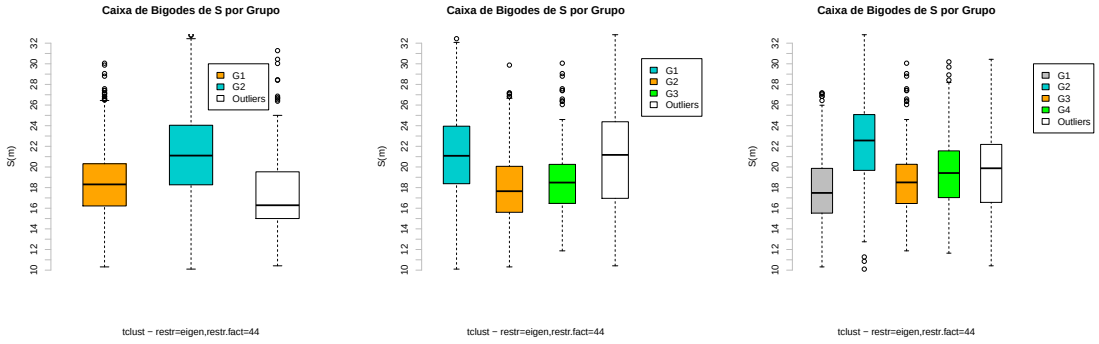
5. A reduzida frequência de algumas combinações clima-solo deverá ser a razão pela qual o algoritmo as considera como atípicas. Eventualmente, a exclusão destas observações poderá beneficiar a modelação do *Site index* para as combinações clima-solo remanescentes;

A Figura 3.8 ilustra os diferentes agrupamentos em termos geográficos, em (a) pelo método convencional e em (b) pelo método robusto.

Em termos do resultado de agrupamento, no que diz respeito aos valores médios que a variável *Site index* assume em cada grupo:

1. Em termos médios a variável *Site index* apresenta muito boa relação com o grupo produzido, em particular quando o número de grupos é de 2 ou 4, ver Tabela B.13 no Apêndice B;

2. Tal como verificado na Análise de Componentes Principais (PCA), a divisão em dois



(a) $k = 2$ (b) $k = 3$ (c) $k = 4$

Figura 3.13: CA robusta. Caixas de bigodes da variável S por grupo, para $k = 2$, $k = 3$ e $k = 4$ grupos com método *Trimmed Cluster*.

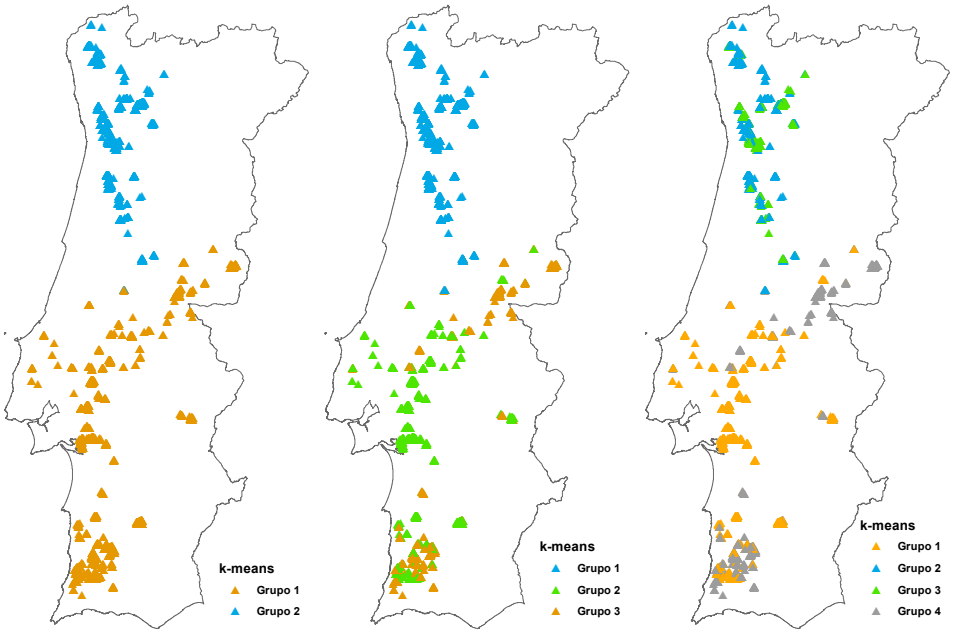
grupos é determinada pelas variáveis climáticas, em particular pela precipitação, opondo regiões com e sem *deficit hídrico*. Em termos médios a precipitação assume valores de cerca de 700 mm no grupo com *deficit hídrico* e de cerca de 1600 no grupo sem *deficit hídrico*.

3. Os 4 grupos resultam da combinação dos dois grupos climáticos anteriormente referidos, com outros dois grupos determinados por características de solo, com elevado ou reduzido volume de solo.

4. Resultados que considerem mais de 4 grupos não se traduziram em maior descriminação de indivíduos, passando a existir enorme sobreposição entre grupos. Em termos médios a variável *Site index* ainda apresenta elevada correlação com cada um dos grupos, no entanto, a nível de sobreposição entre grupos é extremamente elevado.

No seguimento deste trabalho, e de acordo com o que foi exposto, optou-se por continuar o estudo considerando apenas dois grupos. Para o efeito, o conjunto das observações incluídas em cada um dos agrupamentos foi submetido a uma nova análise exploratória para re-analisar a existência de correlações entre variáveis e as suas distribuições.

(a) Método *k-means*



(b) Método *Trimmed tclust*

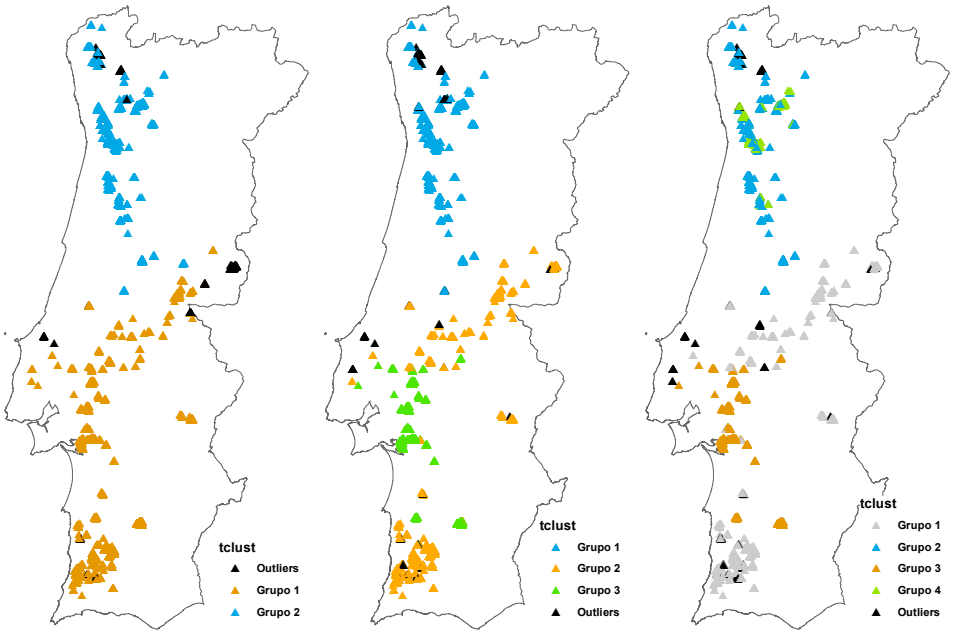


Tabela 3.8: CA Convencional e robusta. Distribuição geográfica das soluções de agregação com os métodos *k-means* e *Trimmed Cluster*, respetivamente.

3.2.4 Análise de Componentes Principais por grupo de observações

Uma vez selecionados dois grupos voltou a efetuar-se uma análise de componentes principais sobre cada um dos grupos em separado.

Previamente, para cada um dos grupos, efetuou-se uma análise exploratória, visando a verificação da nova distribuição das variáveis e a correlação entre elas. Como se referiu, os resultados obtidos levaram a que se associasse o agrupamento decorrente das secções anteriores com o *deficit hídrico*, podendo-se identificar o Grupo 1 como constituído por observações com *deficit hídrico* e o Grupo 2 por observações sem *deficit hídrico*. Assim, no Grupo 1 (**com *deficit hídrico***) a correlação das variáveis com o *Site index* piora significativamente em relação ao que tinha sido obtido sem agrupamento dos dados. Pelo contrário, no Grupo 2 (**sem *deficit hídrico***) algumas variáveis apresentam correlações com o *Site index* bastante mais elevadas. Particularmente, tal verifica-se com as variáveis relacionadas com o solo (Profundidade do solo ($Prof$) e logaritmo da quantidade de água no solo disponível para às plantas ($lnAwsc$)). Verifica-se também que este aumento de correlação é superior quando se utilizam métodos robustos. Ver o detalhe desta informação na matriz de correlações robusta relativa ao Grupo 2 na Tabela 3.9 e a matriz de correlações convencional também relativa ao Grupo 2 na Figura B.7 do Apêndice B.

Uma vez agrupadas as observações em dois grupos climáticos distintos, esperar-se-ia que a Análise de Componentes Principais (PCA) realizada dentro de cada grupo permitisse pôr em evidência as diferenças através das variáveis solo. No entanto, no subgrupo de dados **sem *deficit hídrico***, as primeiras componentes principais opõem novamente diferenças climáticas, surgindo as variáveis de solo apenas na 3ª componente principal. No subgrupo de dados **com *deficit hídrico*** as variáveis de solo surgem correlacionadas com as PC 2 e 4 e as variáveis climáticas correlacionadas com as PC 1 e 2.

De um modo geral, as correlações das variáveis com as PCs é inferior à verificada no conjunto dos *dados de calibração*, isto é, sem agrupamentos; para explicar idêntica proporção de variância são necessárias 5 a 6 PCs.

Assim, concluiu-se que não havia vantagem evidente em fazer a Análise em Componentes Principais por grupo.

	S	alt	exp	dcl	tmin	tmax	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	-0.18	-0.07	0.33	0.11	0.09	-0.16	-0.24	-0.22	0.08	0.51	-0.29	0.46	-0.11	-0.18
alt	-0.18	1.00	0.02	-0.08	-0.72	-0.62	0.67	0.01	0.62	-0.08	-0.02	0.01	-0.08	0.24	0.10
exp	-0.07	0.02	1.00	0.09	-0.01	-0.07	0.18	-0.01	0.14	0.02	-0.05	0.21	-0.11	-0.03	-0.01
dcl	0.33	-0.08	0.09	1.00	0.04	0.04	-0.04	-0.14	-0.08	0.21	0.33	0.03	0.22	-0.03	-0.16
tmin	0.11	-0.72	-0.01	0.04	1.00	0.37	-0.38	0.17	-0.43	-0.26	-0.03	-0.03	0.08	-0.69	-0.13
tmax	0.09	-0.62	-0.07	0.04	0.37	1.00	-0.83	-0.12	-0.56	0.32	-0.02	0.20	-0.13	0.37	0.18
prec	-0.16	0.67	0.18	-0.04	-0.38	-0.83	1.00	0.21	0.83	-0.28	-0.09	-0.02	-0.05	-0.20	0.04
dprec	-0.24	0.01	-0.01	-0.14	0.17	-0.12	0.21	1.00	0.40	-0.18	-0.11	0.14	-0.06	-0.04	0.47
prec678	-0.22	0.62	0.14	-0.08	-0.43	-0.56	0.83	0.40	1.00	-0.28	-0.11	0.06	-0.13	0.14	0.55
evap	0.08	-0.08	0.02	0.21	-0.26	0.32	-0.28	-0.18	-0.28	1.00	0.00	0.27	-0.05	0.53	-0.27
Prof	0.51	-0.02	-0.05	0.33	-0.03	-0.02	-0.09	-0.11	-0.11	0.00	1.00	-0.36	0.84	-0.01	-0.07
Pedreg	-0.29	0.01	0.21	0.03	-0.03	0.20	-0.02	0.14	0.06	0.27	-0.36	1.00	-0.71	0.22	0.14
awsc	0.46	-0.08	-0.11	0.22	0.08	-0.13	-0.05	-0.06	-0.13	-0.05	0.84	-0.71	1.00	-0.20	-0.20
x	-0.11	0.24	-0.03	-0.03	-0.69	0.37	-0.20	-0.04	0.14	0.53	-0.01	0.22	-0.20	1.00	0.43
y	-0.18	0.10	-0.01	-0.16	-0.13	0.18	0.04	0.47	0.55	-0.27	-0.07	0.14	-0.20	0.43	1.00

Tabela 3.9: Matriz de Correlações (*Fast MCD*). Grupo de dados *sem deficit hídrico* (com 1350 observações).

3.3 Regressão Linear Múltipla convencional e robusta

Da análise realizada nas secções anteriores, para o conjunto de *dados de calibração*, resultou que: a variável *Site index* apresenta muito baixa correlação linear com as restantes variáveis ambientais; há um conjunto de variáveis ambientais que explicam a grande maioria da variância dos dados, e cuja combinação se traduz em agrupamentos de observações que traduzem fatores importantes.

Os fatores identificados como relevantes são o *Deficit hídrico* e o *volume de solo* (que se traduz em quantidade de água armazenada no solo disponível para a planta). No fator *Deficit hídrico* a variável com maior peso é a precipitação média anual dos meses mais quentes ($prec_{678}$), na análise robusta, e a transformada da precipitação média anual ($tPrec$), no caso da análise convencional. No fator *volume de solo*, é a variável profundidade do solo ($Prof$) que tem maior peso no caso da análise robusta e a variável logaritmo da quantidade de água no solo disponível para a planta ($lnAwsc$) no caso da análise convencional. Embora com menor peso, um conjunto de outras variáveis contribuem ainda para a explicação da variância total, nomeadamente: a transformada da altitude ($tAlt$), a temperatura média máxima anual ($tmax$), o declive do solo (dcl).

O fator *Deficit hídrico* divide, claramente, as observações em dois grupos. Sobre cada um destes dois grupos foi novamente realizada uma análise exploratória, resultando daí que, no Grupo 2 (sem *deficit hídrico*) as correlações das variáveis ambientais com a variável *Site index* aumentaram, assim como a variância explicada pelo conjunto das variáveis. Deste modo, para o Grupo 2 registaram-se aumentos significativos das correlações do *Site index* em relação às variáveis solo, particularmente a profundidade do solo (51%) e quantidade de água no solo disponível para as plantas (46%). Estes resultados encontram-se na Tabela 3.9 da Secção 3.2.4 para análise robusta e Tabela B.7 do Apêndice B para análise convencional. Como não se obtiveram conclusões análogas para o Grupo 1 (com *deficit hídrico*), decidiu-se investigar a modelação da dependência do *Site index* em relação às variáveis ambientais, considerando todo o conjunto de *dados de calibração* e seguidamente, considerando cada um dos grupos. Deste modo, admitiu-se a possibilidade de ser vantajoso ter 2 ou 3 modelos de regressão distintos, dependendo do grupo de observações.

Com o objetivo estrito de comparar resultados obtidos pelos diferentes métodos, e apesar da elevada correlação entre diversas variáveis ambientais, decidiu-se começar por fazer a análise de regressão nos termos convencionais, recorrendo à regressão passo a passo

(*stepwise regression*) para a seleção dos regressores.

Foi testado o comportamento desses modelos com e sem estandardização dos dados, sendo as conclusões semelhantes, quer ao nível da análise de resíduos, quer na avaliação da *performance* do modelo, pelo se optou por apresentar os resultados com valores não estandardizados, devido a serem de mais fácil interpretação.

Para além disso, testaram-se 3 tipos de modelos iniciais: um só incluindo variáveis quantitativas, outro incluindo variáveis quantitativas e categóricas (através de variáveis *mudas*) e um outro com transformações polinomiais das variáveis quantitativas e com variáveis categóricas. Este último cenário foi motivado pelas baixas correlações do *Site index* relativamente aos regressores, o que sugeriu que a dependência fosse melhor traduzida por termos não lineares nas variáveis ambientais.

Depois de realizada uma análise de regressão prévia dos modelos que includiram as variáveis *mudas* e/ou transformadas verificou-se que a distribuição dos resíduos se afastava fortemente da distribuição *Normal*. Assim, não se verificavam as condições que permitissem considerar credíveis os resultados encontrados na seleção de variáveis.

Assim, um modelo que inclui apenas as variáveis quantitativas foi o escolhido, formalizado pela equação 3.1.

$$\begin{aligned} S_i = \beta_0 &+ \beta_1 \times prec_i^{-0.44} + \beta_2 \times prec_{678i} + \beta_3 \times dprec_i + \beta_4 \times tmin_i + \\ &\beta_5 \times tmax_i + \beta_6 \times evap_i^{0.18} + \beta_7 \times alt_i^{0.39} + \beta_8 \times dcl_i + \\ &\beta_9 \times \ln(awsc)_i + \beta_{10} \times Pedreg_i + \beta_{11} \times Prof_i + \\ &\beta_{12} \times x_i + \beta_{13} \times y_i + \epsilon_i \end{aligned} \tag{3.1}$$

onde os ϵ_i devem ter esperança nula, variância constante e ser não correlacionadas para diferentes índices.
Em geral, os regressores foram já identificados. Os 4 regressores que envolvem transformação de variáveis decorrem de recomendação de literatura específica (ver, por exemplo [18]), nomeadamente o *lnAwsc* e as potências *prec*, *evap* e *alt* resultantes de transformação de *Box-Cox*.

Depois da seleção dos regressores, com a metodologia *stepwise regression*, o modelo obtido para o conjunto dos *dados de calibração*, sugere a exclusão das seguintes variáveis: precipitação média anual *prec* na sua forma transformada; número médio de dias com precipitação superior a 1 mm *dprec*; temperatura média mínima anual (*tmin*) e a Pro-

fundidade do solo $Prof$. Mas, infelizmente o erro padrão dos resíduos foi enorme (é de 3.1) e o coeficiente de determinação (R^2) foi de 0.33.

Com este último conjunto de regressores significativos foram validados os pressupostos do modelo.

Os resíduos não rejeitaram a distribuição *Normal* (com *p-value*=0.08 para o teste de *Lilliefors K-S*) e, analisados em termos geográficos, não revelaram qualquer tendência.

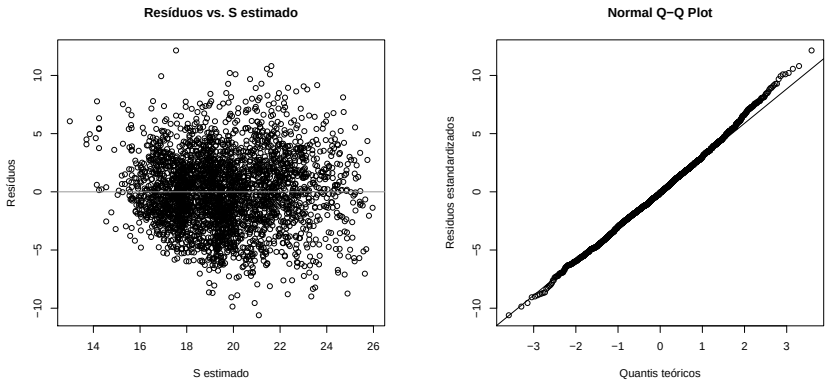
A análise sumária dos resíduos permitiu verificar que na zona central da distribuição (para valores medidos do *Site index* entre 17 e 22), os resíduos não são irrealistas pois apresentam resíduo médio inferior a 2 metros. Quanto aos pressupostos de variância constante dos erros do modelo não foi rejeitada essa hipótese apesar da Figura 3.14(a) levantar algumas dúvidas.

Na Equação 3.2 apresenta-se a equação de regressão estimada com o melhor modelo de regressão linear encontrado pelo método convencional.

$$\begin{aligned} E[S] = {} & 56.9 + 0.294 \times alt^{0.39} + 0.101 \times dcl + 0.410 \times tmax \\ & - 0.039 \times prec_{678} - 27.870 \times evap^{0.18} \\ & - 0.04 \times Pedreg + 1.661 \times \ln awsc + 0.00001 \times y \end{aligned} \tag{3.2}$$

De seguida, usou-se o mesmo modelo, com os mesmos regressores iniciais, estimando os coeficientes para o subconjunto de dados na região **com *deficit hídrico***, correspondente ao Grupo 1. Também se verificaram os pressupostos do modelo com conclusões semelhantes às anteriores. O erro padrão dos resíduos foi de 2.7, e o coeficiente de determinação foi de 0.22. Os resíduos não rejeitam a distribuição *Normal* (com *p-value* igual a 0.28). Neste caso as variáveis quantitativas significativas foram: a altitude (alt) do local na sua forma transformada; a temperatura média máxima ($tmax$); a evapotranspiração($evap$) na sua forma transformada; a pedregosidade do solo ($Pedreg$); a quantidade de água no solo disponível para a planta ($lnAwsc$) na sua forma logarítmica e ainda as variáveis latitude e longitude do local.

Para o subconjunto de dados na região **sem *deficit hídrico***, correspondente ao Grupo 2, a validação dos pressupostos obteve-se idênticas conclusões. O erro padrão dos resí-



(a) Resíduos vs. S estimado (b) Normal Q-Q plot

Figura 3.14: MLR convencional. Gráficos de Diagnóstico para dados de calibração.

duos foi de 3.2 e o coeficiente de determinação foi de 0.35. Os resíduos não rejeitam a distribuição *Normal* (com *p-value* igual a 0.10). Neste caso, as variáveis quantitativas significativas foram: a altitude do local (alt) na sua forma transformada; o declive do local (dcl); a temperatura média máxima ($tmax$); a precipitação média anual ($prec$) na sua forma transformada; o número de dias médio com precipitação superior a 1mm ($dprec$); a precipitação média dos meses mais quentes ($prec_{678}$); a profundidade do solo ($Prof$); a pedregosidade do solo ($Pedreg$) e a variável latitude do local (y).

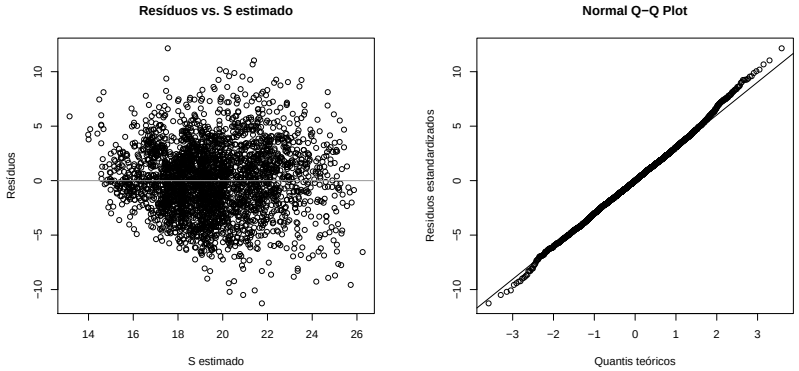
Na Figura 3.14 apresentam-se os gráficos de diagnóstico associados ao melhor modelo de regressão convencional obtido.

Resumindo, com a regressão convencional, e usando a regressão passo-a-passo, o conjunto de regressores significativos não foi coincidente nos dois grupos, nem entre estes e o conjunto total dos dados. Para além disso, o melhor coeficiente de determinação foi encontrado no Grupo 2, com $R^2 = 0.35$, mas acompanhado de um erro padrão de regressão demasiado elevado.

Numa tentativa de melhorar a modelação, de seguida usaram-se métodos robustos para estimar os coeficientes de regressão.

Com a aplicação de estimadores robusto conseguiu-se apenas uma melhoria ligeira do erro padrão da estimativa. Ao contrário do que seria de esperar, não foram detetadas quaisquer observações atípicas quer se consider o conjunto dos *dados de calibração* quer os dados de cada um dos grupos.

No modelo estimado com o conjunto de ***dados de calibração***, aplicando a versão



(a) Resíduos vs. S estimado (b) Normal Q-Q plot

Figura 3.15: MLR robusta. Gráficos de Diagnóstico para dados de calibração.

robusta, o erro padrão dos resíduos foi enorme 3.0, o coeficiente de determinação manteve-se pelos 0.32. Neste caso, as variáveis quantitativas significativas foram: a altitude do local (alt) na sua forma transformada; o declive do local (dcl); a temperatura média máxima ($tmax$); a evapotranspiração média anual ($evap$) na sua forma transformada; a pedregosidade do solo ($Pedreg$) e a quantidade de água no solo disponível para a planta ($awsc$) na sua forma logarítmica.

Para o subconjunto de dados na região **com *deficit hídrico*** (Grupo 2), aplicando a versão robusta da MLR, o desvio padrão dos resíduos é de 2.7, o coeficiente de determinação é de 0.20. Neste caso as variáveis quantitativas mais significativas são: a altitude do local (alt) na sua forma transformada; o declive do local (dcl); a temperatura média máxima($tmax$); o número de dias médio com precipitação superior a 1 mm ($dprec$), a evapotranspiração média anual ($evap$) na sua forma transformada; a pedregosidade do solo ($Pedreg$) e a quantidade de água no solo disponível para a planta ($awsc$) na sua forma logarítmica.

No modelo obtido, para o subconjunto de dados na região **sem *deficit hídrico***, aplicando a versão robusta da MLR, o desvio padrão dos resíduos é de 3.1, o coeficiente de determinação é de 0.34.

Na Equação 3.3 apresenta-se o melhor modelo obtido com aplicação da MLR robusta. E na Figura 3.15 encontram-se os gráficos de diagnóstico associados a este modelo.

$$\begin{aligned} E[S] = 51.7 + 0.286 \times alt^{0.39} + 0.107 \times dcl + 0.543 \times tmax - \\ 26.3 \times evap^{0.18} - 0.042 \times Pedreg + 1.549 \times \ln awsc \end{aligned} \tag{3.3}$$

As variáveis que surgem sempre como altamente significativas em qualquer modelo testado são: a evapotranspiração na sua forma transformada ($tEvap$); a temperatura média máxima ($tmax$); a quantidade de água no solo disponível para a planta na forma logaritmica ($lnAwsc$); a Pedregosidade do solo ($Pedreg$); a altitude na sua forma transformada ($tAlt$) e declive do local (dcl). De entre estas as com maior peso na modelação do *Site index* são: a evapotranspiração ($tEvap$) e a quantidade de água no solo disponível para a planta ($lnAwsc$). A variável temperatura mínima média anual foi sempre excluída do modelo final. A latitude e a longitude do local foram sempre excluídas ao usar estimadores robustos.

Na predição do *Site index*, a variável climática com maior expressão na MLR é a evapotranspiração e a variável de solo com maior peso é a quantidade de água no solo disponível para a planta. De referir que, nos resultados obtidos na PCA, a precipitação média anual nos meses mais quentes e a profundidade do solo eram as variáveis mais importantes na caraterização da distribuição dos indivíduos.

3.4 Regressão em Componentes Principais convencional e robusta

Na secção 2.5 chamou-se a atenção para as vantagens que podem existir ao considerar a regressão linear em componentes principais (PCR), quando o conjunto de potenciais regressores contém variáveis correlacionadas entre si. Na presente secção apresentam-se os resultados obtidos com esse tipo de modelação.

Tendo em conta a importância dos processos computacionais na aplicação dos diversos métodos, procurou-se preferencialmente efetuar os cálculos recorrendo a *packages* já testadas e divulgadas. Não existindo uma *package* do programa R específica para a realização da regressão robusta em Componentes Principais (RPCR), e de modo a viabilizar a comparação entre os resultados encontrados pelo método convencional e pelo robusto, optou-se por implementar a mesma metodologia, tanto na aplicação da PCR como na da RPCR.

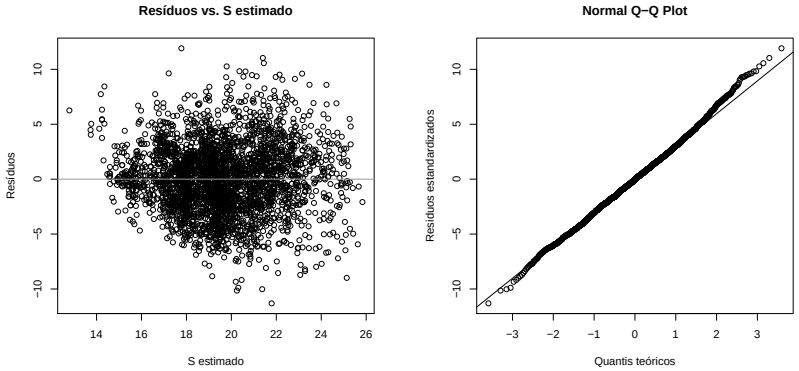
Assim, a metodologia implementada na PCR foi a seguinte:

1. Foram geradas as componentes principais com base na função *PCA* da *package* *FactoMineR*, a qual realiza Análise de Componentes Principais (PCA) utilizando a matriz de covariância convencional, tal como foi descrito na secção 3.2.2.
2. Aplicou-se a Regressão Linear Múltipla Robusta (MLR) sobre as PCs derivadas na alínea anterior.

Do mesmo modo, a metodologia implementada na RPCR foi a seguinte:

1. Foram geradas as componentes principais com base na função *PcaCov* da *package* do R *rrcov*, que realiza Análise de Componentes Principais (PCA) utilizando uma matriz de covariância robusta e tendo por base o estimador *Fast MCD*, de acordo com a secção 3.2.2.
2. Aplicou-se a Regressão Linear Múltipla Robusta (RMLR) sobre as PCs derivadas na alínea anterior. Na aplicação da RMLR utilizou-se a função *lmrob* da *package* do R *robustbase*.

A metodologia acima descrita foi aplicada nos três conjuntos de dados, nomeadamente, no conjunto de *dados de calibração* (com 3022 observações) e nos dois subconjuntos deste, que resultaram da Análise de Agrupamentos (CA) descrita na Secção 3.2.3, um deles representando a região com *deficit hídrico*, e o outro representando a região sem *deficit hídrico* (respetivamente, com 1672 e 1350 observações).



(a) Resíduos vs. S estimado (b) Normal Q-Q plot

Figura 3.16: PCR convencional. Gráficos de diagnóstico para os dados de calibração.

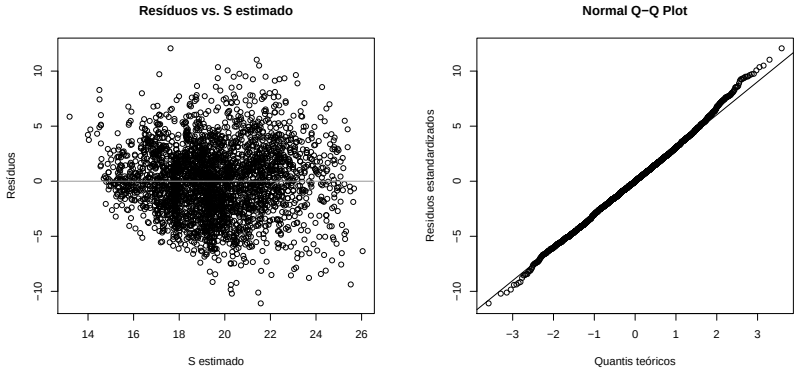
Foram aplicadas as duas metodologias em dados estandardizados e não estandardizados. Tendo-se obtido idênticas conclusões, optou-se por apresentar os resultados produzidos com dados não estandardizados, uma vez que facilita a interpretação dos resultados.

Na aplicação da Regressão em Componentes Principais (PCR), o número de componentes principais começou por ser de 11, 9 e 9, respetivamente, para o total dos dados de calibração, dos dados na região com *deficit hídrico* e dos dados na região sem *deficit hídrico*.

Em termos de erro padrão dos resíduos, para aqueles conjuntos de dados, os resultados foram, respetivamente, de 3.1, 2.7 e 3.3. e, portanto, nada satisfatórios. No entanto, da análise de resíduos (ver Tabela B.16 no Apêndice B), resulta ainda que cerca de 50% dos dados apresentam resíduos inferiores a cerca de 2 metros, o que é uma boa estimativa; mas, para os restantes, as estimativas são em média muito superiores, chegando mesmo a ser superiores a 10 metros - o que é inaceitável!

Note-se que foi efetuado a validação de pressupostos do modelo, conforme se ilustra com a Figura 3.16 (para os *dados de calibração*). Embora o gráfico 3.16(a) possa sugerir variância não constante, considerou-se não haver motivos sérios para rejeitar esse pressuposto (apesar de não ter sido feito nenhum teste formal). Também não houve motivo para rejeitar a distribuição *Normal*.

Na aplicação da Regressão Robusta em Componentes Principais (RPCR), o número de componentes principais selecionadas foi de 7, 6 e 5, respetivamente, para o total dos



(a) Resíduos vs. S estimado (b) Normal Q-Q plot

Figura 3.17: PCR robusta. Gráficos de diagnóstico para os dados de calibração.

dados de calibração, dos dados na região com *deficit hídrico* e dos dados na região sem *deficit hídrico*; pelo que se destaca, desde já, a diminuição do número de regressores.
Infelizmente, ao utilizar o método robusto (que, à partida é menos sensível ao afastamento das premissas do modelo), persistiram os maus resultados. Com efeito, em termos de erro padrão dos resíduos e para os três conjuntos de dados, obtiveram-se resultados muito semelhantes aos que se tinham encontrado com a análise convencional (respetivamente, de 3.1, 2.6 e 3.1).
Embora tenha deixado de existir a mesma preocupação com a validação dos pressupostos do modelo, fez-se uma análise de diagnóstico como usualmente e verificou-se que não havia motivos para rejeitar a distribuição *Normal*, em nenhum dos grupos. Na Figura 3.17 apresentam-se gráficos de dispersão de resíduos e de quantis (para a *Normal*) obtidos com regressão robusta para os *dados de calibração*. Para os dados agrupados obtiveram-se conclusões análogas e nenhum dos casos foi rejeitada a distribuição *Normal* dos resíduos.

Em termos de contributo das variáveis para cada componente principal (PC), no modelo para o conjunto dos *dados de calibração*, as 7 PCs selecionadas foram a PC_1, a PC_2, a PC_4, a PC_6, a PC_8, a PC_9 e a PC_{10}. Na PC_1 a variável com maior peso foi a precipitação média dos meses mais quentes ($prec_{678}$); na PC_2 as duas variáveis com maior expressão são a profundidade do solo ($Prof$) e a pedregosidade dos solo ($Pedreg$).

No conjunto de dados na região **com *deficit hídrico***, foram selecionadas 6 PCs. As variáveis com maior peso na PC_1, que explica a maior proporção da variância, são a profundidade do solo ($Prof$) e a pedregosidade do solo ($Pedreg$), ao contrário da análise dos *dados de calibração*, em que estas variáveis apenas aparecem na 2ª PC. Outras variáveis

com expressão neste conjunto de dados foram, por ordem decrescente da ordem com que entram nas PCs e da variância que traduzem: o número de dias médio com precipitação inferior a 1 mm ($dprec$), a altitude (alt), a quantidade de água no solo disponível para a planta ($awsc$) e a evapotranspiração ($evap$).

No conjunto de dados na região **sem *deficit hídrico***, foram selecionadas 5 PCs. As variáveis com maior peso na PC_1 que explica a maioria da variância são a profundidade do solo ($Prof$) e a pedregosidade do solo ($Pedreg$); a variável com maior peso na PC_2 é a precipitação média dos meses mais quentes ($prec_{678}$), notando-se uma inversão de importância entre variáveis climáticas e relativas ao solo, em comparação com o conjunto dos *dados de calibração*. Outras variáveis com expressão neste conjunto de dados foram, por ordem decrescente de importância: o declive do solo (dcl), a altitude (alt), a quantidade de água no solo disponível para a planta ($awsc$) e a temperatura média máxima ($tmax$).

Concluindo, a principal diferença na aplicação de método convencional e do robusto, está no menor número de PCs necessárias para explicar idêntica proporção de variância, e mantendo o erro padrão dos resíduos muito similar nos dois casos.

3.5 Avaliação do desempenho dos diferentes métodos

Anteriormente já se referiu que um subconjunto de dados foi inicialmente reservado para efeitos de validação e comparação do desempenho dos modelos estudados. Relembra-se que o subconjunto de dados foi obtido por seleção aleatória de 10% do total das observações iniciais, o que corresponde a 336 observações. Esse conjunto de dados designou-se por *dados de validação* e, até agora, nunca foi utilizado para estimar parâmetros, nem na análise das componentes principais, nem na regressão.

Uma vez que se concluiu haver motivos para considerar os dados de calibração agrupados, e que se estimou um modelo de regressão em componentes principais para cada um dos grupos, houve necessidade de começar por classificar cada uma das observações do subconjunto de validação, de modo a saber em que grupo deviam ser classificadas (Grupo 1 - *sem défícit hídrico*, ou Grupo 2 - *com défícit hídrico*), de modo a investigar a capacidade de predição do *Site index* de cada um dos modelos estimados, quando aplicados a indivíduos do correspondente grupo.

Como se referiu no final da Secção 2.3, o discriminante linear de Fisher só poderia ser usado na classificação, se fosse possível assumir que as observações de ambos os grupos provinham de populações com igual matriz de covariâncias. Este cenário não era razoável para a situação em estudo, desde logo porque se verificou que, na PCA, a estrutura de dependência entre variáveis era diferente nos dois grupos; as variáveis que mais contribuíam para a explicação da variância (e para a definição das PCs) não foram coincidentes nos dois grupos. Ao procurar utilizar como alternativa o discriminante quadrático, foi evidente que o pressuposto da distribuição Normal de cada grupo também estava longe de poder ser assumido. Logo, ao efetuar essa tarefa, teve-se presente que não estavam reunidas as condições necessárias para uma boa classificação. Existindo a possibilidade de que os métodos robustos atenuassem o efeito de afastamento dos pressupostos, decidiu-se optar pela violação das hipóteses de distribuição Normal e prosseguir com a classificação determinada pela discriminante quadrática.

Assim, as observações do subconjunto de validação foram classificadas como pertencentes ao Grupo 1 (*sem deficit hídrico*) ou ao Grupo 2 (*com deficit hídrico*) com base em análise quadrática discriminante (QDA), conforme a metodologia explicada na Secção 2.3.

Tratando-se de dados reais, é possível avaliar até que ponto é que os modelos esti-

mados têm a capacidade de predizer o *Site index*, comparando, para cada modelo fixo, o valor predito com o valor observado na realidade. Para além disso, podem comparar-se diferentes métodos, no sentido em que o melhor modelo será aquele que conduz a menor afastamento entre as predições e os valores observados. A medida de afastamento utilizada de acordo com este critério, foi a raiz quadrada do erro quadrático médio ($RMSE$), definido por

$$RMSE = \sqrt{\frac{1}{336}\sum_i (S_i - \hat{S}_i)^2},$$

onde S_i e $\hat{S}_i$ denotam, respetivamente, os valores observados e os valores preditos do *Site index* para os correspondentes indivíduos, de acordo com cada método particular.

Por outro lado, a maior parte das publicações científicas sobre esta matéria usa como critério o coeficiente de determinação R^2, já referido em secções anteriores. Para além disso, como se explicou na Secção 2.4, num modelo de regressão linear, o coeficiente de determinação traduz a proporção de variabilidade que é explicada pelo modelo, com o conjunto de regressores que o define.

Assim, as comparações e conclusões foram estabelecidas com base em dois critérios: no valor do coeficiente de determinação e na raiz quadrada do erro quadrático médio (amostral).

Na Tabela 3.10 apresentam-se valores do $RMSE$ para diferentes modelos testados. Estes

	MLR	RMLR	PCR	RPCR
Mod_{Calib}	3.1	3.2	3.2	4.4
Mod_{G1}	2.9	2.9	3.2	3.2
Mod_{G2}	3.9	3.4	3.6	1.1

Tabela 3.10: Valores do RMSE para os modelos estimados por cada um dos métodos: Regressão linear convencional (MLR), Regressão linear robusta (RMLR), Regressão linear em componentes principais (PCR) e Regressão linear robusta em componentes principais (RPCR); e usando cada um dos conjuntos de dados: dados de calibração (Mod_{Calib}), dados com *deficit hídrico* (Mod_{G1}) e dados sem *deficit hídrico* (Mod_{G2}).

modelos são os resultantes do estudo apresentado em secções anteriores e são denotados na tabela por Mod_{calib}, Mod_{G1} e Mod_{G2}, correspondentes aos modelos estimados com base, respetivamente, no conjunto dos *dados de calibração*, no conjunto de dados *com déficit hídrico* (Grupo 1) e no conjunto de dados *sem déficit hídrico* (Grupo 2). Para cada modelo são apresentados quatro resultados, que indicam o valor do $RMSE$ obtido com cada um dos métodos de regressão: a regressão linear na versão convencional (MLR) e

na versão robusta (RMLR) e a regressão linear em componentes principais, também nas versões convencional (PCR) e robusta (RPCR).

Da análise da Tabela 3.10 pode concluir-se que, atendendo a este critério, os diferentes métodos têm desempenhos semelhantes e pouco satisfatórios, com a exceção da regressão robusta em componentes principais, no Grupo 2. Mas este método também produz maus resultados nos outros conjuntos de dados (indiferenciados ou do Grupo 1).
No que diz respeito aos conjuntos de dados, para o Grupo 1 não se nota nenhuma vantagem relevante em ter havido o agrupamento.
Globalmente, os melhores resultados são encontrados para o Grupo 2 - o dos indivíduos sem *déficit hídrico*, conjuntamente com a regressão robusta em componentes principais.

Por outro lado, o coeficiente de determinação é o indicador mais divulgado para avaliar a proporção de variabilidade explicada pelo conjunto dos regressores. Por isso, foi também considerado nesta avaliação. Este indicador é frequentemente usado na literatura específica sobre o assunto e interessa analisá-lo, na medida em que existem publicações científicas que dão valores de referência.
Desde já é importante realçar que em artigos como, por ex., [4] e [3], os melhores valores de coeficiente de determinação encontrados na modelação, utilizando técnicas estatísticas idênticas, estão situados na ordem dos 25% a 47%. Em artigos como, por ex., [36] e [18] os valores de coeficiente de determinação rondam os 0.60, mas parece ser um caso excecional. Uma outra situação de exceção é apontada em [15], que refere um modelo em que obteve um coeficiente de determinação de 0.90; no entanto, neste último caso, as variáveis ambientais consideradas são de muito difícil e dispendiosa recolha, pelo que não servem para sustentar um modelo que se deseja de fácil aplicação.
De referir ainda que nos artigos anteriormente citados com valores excecionais, a região, a espécie e as variáveis ambientais consideradas não são semelhantes às deste estudo.

	MLR	RMLR	PCR	RPCR
Mod_{Calib}	0.32	0.31	0.31	0.005
Mod_{G1}	0.17	0.10	0.01	0.01
Mod_{G2}	0.24	0.34	0.18	0.03

Tabela 3.11: Valores do R^2 para os modelos estimados por cada um dos métodos: Regressão linear convencional (MLR), Regressão linear robusta (RMLR), Regressão linear em componentes principais (PCR) e Regressão linear robusta em componentes principais (RPCR); e usando cada um dos conjuntos de dados: dados de calibração (Mod_{Calib}), dados com *deficit hídrico* (Mod_{G1}) e dados sem *deficit hídrico* (Mod_{G2}).

Na Tabela 3.11 apresentam-se os resultados organizados de modo idêntico ao que se fez em relação ao indicador $RMSE$.
Analisando a tabela, verifica-se que os valores do coeficiente de determinação são fracos ou muito maus para as diferentes situações consideradas, e que se situam entre 0.03 e 0.34.

A região a que refere o Mod_{G1} carateriza-se por elevado *deficit hídrico*, pelo que, as condições ambientais em presença são muito restritivas para o desenvolvimento das plantas. Os dados de produtividade nesta região, expressa pelo *Site index*, são muito reduzidos e apresentam pouca variabilidade. Já na região a que se refere o Mod_{G2} não existe *deficit hídrico* pelo que as plantas crescem sem o fator principal de limitação ao seu crescimento. Deste modo, podem expressar todo o seu potencial em diferentes condições ambientais. Por este facto, nesta região, os valores de produtividade, de igual modo expressa pelo *Site index*, são superiores e apresentam maior variabilidade. Os resultados diferenciados podem deste modo refletir o o comportamento diferenciado dos povoamentos florestais da *Eucalyptus globulus* em condições de presença ou ausência de fator fortemente limitante ao seu crescimento.

Em termos de verificar a capacidade de predição do modelo foram ainda utilizados critérios "ecológicos de interpretabilidade", de modo a analisar o grau pelo qual o modelo incorpora a relativa importância das variáveis regressoras e de que modo o *Site index* se altera com a alteração de cada variável regressora.

Assim, de um modo geral os regressores incluídos nos modelos testados traduzem o comportamento previsível da variável $SiteIndex$. Há no entanto duas exceções, nomeadamente: a temperatura média máxima anual ($tmax$) e a precipitação média dos meses mais quentes ($prec_{678}$). De facto, estas duas variáveis surgem com sinal contrário ao que seria de esperar, com o *Site index* a aumentar com o aumento da temperatura máxima e a diminuir com a diminuição da precipitação média dos meses mais quentes, apesar de o coeficiente de regressão que lhes está associado diminuir o peso da variável no modelo de regressão.

Verifica-se também que o *Site index* aumenta com o aumento da altitude e do declive (dcl). Esta relação dever-se-á ao facto de a região Norte, com mais precipitação e melhores solos, apresentar relevo mais acidentado, com altitudes e declives médios superiores às da

região sul.

As restantes relações com as variáveis regressoras estão de acordo com o expectável, em particular, com o *Site index* a aumentar com a quantidade de água no solo disponível para as plantas e a diminuir com o aumento da evapotranspiração.

Resumindo, as principais vantagens em usar a análise em Componentes Principais residem nessa análise em si mesma, e não como suporte para a modelação por regressão linear. As versões robustas dos diversos métodos ajudaram na identificação de observações atípicas e foram mostrando vantagens registadas ao longo do trabalho, que são melhor expressas em termos da interpretação qualitativa dos resultados, do que em termos de indicadores numéricos.

Os resultados em termos de $RMSE$ sugerem que a situação melhor modelada por um modelo de regressão é para as observações *sem déficit hídrico*, usando a regressão robusta em componentes principais. Os resultados encontrados para o coeficiente de determinação, começam por sugerir que o modelo de regressão usual é preferível ao das componentes principais, seja na versão tradicional ou robusta.

Conjugando os dois critérios com a análise em componentes principais prévia, os resultados apontam no sentido de que, para as condições das plantas em estudo, o modelo de regressão linear não é o desejável para modelar o *Site index*! Esta interpretação, apesar das questões levantadas sobre o método de classificação, é apontada por três vias:

- Pelos métodos robustos, que dão mais segurança nas apreciações e que produziram péssimos resultados de acordo com o critério do R^2.
- Pelos valores do RMSE, uma vez que mede afastamentos relativos a dados reais e que é mínimo para o par Grupo 2 e RPCR, quando nesse caso particular o R^2 indica que o modelo de regressão não é aceitável.
- Pela análise em componentes principais, onde se viu que, mesmo sem agrupamentos, as duas primeiras PCs já explicam mais de 75% da variância nas observações dos regressores; e que as quatro primeiras PCs (todas elas consideradas como regressores) explicam 87%, na versão convencional, e mais de 88%, na versão robusta.

3.6 Recursos Computacionais

O tratamento dos dados e o estudo efetuado requerem uma grande componente de trabalho computacional. Nesta secção descrevem-se os principais recursos computacionais utilizados e algumas das suas especificidades.

As aplicações estatísticas (e outras) utilizadas na realização deste trabalho correram em ambiente *Windows* (*XP Profissional, service pack p3*), num computador portátil com processadores *duol core* da *Intel* de 2.53 GHz cada e 2Gb de memória RAM.

Os dados base foram processados no Sistema de Informação Geográfica (SIG) *ArcGIS 10* da ESRI [10] e armazenados numa *Personal geodatabase* - sistema de gestão de bases de dados *Access* da (*Microsoft*). O *ArcGIS 10* foi ainda utilizado para visualização geográfica das soluções que foram sendo produzidas.

O programa estatístico escolhido para o desenvolvimento do trabalho foi o programa R. O R está disponível gratuitamente na Internet sobre uma *General Public License(GPL).* É uma ferramenta de estatística poderosa e flexível capaz de processar modelos complexos e suportados em enormes conjuntos de dados, permitindo ao utilizador seguir exatamente o que está ser calculado, possuindo ainda, excelentes facilidades gráficas. Para além do programa base, usaram-se diversas *packages* disponíveis, que desempenharam um papel fundamental na análise multivariada e, sobretudo, na aplicação de métodos robustos.

Foram produzidos 10 ficheiros *scripts* no R, cada um com uma função específica: um para importação dos dados e das *packages* do R e pré-processamento dos dados (transformação dos dados); um para análise exploratória de dados; e um por cada tipo de análise realizada (PCA, RPCA, CA, RCA, MLR, RMLR, PCR e RPCR).

O pré-processamento dos dados incluiu a estandardização e transformação dos dados. Na estandardização, utilizou-se a função *scale* da *package base*. Quando as metodologias aplicadas eram também robustas, a estandardização dos dados foi efectuada através de estimadores robustos, nomeadamente, da mediana e do *Median Absolute Deviation* (MAD). Na transformação dos dados utilizaram-se as funções *boxcoxfit* da *package geoR* [26] e a função *boxCox* da *package car* [16] respetivamente, para estimar o parâmetro para cada transformação (por variável), e para realizar a transformação com base no parâmetro estimado.

Na aplicação da Análise de Componentes Principais (PCA) convencional utilizou-se a *package FactorMineR* [23] e na PCA robusta a *package rrcov* [40]. As funções que determinaram as componentes principais são respetivamente, a *PCA* e a *PcaCov*.

As principais vantagens da *package FactorMineR* prendem-se com o facto de possibilitar a utilização de variáveis suplementares, que não entram na composição das componentes, havendo a possibilidade de as representar graficamente, juntamente com as restantes variáveis. E ainda, a grande qualidade dos gráficos que produz, para além de estar muito bem documentada. Uma particularidade a ter em conta na utilização da *package FactoMineR*, é o facto da função *PCA* não gerar explicitamente os *loadings*, que descrevem os contributos das variáveis para cada componente principal, pelo que foi necessário determiná-los separadamente.

Foi ainda utilizada a *package bpca* [12] para gerar alguns gráficos bidimensionais e tridimensionais para representar variáveis e indivíduos por componente principal.

Para encontrar as componentes principais pela abordagem robusta, usou-se a função *PcaCov* da *package rrcov*. Esta função estima a matriz de covariância a partir de um estimador robusto proposto por Rousseew e VanDrissen 1999, usando um algoritmo designado por *Fast-MCD (Minimum Covariance Determinante)*. A função *getFlag* sinaliza observações atípicas.

Na aplicação da Análise de Agrupamentos convencional, utilizou-se a função *kmeans* da *package stats*. Na versão robusta utilizou-se a *package tclust* [24]. A função *kmeans* contida nesta *package* efetua os agrupamentos pelo método não hierárquico *k-means*.

A *package tclust* implementa diferentes algoritmos de agrupamento robusto não hierárquicos, onde a exclusão de observações (*trimming*) da análise tem um papel fundamental. O processo de exclusão permite a remoção de uma fração α dos dados mais atípicos (com $0 < \alpha < 1$), e torna-se mais robusto pelo facto de evitar a influência destas observações.

Havendo necessidade de classificar o conjunto de *dados de validação* de acordo com a mesma regra definida na análise de agrupamentos, recorreu-se à Análise Discriminante Quadrática. No métodos convencionais utilizou-se a função *qda* da *package MASS* e nos métodos robustos a função *QdaCov* da *package rrcov*. Em ambos os casos, as funções correram sobre os dados de calibração agrupados via análise de agrupamentos (CA), tendo sido especificado o campo que classifica cada individua em cada grupo.

Na análise de regressão linear múltipla (MLR) convencional utilizaram-se as funções *lm* da *package stats* e a função *stepAIC* da *package MASS* [45]. A *lm* realiza a regressão linear, enquanto que a função *stepAIC* efetua a regressão passo-a-passo bidirecional com base no Critério de Informação de *Akaike*.

Na análise de regressão linear robusta, utilizaram-se as funções *lmrob* da *package robustbase*[35] e a função *step.lmRob* da *package robust*, a qual seleciona os regressores passo-a-passo, por eliminação sucessiva (*backward*) [46].

Na análise da regressão em componentes principais, os valores a usar como observações dos regressores têm que ser previamente calculados. Como cada componente principal fica definida pela combinação linear das variáveis, sendo o valor dos coeficientes indicado pelos *loadings*, para o *i-ésimo* individuo, é necessário obter o valor de cada componente principal formando a soma dos produtos dos *loadings* multiplicados pelos valores observados dos respetivos regressores (originais). Depois de obtida essa matriz, ela vai desempenhar, na regressão em componentes principais, o papel da matriz de observações dos regressores. Essa matriz vai coincidir com a matriz dos *scores* obtida com a função que faz a análise em componentes principais.

Na Tabela 3.12 apresenta um resumo das principais *packages* usadas e dos nomes dos comandos (para além das *packages* e dos comandos mais generalistas).

Package	Função	Descrição
MASS	*stepAIC()*	Seleciona um modelo com base no Critério de Informação de Akaike (AIC) num algoritmo de Stepwise regression
	qda()	Análise discriminante quadrática (*Quadratic Discriminant Analysis*)
nortest	*lillie.test()*	*Lilliefors (Kolmogorov-Smirnov) test for normality*
car	*boxCox()*	Transformação de Box-Cox para modelos lineares
geoR	*boxcoxfit()*	Estima o parâmetro para a transformação de Box-Cox
FactoMineR	*PCA()*	Realiza a Análise de Componentes Principais (PCA) convencional
	plot.PCA()	Desenha os gráficos da Análise de Componentes principais (PCA)
cluster	*clusplot()*	Gera gráfico bivariado com base em objeto que deriva da aplicação de função de agrupamento
rrcov	*biplot()*	Representa simultânea de observações e variáveis de uma matriz multivariada por componentes derivadas da PCA.
	CovControlMcd()	Cria um objeto de controlo contendo os parâmetros para a função CovMcd()
	CovMcd()	Estimador robusto da posição e dispersão utilizando o estimador Fast-MCD (Minimum Covariance Determinant)
	PcaCov()	Realiza a PCA com base numa matriz de covariância robusta.
	QdaCov()	Realiza a análise discriminante quadrática com base numa matriz de covariância robusta.
	screeplot()	Representação gráfica da variância por Componente gerada via PCA.
tclust	*ctlcurves()*	*Classification Trimmed Likelihood* Curves. Aplica a (*trimmed cluster*) várias vezes enquanto os parâmetros alpha e k se alteram. O objecto resultante dá uma ideia da percentagem ótima de observações atípicas a remover e número de grupos a considerar para um dado conjunto de dados.
	DiscrFact()	*Descriminant Factor Analysis* para objetos tclust.
	tclust()	Abordagem trimming à Análise de Agrupamentos (CA) robusta.
robust	*lmRob()*	Realiza a regressão linear robusta.
	step.lmRob()	Realiza a regressão linear robusta passo-a-passo por eliminação (*stepwise regression backward*).

Tabela 3.12: *Packages* do R e respetivas funções que foram utilizadas na aplicação de metodologias específicas.

Capítulo 4

Conclusões

O trabalho consistiu no estudo da dependência do *Site index* - uma variável biométrica indicadora da produtividade em povoamentos da *Eucalyptus globulus* -, relativamente a um conjunto de variáveis ambientais. O estudo foi delineado em quatro fases de desenvolvimento:

1. De entre um conjunto de variáveis ambientais disponíveis, selecionar as que são mais relevantes para caraterizar as diferentes condições dos povoamentos, tendo como objetivo a futura modelação do *Site index*.

2. Investigar a existência de grupos de indivíduos que justificassem diferentes modelos, consoante as características ambientais.

3. Modelar a dependência do *Site index* em função das variáveis identificadas como relevantes, eventualmente por diferentes modelos de predição perante um cenário de diferentes grupos;

4. Investigar o desempenho dos modelos estatísticos usuais na predição do *Site index* e dos estimadores utilizados, nomeadamente, considerando métodos robustos de estimação, como alternativa e complemento aos métodos tradicionais.

Na primeira fase de desenvolvimento do trabalho, através da aplicação da Análise de Componentes Principais, identificaram-se duas componentes principais (PC) que explicam a maioria da variância das observações. Relacionando-as com o *Site index*, identificaram-se as PC tendo em conta dois fatores limitantes do crescimento das árvores, nomeadamente: o *deficit hídrico* e o *volume de solo*.

Na 1ª PC, as variáveis determinantes foram as climáticas, destacando-se entre elas a precipitação média nos meses mais quentes ($prec_{678}$). A 1ª PC determinou claramente a divisão das observações em dois grupos, confirmados através da aplicação da Análise de Agrupamentos. Num grupo, a disponibilidade de água é o fator limitante ao crescimento das árvores (com *deficit hídrico*); no outro grupo a disponibilidade de água não é fator limitante ao crescimento das árvores (sem *deficit hídrico*).

Na 2ª PC, as variáveis determinantes foram as relacionadas com o solo. Nesta PC, não foi evidente o agrupamento de indivíduos, mas apenas um gradiente de distribuição, essencialmente determinado pela profundidade do solo ($Prof$) e pela quantidade de água no solo disponível para as plantas ($awsc$).

Comparando a aplicação da metodologia estatística convencional com a robusta, os métodos robustos discriminaram melhor a importância de determinada variável dentro de cada componente principal; por exemplo, enquanto no método convencional o peso atribuído a todas as variáveis climáticas foi sensivelmente o mesmo, os métodos robustos pesaram de modo diferente as variáveis, sendo deste modo possível identificar as mais relevantes.

A aplicação de métodos robustos possibilitou ainda a identificação de observações atípicas (*outliers*). No sul do país, estas observações, estão associadas à melhoria das condições climáticas com o aumento da altitude (Serras de Monchique e Açor) e com condições extremas de *deficit hídrico* e reduzido *volume de solo*.

Na região centro as observações atípicas estão associadas a duas condições distintas, na região Oeste, a reduzida precipitação média anual é compensada pelo facto de estar bem distribuída ao longo do ano. De facto, nesta região, o número médio de dias com precipitação superior a 1 mm é particularmente elevado. A região mais interior, junto à fronteira, carateriza-se por apresentar condições extremas de *deficit hídrico* e reduzido *volume de solo*.

Na região Centro-Norte as observações atípicas estão associadas a condições de maior altitude, em que a temperatura mínima é particularmente baixa ou, a situações de muito elevada precipitação, superior a 2000 mm, bem distribuída ao longo do ano e com elevado *volume de solo*.

As observações identificadas como atípicas, traduzem algumas condições edafo-climáticas que se desviam de outras observações mais frequentes, mas, que representam situações reais que interessa incluir na modelação. No entanto, há que considerar que a sua exclusão poderá beneficiar a modelação do *Site index* nas restantes condições ambientais.

Concluindo, esta primeira fase do trabalho foi muito bem conseguida, tendo sido selecionado um conjunto de variáveis que, combinadas em apenas duas componentes principais, explicam cerca de 75% da variância, com vantagem para os métodos robustos. Considerando uma 3ª componente principal, a variância explicada ronda os 80%. As variáveis ambientais disponíveis e consideradas na caraterização ambiental das observações explicam a grande maioria da sua variância, em particular, as variáveis climáticas (cerca de 50%), destacando-se de entre elas a precipitação média anual nos meses mais quentes. As variáveis relativas ao solo explicam ainda uma parte importante da variância dos dados (cerca de 20%), e de entre estas destacam-se a profundidade do solo e a quantidade de água no solo disponível para as plantas. As variáveis altitude e declive explicam ainda uma proporção da variância total relativamente importante.

Na segunda fase do trabalho efetuou-se uma Análise de Agrupamentos aplicada ao conjunto de dados de calibração e com as variáveis selecionadas pela Análise de Componentes Principais. O número máximo de grupos a distinguir com interesse foi de 4; no entanto, as soluções de agrupamento com 3 ou 4 grupos apresentavam elevada sobreposição entre 2 ou mais grupos. Apenas a solução com dois grupos permitiu a maximização da homogeneidade dentro dos grupos e a maximização da heterogeneidade entre grupos. Decidiu-se optar pelo agrupamento em 2 grupos, que foram determinados pelas variáveis climáticas.

Em termos comparativos entre métodos convencionais e métodos robustos, as soluções de agrupamento foram idênticas. No entanto, as diferenças metodológicas são muito diferentes: na aplicação convencional, apenas é necessário definir à partida o número de grupos que se pretendem formar, estando disponíveis algumas ferramentas que ajudam na determinação desse número de grupos; na aplicação do método robusto, os processos computacionais necessários são mais exigentes e levantam muito mais dificuldades. Através da abordagem robusta, foram detetadas e identificadas observações atípicas (cerca de 5%) que coincidem, em termos geográficos, com as detetadas na Análise de Componentes Principais.

Esta fase também foi muito bem conseguida, com vantagens que resultam do agrupamento das observações, sobretudo, na região sem *deficit hídrico*.

A terceira fase do trabalho foi dedicada à predição da variável *Site index* em função de variáveis ambientais. Para isso, aplicou-se o modelo de Regressão Linear Múltipla.

Foram testadas diversas equações de regressão, para além dos modelos apresentados neste texto, incluindo modelos apenas com variáveis quantitativas (transformadas ou não) e modelos com variáveis quantitativas e categóricas. A escolha final recaiu num modelo de regressão linear contendo apenas variáveis quantitativas, algumas delas transformadas. As transformações consideradas e utilizadas na modelação foram a transformação logarítmica da quantidade de água no solo disponível para as plantas e transformações exponenciais da altitude e da evapotranspiração. Também foram investigadas transformações polinomiais de 2º e 3º grau, que foram abandonadas, por não se terem traduzido em benefícios na predição do *Site index*.

O melhor modelo obtido estima o *Site index* com um coeficiente de determinação (R^2) igual a 0.32. Este valor é perfeitamente aceitável, quando comparado com os resultados usuais nesta área específica, mas é pouco satisfatório, em termos de modelação. As variáveis selecionadas via *stepwise regression* foram: a altitude, o declive, a temperatura média máxima, a precipitação média dos meses mais quentes, a evapotranspiração média anual, a pedregosidade do solo, a quantidade de água no solo disponível para as plantas e a latitude do local.

A aplicação de métodos robustos permitiu melhorar apenas ligeiramente o erro padrão dos resíduos e o coeficiente de determinação, mas teve a vantagem de simplificar o modelo, ao selecionar um menor número de variáveis, nomeadamente: a altitude, o declive, a temperatura média máxima, a evapotranspiração, a pedregosidade e a quantidade de água no solo disponível para a planta.

Tendo em conta a grande diversidade de modelos testados sem bons resultados, e o correlacionamento indesejável entre diversas variáveis ambientais enquanto potenciais regressores, optou-se por considerar um modelo de regressão linear em componentes principais, uma vez que as componentes principais são ortogonais entre si e que desse modo, seria de esperar melhores resultados.

Os resultados ficaram muito aquém do esperado. Ignorando o agrupamento sugerido nas fases anteriores do trabalho, não se verificou nenhuma melhoria com a nova metodologia, apesar das componentes principais, por uma lado, serem não correlacionadas entre si e, por outro, integrarem indiretamente no modelo as variáveis originais mais significativas para descrever a variabilidade das características ambientais (obtidas pela Análise de Componentes Principais). Os resultados foram ainda piores, em termos do coeficiente de determinação, ao modelar separadamente os dois grupos de observações. Os procedi-

mentos robustos também não deram resposta aos problemas levantados e concluiu-se que os modelos de regressão linear eram ineficazes para a modelação e predição do *Site index* em função de variáveis ambientais.
Resumindo, nesta terceira fase do trabalho, não foram obtidos bons resultados; no entanto, para o modelo de regressão linear usual, esses resultados enquadram-se dentro de níveis comuns nesta área específica (com R^2 entre 25% e 47%). Considerando a modelação separadamente por grupos, obteve-se a mesma conclusão para um dos grupos. Ou seja, a utilização de modelos lineares, revelou-se pouco eficaz na predição do *Site index* a partir de variáveis ambientais.

Finalmente, numa última fase estudou-se o desempenho dos modelos e dos estimadores. Para este fim, tinha sido inicialmente reservado um subconjunto de dados, selecionado aleatoriamente do conjunto total dos dados, e que nunca interveio nas análises estatísticas anteriores, de modo a poder validar e comparar os resultados encontrados pelos diferentes métodos. Este procedimento não é comum na bibliografia específica sobre o assunto, mas havendo um número de observações suficiente, é claramente recomendado, uma vez que, para cada modelo e cada método de estimação fixos, permite comparar os valores preditos pelo modelo, com os valores observados na realidade.
Existindo condições de validação como as que foram criadas, o critério de comparação de modelos mais adequado é o da raiz quadrada do erro quadrático médio ($RMSE$), o qual dá informação sobre a diferença média entre valores preditos e valores observados. Neste âmbito, este critério é mais interessante do que o coeficiente de determinação.
Assim, as observações do subconjunto de validação foram tratadas como um todo, para avaliar o modelo global e também foram classificadas num dos grupos (com ou sem *déficit hídrico*), com base em análise quadrática discriminante (QDA), convencional e robusta, de modo a aplicar os modelos de regressão estimados para cada grupo em particular.
Usando o $RMSE$ como critério, os resultados da modelação do *Site index* foram generalizadamente fracos e nenhum dos modelos/métodos se mostrou satisfatório (com uma exceção).
O desconhecimento (ou a provável inexistência) de artigos científicos que avaliem a modelação do *Site index* pelo $RMSE$, impede que se estabeleçam comparações com outros trabalhos relativamente à utilização simultânea dos dois critérios, mas é de notar que se o modelo de regressão linear for adequado à modelação, o coeficiente de determinação é útil na avaliação do modelo; mas que, se a relação de dependência for melhor modelado

por outros modelos estatísticos, deixa de fazer sentido usar o coeficiente de determinação para avaliar o ajustamento do modelo.

Uma reflexão sobre a conjugação dos resultados obtidos nas primeiras fases do trabalho e as conclusões encontradas na modelação por regressão, complementada com o conhecimento da área florestal sobre a dependência do *Site index* em relação às variáveis ambientais disponíveis, aponta para a ineficácia do modelo de regressão linear na modelação do *Site index*, apesar de ser o modelo estatístico mais divulgado nas publicações sobre este tema.

Concluindo, a última fase do trabalho sugere que os modelos de regressão linear não são eficazes na predição do *Site index* em função de variáveis ambientais, e que será vantajoso a investigação futura de outros modelos.

Referências Bibliográficas

[1] HAGLOF SWEDEN AB. Vertex iv and transponder t3 manual. http://www.haglofcg.com/, 2007. Consultado em Agosto de 2010.

[2] Wim Aertsen, Vincent Kint, Bart Muys, and Jos Van Orshoven. Effects of scale and scaling in predictive modelling of forest site productivity. *Environmental Modelling Software*, 31:19 – 27, 2012.

[3] Wim Aertsen, Vincent Kint, Jos van Orshoven, Kursad Ozkan, and Bart Muys. Comparison and ranking of different modelling techniques for prediction of site index in mediterranean mountain forests. *Ecological Modelling*, 221(8):1119 – 1130, 2010.

[4] M. Albert and M. Schmidt. Climate-sensitive modelling of site-productivity relationships for norway spruce [picea abies (l.) karst.] and common beech (fagus sylvatica l.). *Forest Ecology and Management*, 259(4):739–749, 2010.

[5] T. W. Beers, P. E. Dress, and L. C. Wensell. Aspect transformation in site productivity research. *Journal of Forestry*, 64:691–692, 1966.

[6] J. T. Brawner, D. J. Lee, M. Hunt, and Auro Almeida. Environmental drivers of eucalypt productivity. In *Proceedings of IUFRO 2011, Brazil - Improvement and Culture of Eucalypts*, pages 139–142. IUFRO, 2011.

[7] J.B. da COSTA, A. L. Azevedo, and R.P. Ricardo. *Caracterização e Constituição do Solo*. Fundação Calouste Gulbenkian, 2004.

[8] David B. Dahl. xtable: Export tables to latex or html. http://CRAN.R-project.org/package=xtable, Consultado em Agosto 2012. R package version 1.7-0.

[9] APA Agencia Portuguesa do Ambiente. Atlas do ambiente. http://sniamb.apambiente.pt/webatlas/. Consultado em Julho de 2011.

[10] ESRI. Arcgis for desktop basic 10. http://www.esriportugal.pt/solucoes. Consultado em Agosto 2012.

[11] FAO. World reference base for soil resources 2006 - a framework for international classification, correlation and communication. ftp://ftp.fao.org/agl/agll/docs/wsrr103e.pdf, 2006. Consultado em Agosto de 2012.

[12] Jose Claudio Faria and Clarice Garcia Borges Demetrio. *bpca: Biplot of Multivariate Data Based on Principal Components Analysis.* UESC and ESALQ, Ilheus, Bahia, Brasil and Piracicaba, Sao Paulo, Brasil, 2012.

[13] Peter Filzmoser and Valentin Todorov. Review of robust multivariate statistical methods in high dimension. *Analytica Chimica Acta*, In Press, Corrected Proof:–, 2011.

[14] L. Fontes, J. D. Bontemps, H. Bugmann, M. Van Oijen, C. Gracia, K. Kramer, M. Lindner, T Rotzer, and J. P. Skovsgaard. Models for supporting forest management in a changing environment. *Forest Systems*, 19(SI):8–20, Setembro 2010.

[15] L. Fontes, M. Tomé, F. Thompson, A. Yeomans, J. Sales Luís, and P.Savill. Modelling the douglas-fir (pseudotsuga menziesii (mirb.) franco) site index from site factores in portugal. *Forestry*, 76(5):491 – 507, 2003.

[16] John Fox and Sanford Weisberg. *An R Companion to Applied Regression.* Sage, Thousand Oaks CA, second edition, 2011.

[17] L. Garciaa-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, 21:585–599, 2011. 10.1007/s11222-010-9194-z.

[18] John C. Grant, J. Doland Nichols, R. Geoff B. Smith, Paul Brennan, and Jerome K. Vanclay. Site index prediction of *Eucalyptus dunni* maiden plantations with soil and site parameters in sub-tropical eastern australia. *Australian Forestry*, 73(4):234–245, February 2010.

[19] Juergen Gross and bug fixes by Uwe Ligges. *nortest: Tests for Normality*, 2012. R package version 1.0-2.

[20] R. J. Harper, J.G Edwards., J. F. McGrath, T. J. Reilly, and S. L. Ward. Performance of eucalyptus globulus plantations in south-western australia in relation to soils and climate. In *Balancing Productivity and Drought Risk in Blue Gum Plantations: a Plantation Management Workshop*, pages 1–6, Karri Valley Resort, Pemberton, W.A, 9 - 10 November 1999. Bunnings Tree Farms, Department of Conservation and Land Management, CSIRO Forestry and Forest Products, Timbercorp Eucalypts Limited, Selected Experiments and Operational Plantations Bunnings Tree Farms.

[21] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[22] M. Hubert and M. Debruyne. Minimum covariance determinant. *Advanced Review*, 2:36–43, Janeiro/Fevereiro 2010.

[23] Francois Husson, Julie Josse, Sebastien Le, and Jeremy Mazet. *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R*, 2012. R package version 1.18.

[24] Agustin Mayo Iscar, Luis Angel Garcia Escudero, and Heinrich Fritz. tclust: Robust trimmed clustering. http://CRAN.R-project.org/package=tclust, 2011. R package version 1.1.

[25] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, sixth edition, 2007.

[26] Paulo J. Ribeiro Jr and Peter J. Diggle. geor: a package for geostatistical analysis. *R-NEWS*, 1(2):15–18, June, 2001.

[27] LaTex. Latex - a document preparation system. http://www.latex-project.org. Consultado em Agosto 2011.

[28] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2012. R package version 1.14.2 — For new features, see the 'Changelog' file (in the package source).

[29] Ricardo A. Maronna, R.Douglas Martin, and Víctor J. Yohai. *Robust Statistics*. John Wiley & Sons, Lda, 2006.

[30] MikTex. Miktex - typesetting beautiful documents. http://miktex.org/about. Consultado em Agosto de 2011.

[31] S. Frosch Moller, J. von Frese, and R. Bro. Robust methods for multivariate data analysis. *Journal of Chemometrics*, 19(10):549–563, 2005.

[32] Ana M. Pires and João A. Branco. *Introdução aos Métodos Estatísticos Robustos*. Congresso Anual da Sociedade Portuguesa de Estatística, 2007.

[33] N. M. Razali and Y. B. Wah. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling test. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[34] Brian Ripley and from 1999 to Oct 2002 Michael Lapsley. *RODBC: ODBC Database Access*, 2012. R package version 1.3-5.

[35] Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, and Martin Maechler. *robustbase: Basic Robust Statistics*, 2012. R package version 0.9-4.

[36] I. Seynave, J.-C. Gégout, J.-C. Hervé, J.-F. Dhôte, J. Drapier, E. Bruno, and Gérard Dumé. Picea abies site index prediction by environmental factors and understorey vegetation: a two-scale approach based on survey databases. *Canadian Journal of Forest Research*, 35:1669–1678, 2005.

[37] J. P. Skovsgaard and J. K. Vanclay. Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands. *Forestry*, 81(1):13–31, 2008.

[38] R Development Core Team. R: A language and environment for statistical computing. http://www.R-project.org/, 2012. Consultado em Agosto de 2012.

[39] TeXnicCenter. Texniccenter - the center of your latex universe. http://www.texniccenter.org/about/about-texniccenter. Consultado em Agosto de 2011.

[40] Valentin Todorov and Peter Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009.

[41] M. Tomé, T. Oliveira, and P. Soares. O modelo globulus 3.0. Publicações GIMREF RC2/2006, Universidade Técnica de Lisboa. Instituto Superior de Agronomia. Centro de Estudos Florestais., Lisboa, 2006.

[42] Luis Torgo. *A Linguagem R. Programação para a análise de dados*. Escolar Editora, 2009.

[43] A. L. Tyler, D. C. Macmillan, and J. Dutch. Models to predict the general yield class of douglas fir, japanese larch and scots pine on better quality land in scotland. *Forestry*, 69:13–24, 1996.

[44] Kurt Varmuza and Peter Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Taylor & Francis Goup, 2008.

[45] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

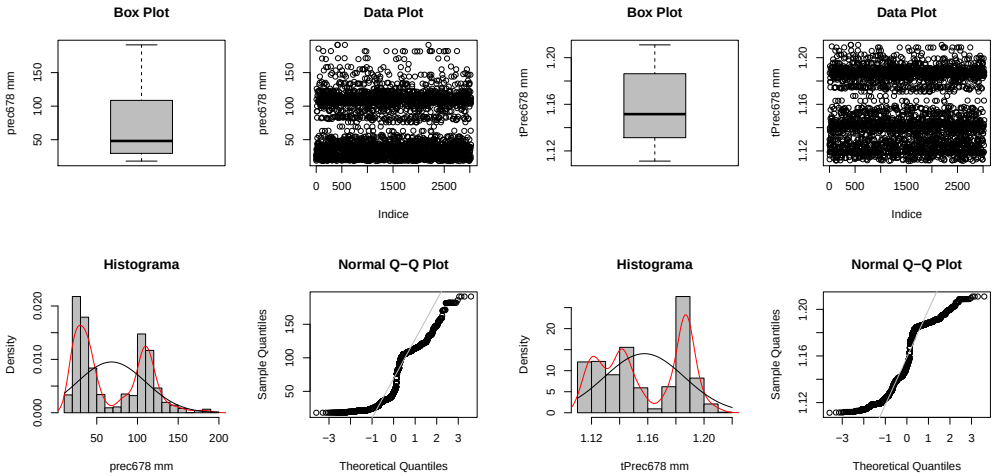
[46] Jiahui Wang, Ruben Zamar, Alfio Marazzi, Victor Yohai, Matias Salibian-Barrera, Ricardo Maronna, Eric Zivot, David Rocke, Doug Martin, Martin Maechler, and Kjell Konis. *robust: Insightful Robust Library*, 2012. R package version 0.3-19.

[47] P. W. West. *Tree and Forest Measurement*. Springer, 2ª edition, 2009.

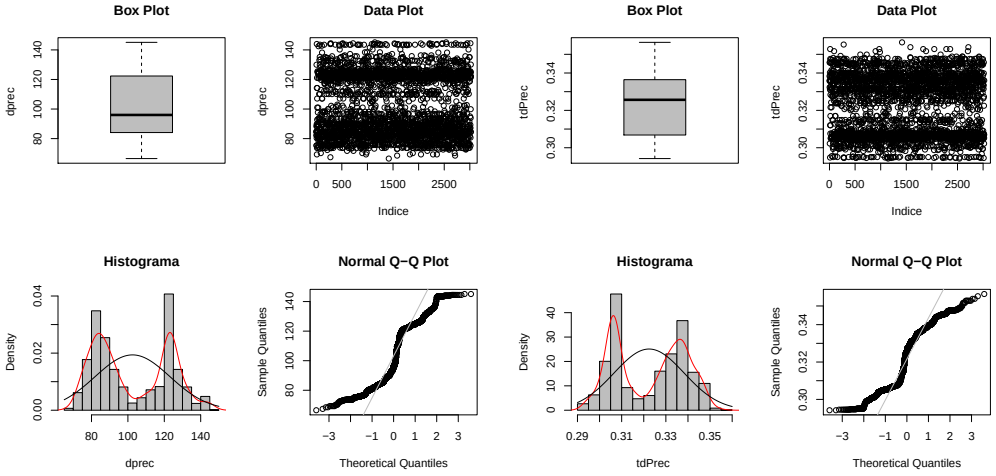
[48] Toshie Yamashita, Keizo Yamashita, and Ryotaro Kamimura. A Stepwise AIC method for variable selection in linear regression. *Communications in Statistics - Theory and Methods*, 36(13):2395–2403, 2007.

Apêndice A

Figuras adicionais

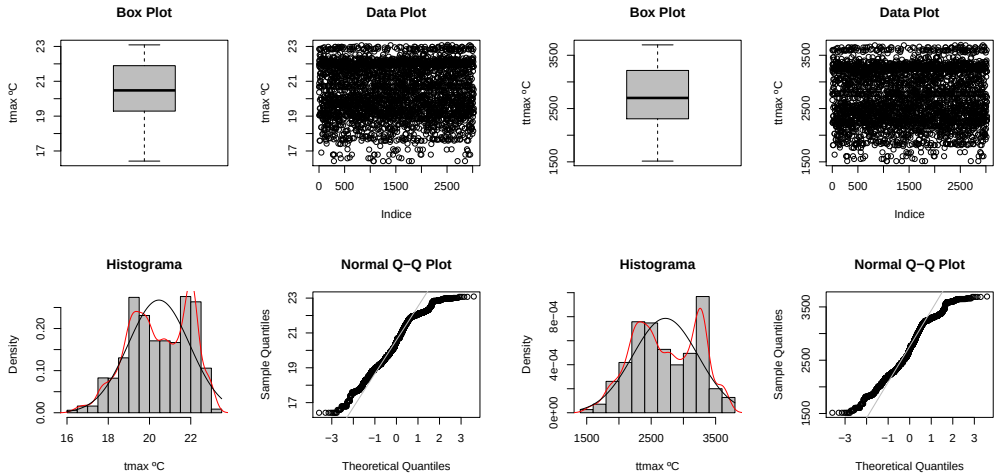


(a) Precipitação média anual dos meses Junho, Julho e Agosto (mm)

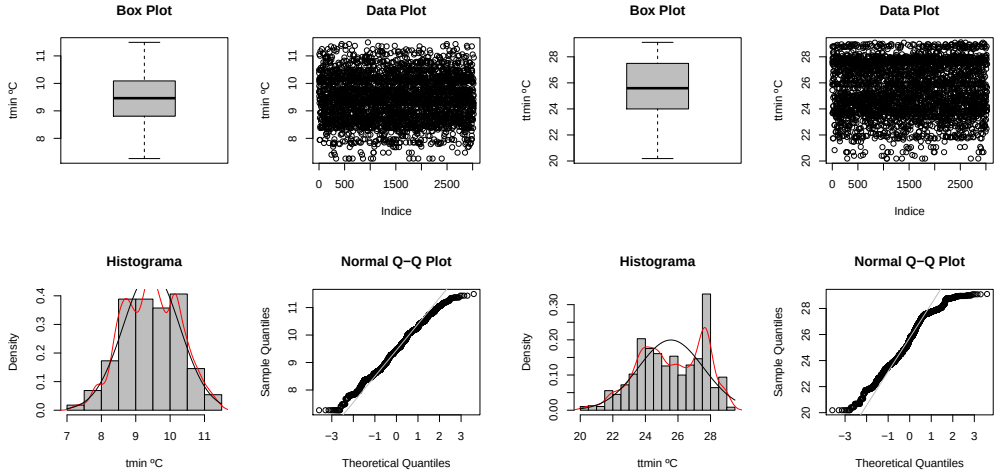


(b) Número médio de dias com Precipitação superior a 1 mm

Figura A.1: Representação gráfica univariada de variáveis.

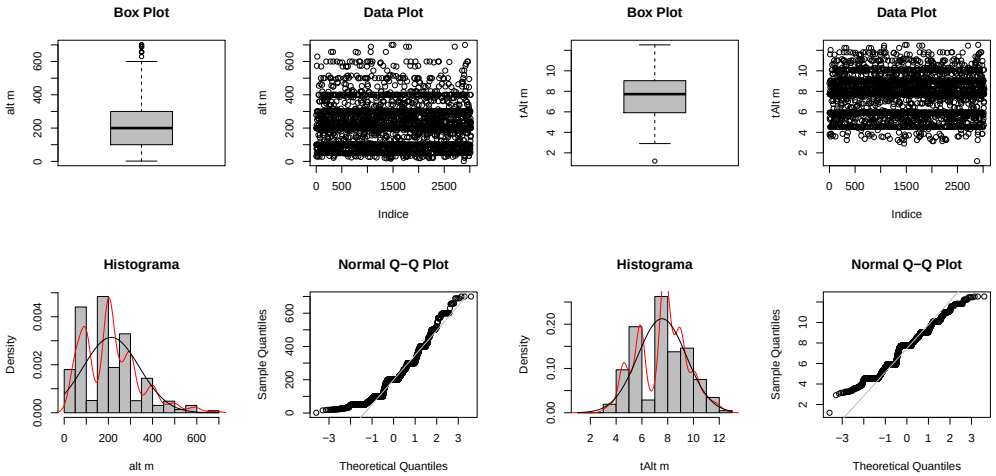


(a) Temperatura média máxima anual (°C)

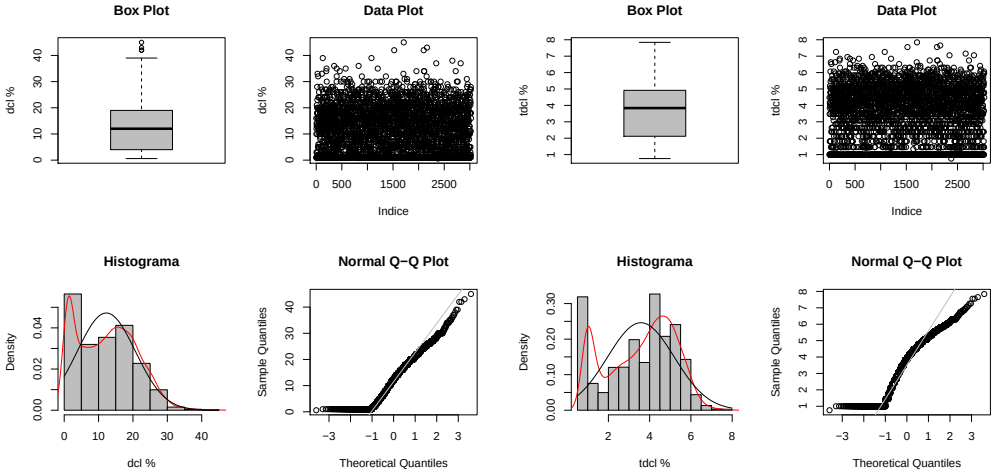


(b) Temperatura média mínima anual (°C)

Figura A.2: Representação gráfica univariada de variáveis.

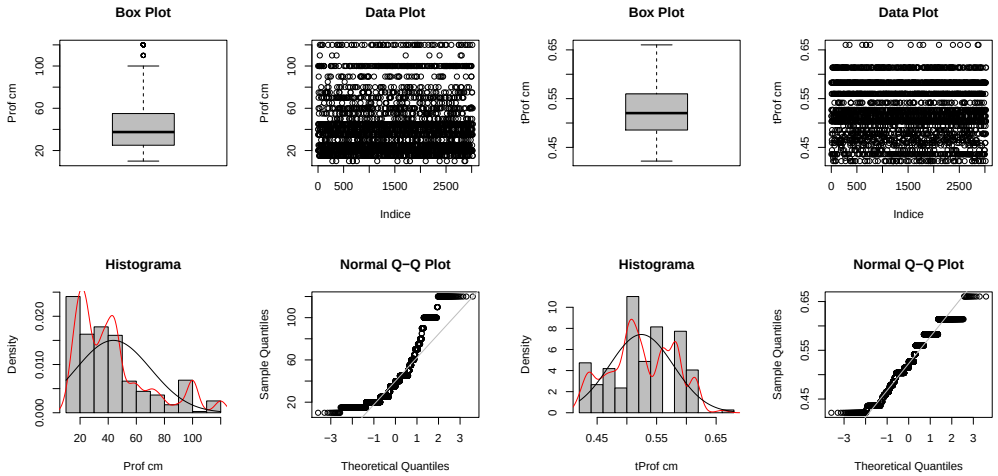


(a) Altitude (m)

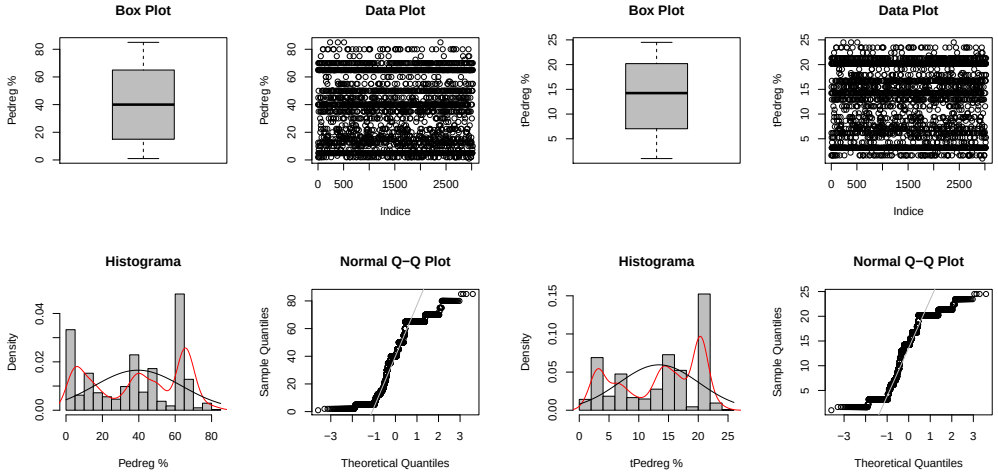


(b) Declive (%)

Figura A.3: Representação gráfica univariada de variáveis.



(a) Profundidade do solo (cm)



(b) Pedregosidade do solo (%)

Figura A.4: Representação gráfica univariada de variáveis.

Apêndice B

Tabelas adicionais

	Mínimo	1.º Quartil	Mediana	Media	3.ºQuartil	Máximo	Desvio Padrão
S	10.1	16.7	19.3	19.6	22.0	32.8	3.8
alt	1.6	100.0	200.0	212.9	300.0	700.0	126.9
dcl	0.0	4.0	12.0	12.1	19.0	45.0	8.7
exp	-1.0	94.8	168.7	170.6	240.8	357.9	93.8
tmin	7.3	8.8	9.5	9.5	10.1	11.5	0.8
tmax	16.4	19.3	20.5	20.5	21.9	23.1	1.5
tmed	11.9	14.2	15.0	15.0	15.8	16.8	1.0
prec	548.2	713.9	868.1	1126.5	1524.2	2675.1	477.8
dprec	66.5	84.1	96.0	102.5	122.3	145.1	20.6
prec678	18.1	29.6	48.1	68.1	108.6	191.3	41.9
evap	27.7	33.2	42.4	40.9	48.8	59.5	8.3
Prof	10.0	25.0	37.5	43.7	55.0	120.0	26.6
Pedreg	0.0	15.0	40.0	39.5	65.0	85.0	24.1
awsc	7.6	27.0	41.7	49.9	62.0	202.7	32.2
x	105907.4	163038.5	177303.5	184131.3	203477.2	295704.0	29523.4
y	31179.3	170302.5	306739.8	297583.7	460735.2	559073.6	164140.4

Tabela B.1: Estatísticas Sumárias. Dados utilizados na modelação (com 3022 observações).

	Mínimo	1.º Quartil	Mediana	Media	3.ºQuartil	Máximo	Desvio Padrão
S	10.1	16.6	19.0	19.3	21.4	31.6	3.8
alt	19.5	100.0	200.0	196.3	266.3	629.7	120.1
dcl	0.0	5.0	13.0	12.8	18.0	50.0	9.2
exp	-1.0	118.4	186.3	183.9	256.7	359.2	94.5
tmin	7.3	8.9	9.5	9.6	10.2	11.4	0.8
tmax	16.6	19.5	20.7	20.6	22.0	23.0	1.4
tmed	12.3	14.4	15.2	15.1	15.9	16.8	1.0
prec	548.6	704.5	812.5	1080.7	1494.3	2675.1	478.2
dprec	72.5	83.5	93.0	100.7	121.6	144.5	20.2
prec678	18.2	28.5	41.6	63.8	107.7	191.3	42.0
evap	27.7	33.6	42.7	41.7	49.1	59.2	8.2
Prof	10.0	25.0	40.0	43.0	50.0	120.0	25.5
Pedreg	2.0	12.5	40.0	39.1	65.0	75.0	24.1
awsc	9.0	29.4	42.5	52.4	66.8	194.2	33.5
x	117372.6	162113.0	176635.5	182615.4	200791.5	294439.9	29607.9
y	37826.9	99195.2	270431.2	280293.7	449340.9	558558.5	164565.9

Tabela B.2: Estatísticas Sumárias. Dados utilizados na validação (com 336 observações).

	Shapiro-Wilk		Lilliefors (K-S)	
Variável	Statistica W	p-value	Statistica D	p-value
S	0.9849929	2.28E-17	0.03647522	6.27E-10
alt	0.9320793	4.84E-35	0.12352806	6.65E-127
dcl	0.9548341	1.27E-29	0.08088279	7.23E-53
exp	0.9751947	1.59E-22	0.04567065	5.85E-16
tmin	0.992779	3.77E-11	0.05465385	2.71E-23
tmax	0.9659236	3.71E-26	0.09471697	2.31E-73
tmed	0.9663332	5.16E-26	0.09590173	2.83E-75
prec	0.8789105	1.52E-43	0.21369834	0.00E+00
dprec	0.9004804	1.50E-40	0.15659712	1.94E-206
prec678	0.8710386	1.58E-44	0.19790416	0.00E+00
evap	0.9206919	3.16E-37	0.18989907	4.59E-306
Prof	0.859955	7.82E-46	0.20140403	0.00E+00
Pedreg	0.8949899	2.33E-41	0.179746	1.30E-273
awsc	0.8737722	3.43E-44	0.14407443	4.78E-174
lnAwsc	0.9897002	5.60E-14	0.06857656	1.77E-37
x	0.8915409	7.53E-42	0.13201833	1.76E-145
y	0.9027301	3.29E-40	0.15033521	6.78E-190

Tabela B.3: Teste de aproximação à distribuição Normal de Shapiro-Wilk e de Lilliefors (K-S)

(a) Grupo 1 - com *deficit hídrico* (com 1672 observações

	Media	Mediana	Desvio Padrão	MAD
S	18.26	18.11	3.04	3.17
alt	159.63	100.00	105.08	87.08
dcl	9.14	7.00	7.98	8.90
exp	173.61	173.27	96.24	111.42
tmin	9.96	10.05	0.65	0.59
tmax	21.50	21.77	0.94	0.72
tmed	15.73	15.77	0.53	0.57
prec	736.94	726.10	114.45	74.72
dprec	85.35	84.58	7.40	7.30
prec678	32.97	32.80	10.28	12.01
evap	47.48	48.27	4.49	4.71
Prof	46.66	40.00	30.26	29.65
Pedreg	32.50	30.00	25.67	37.06
awsc	48.00	39.60	29.33	23.68
x	181701.91	166336.74	36827.06	19506.44
y	166503.11	174397.57	91746.44	119254.96

(b) Grupo 2 - sem *deficit hídrico* (com 1350 observações).

	Media	Mediana	Desvio Padrão	MAD
S	21.23	21.13	4.00	4.28
alt	278.89	251.42	120.53	76.24
dcl	16.12	16.00	7.35	7.41
exp	166.81	165.62	90.62	101.94
tmin	8.83	8.86	0.53	0.62
tmax	19.16	19.21	0.92	0.85
tmed	14.00	14.17	0.63	0.57
prec	1609.08	1561.64	271.79	204.95
dprec	123.84	123.38	7.82	4.02
prec678	111.61	109.52	19.57	10.78
evap	32.74	32.56	2.73	2.27
Prof	39.96	35.00	20.61	14.83
Pedreg	48.19	50.00	18.67	22.24
awsc	52.22	42.52	35.38	25.52
x	187140.14	183969.40	15988.45	17341.86
y	459929.34	467605.29	47293.42	28366.95

Tabela B.4: Medidas de tendência central e de dispersão.

(a) Coeficiente de Correlação de *Pearson*

	S	alt	exp	dcl	tmin	tmax	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	0.14	-0.03	0.25	-0.31	-0.28	0.34	0.35	0.35	-0.34	0.15	-0.11	0.36	-0.05	0.34
alt	0.14	1.00	-0.09	0.33	-0.47	**-0.62**	**0.52**	0.38	**0.50**	-0.38	-0.25	0.43	-0.03	**0.53**	0.38
exp	-0.03	-0.09	1.00	0.06	0.04	0.00	-0.00	-0.03	-0.02	0.05	0.10	0.01	-0.01	-0.11	-0.04
dcl	0.25	0.33	0.06	1.00	-0.27	**-0.52**	0.35	0.39	0.31	-0.45	-0.12	**0.52**	-0.04	0.07	0.23
tmin	-0.31	-0.47	0.04	-0.27	1.00	0.49	**-0.64**	**-0.63**	**-0.71**	**0.52**	-0.04	-0.13	-0.09	-0.31	**-0.75**
tmax	-0.28	**-0.62**	0.00	**-0.52**	0.49	1.00	**-0.84**	**-0.78**	**-0.76**	**0.82**	0.05	-0.43	-0.12	0.07	**-0.59**
prec	0.34	**0.52**	-0.00	0.35	**-0.64**	**-0.84**	1.00	**0.92**	**0.97**	**-0.87**	-0.02	0.23	0.18	-0.01	**0.84**
dprec	0.35	0.38	-0.03	0.39	**-0.63**	**-0.78**	**0.92**	1.00	**0.92**	**-0.89**	-0.04	0.31	0.11	0.00	**0.87**
prec678	0.35	**0.50**	-0.02	0.31	**-0.71**	**-0.76**	**0.97**	0.92	1.00	-0.83	-0.05	0.21	0.14	0.14	0.93
evap	-0.34	-0.38	0.05	-0.45	**0.52**	**0.82**	**-0.87**	**-0.89**	**-0.83**	1.00	0.11	-0.41	-0.07	0.11	**-0.73**
Prof	0.15	-0.25	0.10	-0.12	-0.04	0.05	-0.02	-0.04	-0.05	0.11	1.00	-0.42	**0.50**	-0.36	-0.04
Pedreg	-0.11	0.43	0.01	**0.52**	-0.13	-0.43	0.23	0.31	0.21	-0.41	-0.42	1.00	-0.49	0.34	0.17
awsc	0.36	-0.03	-0.01	-0.04	-0.09	-0.12	0.18	0.11	0.14	-0.07	**0.50**	-0.49	1.00	-0.30	0.07
x	-0.05	**0.53**	-0.11	0.07	-0.31	0.07	-0.01	0.00	0.14	0.11	-0.36	0.34	-0.30	1.00	0.28
y	0.34	0.38	-0.04	0.23	**-0.75**	**-0.59**	**0.84**	**0.87**	**0.93**	**-0.73**	-0.04	0.17	0.07	0.28	1.00

(b) *Fast MCD - Minimum Covariance Determinant*

	S	alt	exp	dcl	tmin	tmax	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	0.25	-0.10	0.29	-0.31	-0.31	0.37	0.35	0.37	-0.37	0.05	-0.01	0.26	0.16	0.36
alt	0.25	1.00	-0.05	0.43	**-0.57**	**-0.68**	**0.67**	**0.58**	**0.62**	**-0.63**	-0.40	**0.50**	-0.12	**0.58**	0.48
exp	-0.10	-0.05	1.00	0.04	0.07	0.04	-0.05	-0.05	-0.07	0.06	0.02	0.10	-0.10	-0.06	-0.08
dcl	0.29	0.43	0.04	1.00	-0.30	**-0.58**	0.49	**0.52**	0.42	**-0.60**	-0.21	**0.61**	-0.04	0.18	0.33
tmin	-0.31	**-0.57**	0.07	-0.30	1.00	**0.53**	**-0.71**	**-0.68**	**-0.77**	**0.58**	0.06	-0.20	0.04	**-0.65**	**-0.81**
tmax	-0.31	**-0.68**	0.04	**-0.58**	**0.53**	1.00	**-0.88**	**-0.84**	**-0.79**	**0.89**	0.28	**-0.57**	0.05	-0.14	**-0.67**
prec	0.37	**0.67**	-0.05	0.49	**-0.71**	**-0.88**	1.00	**0.95**	**0.98**	**-0.92**	-0.27	0.46	-0.09	0.39	**0.91**
dprec	0.35	**0.58**	-0.05	**0.52**	**-0.68**	**-0.84**	**0.95**	1.00	**0.95**	**-0.92**	-0.26	**0.51**	-0.10	0.42	**0.91**
prec678	0.37	**0.62**	-0.07	0.42	**-0.77**	**-0.79**	**0.98**	**0.95**	1.00	**-0.88**	-0.24	0.38	-0.08	0.48	**0.97**
evap	-0.37	**-0.63**	0.06	**-0.60**	**0.58**	**0.89**	**-0.92**	**-0.92**	**-0.88**	1.00	0.32	**-0.60**	0.10	-0.28	**-0.80**
Prof	0.05	-0.40	0.02	-0.21	0.06	0.28	-0.27	-0.26	-0.24	0.32	1.00	-0.45	**0.53**	-0.30	-0.14
Pedreg	-0.01	**0.50**	0.10	**0.61**	-0.20	**-0.57**	0.46	**0.51**	0.38	**-0.60**	-0.45	1.00	-0.44	0.24	0.25
awsc	0.26	-0.12	-0.10	-0.04	0.04	0.05	-0.09	-0.10	-0.08	0.10	**0.53**	-0.44	1.00	-0.17	-0.06
x	0.16	0.58	-0.06	0.18	**-0.65**	-0.14	0.39	0.42	0.48	-0.28	-0.30	0.24	-0.17	1.00	0.49
y	0.36	0.48	-0.08	0.33	**-0.81**	**-0.67**	**0.91**	**0.91**	**0.97**	**-0.80**	-0.14	0.25	-0.06	0.49	1.00

Tabela B.5: Matriz de Correlações (com 3022 observações). Variáveis originais

(a) *Método de Pearson*

	S	alt	exp	dcl	tmin	tmax	tmed	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	-0.06	-0.09	0.01	0.00	-0.04	-0.04	0.12	0.14	-0.03	-0.16	0.12	-0.20	0.26	-0.22	-0.04
alt	-0.06	1.00	-0.01	0.36	-0.07	-0.43	-0.42	0.27	0.04	0.27	-0.06	-0.50	0.59	-0.26	0.69	0.04
exp	-0.09	-0.01	1.00	0.06	-0.04	-0.02	-0.04	0.03	0.05	0.02	-0.02	0.02	0.08	-0.10	0.02	0.02
dcl	0.01	0.36	0.06	1.00	0.11	-0.53	-0.40	0.36	0.36	-0.05	-0.49	-0.22	0.66	-0.15	0.05	-0.21
tmin	0.00	-0.07	-0.04	0.11	1.00	-0.16	0.48	-0.04	-0.05	-0.54	-0.14	-0.15	0.12	-0.01	-0.35	-0.62
tmax	-0.04	-0.43	-0.02	-0.53	-0.16	1.00	0.79	-0.63	-0.54	-0.03	0.64	0.14	-0.54	0.03	0.14	0.24
tmed	-0.04	-0.42	-0.04	-0.40	0.48	0.79	1.00	-0.58	-0.51	-0.36	0.49	0.04	-0.40	0.02	-0.09	-0.17
prec	0.12	0.27	0.03	0.36	-0.04	-0.63	-0.58	1.00	0.55	0.41	-0.51	0.09	0.24	0.10	-0.18	0.08
dprec	0.14	0.04	0.05	0.36	-0.05	-0.54	-0.51	0.55	1.00	0.26	-0.43	0.11	0.25	-0.01	-0.17	0.13
prec678	-0.03	0.27	0.02	-0.05	-0.54	-0.03	-0.36	0.41	0.26	1.00	0.21	0.06	0.01	-0.10	0.50	0.90
evap	-0.16	-0.06	-0.02	-0.49	-0.14	0.64	0.49	-0.51	-0.43	0.21	1.00	0.01	-0.47	-0.00	0.39	0.35
Prof	0.12	-0.50	0.02	-0.22	-0.15	0.14	0.04	0.09	0.11	0.06	0.01	1.00	-0.44	0.49	-0.42	0.12
Pedreg	-0.20	0.59	0.08	0.66	0.12	-0.54	-0.40	0.24	0.25	0.01	-0.47	-0.44	1.00	-0.43	0.36	-0.15
awsc	0.26	-0.26	-0.10	-0.15	-0.01	0.03	0.02	0.10	-0.01	-0.10	-0.00	0.49	-0.43	1.00	-0.37	-0.12
x	-0.22	0.69	0.02	0.05	-0.35	0.14	-0.09	-0.18	-0.17	0.50	0.39	-0.42	0.36	-0.37	1.00	0.50
y	-0.04	0.04	0.02	-0.21	-0.62	0.24	-0.17	0.08	0.13	0.90	0.35	0.12	-0.15	-0.12	0.50	1.00

(b) *Fast MCD - Minimum Covariance Determinant*

	S	alt	exp	dcl	tmin	tmax	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	0.11	-0.06	-0.01	-0.11	0.02	0.08	0.05	0.07	-0.06	0.00	-0.11	0.17	0.01	0.04
alt	0.11	1.00	0.01	0.40	0.05	-0.29	-0.08	-0.04	-0.49	-0.27	-0.46	0.54	-0.12	0.53	-0.61
exp	-0.06	0.01	1.00	0.11	0.03	-0.06	0.06	0.10	-0.01	-0.07	0.03	0.11	-0.11	-0.01	-0.01
dcl	-0.01	0.40	0.11	1.00	0.04	-0.63	0.39	0.39	-0.26	-0.69	-0.18	0.77	-0.13	-0.05	-0.39
tmin	-0.11	0.05	0.03	0.04	1.00	-0.18	-0.38	-0.25	-0.74	-0.14	-0.27	0.23	-0.12	-0.33	-0.73
tmax	0.02	-0.29	-0.06	-0.63	-0.18	1.00	-0.59	-0.52	0.28	0.79	0.10	-0.67	0.02	0.51	0.47
prec	0.08	-0.08	0.06	0.39	-0.38	-0.59	1.00	0.49	0.49	-0.55	0.26	0.26	0.11	-0.51	0.27
dprec	0.05	-0.04	0.10	0.39	-0.25	-0.52	0.49	1.00	0.16	-0.50	0.19	0.34	-0.06	-0.19	0.11
prec678	0.07	-0.49	-0.01	-0.26	-0.74	0.28	0.49	0.16	1.00	0.28	0.47	-0.46	0.19	-0.09	0.94
evap	-0.06	-0.27	-0.07	-0.69	-0.14	0.79	-0.55	-0.50	0.28	1.00	0.11	-0.75	0.10	0.40	0.44
Prof	0.00	-0.46	0.03	-0.18	-0.27	0.10	0.26	0.19	0.47	0.11	1.00	-0.31	0.38	-0.25	0.50
Pedreg	-0.11	0.54	0.11	0.77	0.23	-0.67	0.26	0.34	-0.46	-0.75	-0.31	1.00	-0.30	0.00	-0.59
awsc	0.17	-0.12	-0.11	-0.13	-0.12	0.02	0.11	-0.06	0.19	0.10	0.38	-0.30	1.00	-0.16	0.17
x	0.01	0.53	-0.01	-0.05	-0.33	0.51	-0.51	-0.19	-0.09	0.40	-0.25	0.00	-0.16	1.00	-0.01
y	0.04	-0.61	-0.01	-0.39	-0.73	0.47	0.27	0.11	0.94	0.44	0.50	-0.59	0.17	-0.01	1.00

Tabela B.6: Matriz de Correlações. Grupo de dados *com deficit hídrico* (com 1672 observações)

	S	alt	exp	dcl	tmin	tmax	tmed	prec	dprec	prec678	evap	Prof	Pedreg	awsc	x	y
S	1.00	-0.12	-0.11	0.23	0.06	0.08	0.08	-0.05	-0.11	-0.06	-0.03	0.42	-0.35	0.41	-0.05	-0.04
alt	-0.12	1.00	0.01	0.01	-0.81	-0.63	-0.80	0.27	-0.23	0.17	0.29	0.23	-0.02	0.03	0.39	-0.15
exp	-0.11	0.01	1.00	0.04	0.01	-0.03	-0.01	0.13	0.07	0.10	0.06	-0.06	0.17	-0.08	-0.07	-0.01
dcl	0.23	0.01	0.04	1.00	-0.05	0.06	0.02	-0.22	-0.23	-0.24	0.24	0.17	0.15	0.04	0.15	-0.27
tmin	0.06	-0.81	0.01	-0.05	1.00	0.46	0.76	-0.08	0.38	-0.02	-0.47	-0.23	-0.01	0.01	-0.67	0.22
tmax	0.08	-0.63	-0.03	0.06	0.46	1.00	0.93	-0.65	-0.13	-0.49	0.15	-0.19	0.23	-0.20	0.23	-0.06
tmed	0.08	-0.80	-0.01	0.02	0.76	0.93	1.00	-0.51	0.06	-0.37	-0.08	-0.24	0.16	-0.14	-0.11	0.05
prec	-0.05	0.27	0.13	-0.22	-0.08	-0.65	-0.51	1.00	0.61	0.95	-0.45	0.17	-0.30	0.28	-0.50	0.57
dprec	-0.11	-0.23	0.07	-0.23	0.38	-0.13	0.06	0.61	1.00	0.66	-0.44	0.02	-0.09	0.19	-0.53	0.71
prec678	-0.06	0.17	0.10	-0.24	-0.02	-0.49	-0.37	0.95	0.66	1.00	-0.52	0.13	-0.29	0.25	-0.40	0.75
evap	-0.03	0.29	0.06	0.24	-0.47	0.15	-0.08	-0.45	-0.44	-0.52	1.00	0.10	0.30	-0.10	0.57	-0.61
Prof	0.42	0.23	-0.06	0.17	-0.23	-0.19	-0.24	0.17	0.02	0.13	0.10	1.00	-0.42	0.71	-0.01	0.03
Pedreg	-0.35	-0.02	0.17	0.15	-0.01	0.23	0.16	-0.30	-0.09	-0.29	0.30	-0.42	1.00	-0.70	0.27	-0.18
awsc	0.41	0.03	-0.08	0.04	0.01	-0.20	-0.14	0.28	0.19	0.25	-0.10	0.71	-0.70	1.00	-0.24	0.15
x	-0.05	0.39	-0.07	0.15	-0.67	0.23	-0.11	-0.50	-0.53	-0.40	0.57	-0.01	0.27	-0.24	1.00	-0.25
y	-0.04	-0.15	-0.01	-0.27	0.22	-0.06	0.05	0.57	0.71	0.75	-0.61	0.03	-0.18	0.15	-0.25	1.00

Tabela B.7: Matriz de Correlações (*Método de Pearson*). Grupo de dados *sem deficit hídrico* (com 1350 observações).

(a) *Método de Pearson*

	S2	alt2	dcl2	tmin2	tmax2	tmed2	prec2	dprec2	prec6782	evap2	Prof2	Pedreg2	awsc2	x2	y2
S2	1.00	0.12	0.22	-0.24	-0.28	-0.30	-0.36	-0.36	0.34	-0.37	0.12	0.02	0.37	0.05	0.31
alt2	0.12	1.00	0.43	-0.53	-0.65	-0.67	-0.53	-0.45	0.51	-0.45	0.11	0.41	-0.13	-0.57	0.40
dcl2	0.22	0.43	1.00	-0.27	-0.54	-0.50	-0.47	-0.47	0.37	-0.49	0.20	0.43	-0.06	-0.17	0.28
tmin2	-0.24	-0.53	-0.27	1.00	0.56	0.80	0.65	0.60	-0.74	0.53	-0.03	-0.21	-0.02	0.44	-0.75
tmax2	-0.28	-0.65	-0.54	0.56	1.00	0.94	0.88	0.82	-0.76	0.84	-0.20	-0.40	-0.06	0.01	-0.62
tmed2	-0.30	-0.67	-0.50	0.80	0.94	1.00	0.89	0.84	-0.84	0.82	-0.16	-0.37	-0.05	0.17	-0.75
prec2	-0.36	-0.53	-0.47	0.65	0.88	0.89	1.00	0.94	-0.94	0.92	-0.15	-0.33	-0.08	0.08	-0.84
dprec2	-0.36	-0.45	-0.47	0.60	0.82	0.84	0.94	1.00	-0.90	0.90	-0.15	-0.33	-0.06	0.08	-0.84
prec6782	0.34	0.51	0.37	-0.74	-0.76	-0.84	-0.94	-0.90	1.00	-0.83	0.05	0.28	0.02	-0.27	0.97
evap2	-0.37	-0.45	-0.49	0.53	0.84	0.82	0.92	0.90	-0.83	1.00	-0.19	-0.36	-0.05	-0.01	-0.74
Prof2	0.12	0.11	0.20	-0.03	-0.20	-0.16	-0.15	-0.15	0.05	-0.19	1.00	0.24	0.09	0.12	-0.02
Pedreg2	0.02	0.41	0.43	-0.21	-0.40	-0.37	-0.33	-0.33	0.28	-0.36	0.24	1.00	-0.20	-0.29	0.22
awsc2	0.37	-0.13	-0.06	-0.02	-0.06	-0.05	-0.08	-0.06	0.02	-0.05	0.09	-0.20	1.00	0.30	-0.03
x2	0.05	-0.57	-0.17	0.44	0.01	0.17	0.08	0.08	-0.27	-0.01	0.12	-0.29	0.30	1.00	-0.32
y2	0.31	0.40	0.28	-0.75	-0.62	-0.75	-0.84	-0.84	0.97	-0.74	-0.02	0.22	-0.03	-0.32	1.00

(b) *Fast MCD - Minimum Covariance Determinant*

	S2	alt2	exp_pond2	dcl2	tmin2	tmax2	prec2	dprec2	prec6782	evap2	Prof2.1	Pedreg2.1	awsc2	y2
S2	1.00	0.22	-0.01	0.29	-0.25	-0.32	-0.35	-0.34	0.35	-0.34	-0.19	0.03	0.32	0.33
alt2	0.22	1.00	0.11	0.50	-0.63	-0.71	-0.63	-0.58	0.57	-0.63	0.31	0.54	-0.08	0.46
exp_pond2	-0.01	0.11	1.00	0.07	-0.17	-0.07	-0.09	-0.09	0.09	-0.07	0.06	0.15	-0.07	0.10
dcl2	0.29	0.50	0.07	1.00	-0.41	-0.60	-0.55	-0.55	0.45	-0.61	0.11	0.66	-0.09	0.38
tmin2	-0.25	-0.63	-0.17	-0.41	1.00	0.67	0.72	0.70	-0.77	0.58	-0.01	-0.31	-0.01	-0.76
tmax2	-0.32	-0.71	-0.07	-0.60	0.67	1.00	0.94	0.88	-0.85	0.91	-0.14	-0.53	0.01	-0.77
prec2	-0.35	-0.63	-0.09	-0.55	0.72	0.94	1.00	0.96	-0.97	0.94	-0.13	-0.48	0.02	-0.92
dprec2	-0.34	-0.58	-0.09	-0.55	0.70	0.88	0.96	1.00	-0.95	0.93	-0.14	-0.52	0.06	-0.93
prec6782	0.35	0.57	0.09	0.45	-0.77	-0.85	-0.97	-0.95	1.00	-0.89	0.10	0.37	0.00	0.98
evap2	-0.34	-0.63	-0.07	-0.61	0.58	0.91	0.94	0.93	-0.89	1.00	-0.21	-0.58	0.08	-0.83
Prof2.1	-0.19	0.31	0.06	0.11	-0.01	-0.14	-0.13	-0.14	0.10	-0.21	1.00	0.40	-0.60	0.07
Pedreg2.1	0.03	0.54	0.15	0.66	-0.31	-0.53	-0.48	-0.52	0.37	-0.58	0.40	1.00	-0.49	0.30
awsc2	0.32	-0.08	-0.07	-0.09	-0.01	0.01	0.02	0.06	0.00	0.08	-0.60	-0.49	1.00	-0.00
y2	0.33	0.46	0.10	0.38	-0.76	-0.77	-0.92	-0.93	0.98	-0.83	0.07	0.30	-0.00	1.00

Tabela B.8: Matriz de Correlações (com 3022 observações). Dados transformação *Box-cox transformation.*

	Shapiro-Wilk		Lilliefors (K-S)	
Variável	Statistica W	p-value	Statistica D	p-value
tS	0.997974	6.22E-04	0.016799	4.79E-02
tAlt	0.97268	1.34E-23	0.141847	1.42E-168
tdcl	0.935073	2.03E-34	0.104259	1.80E-89
texp	0.894888	2.25E-41	0.132792	3.10E-147
ttmin	0.966129	4.38E-26	0.09468	2.65E-73
ttmax	0.967396	1.24E-25	0.094573	3.93E-73
tPrec	0.900766	1.65E-40	0.170705	3.00E-246
tdPrec	0.908314	2.47E-39	0.169145	1.13E-241
tPrec678	0.89582	3.07E-41	0.177986	3.57E-268
tEvap	0.918934	1.53E-37	0.179771	1.08E-273
tProf	0.963825	7.14E-27	0.118964	1.92E-117
tPedreg	0.888283	2.66E-42	0.169336	3.11E-242
tAwsc	0.990329	1.89E-13	0.063431	7.76E-32
tx	0.957147	5.87E-29	0.073473	2.79E-43
ty	0.901703	2.29E-40	0.149305	3.07E-187

Tabela B.9: Teste de aproximação à distribuição Normal de *Shapiro-Wilk e de Lilliefors* (K-S). Com transformação de *Box-Cox*.

(a) Grupo 1 - com *deficit* hídrico

	Valores Próprios	Percentagem de Variância	Percentagem de variância acumulada
PC 1	3.8	**34.7**	34.7
PC 2	2.2	**19.6**	54.3
PC 3	1.8	**16.5**	70.9
PC 4	0.7	**6.4**	77.3
PC 5	0.7	**5.9**	83.2
PC 6	0.5	**4.7**	**87.9**
PC 7	0.5	4.2	92.1
PC 8	0.4	3.3	95.4
PC 9	0.2	2.1	97.5
PC 10	0.2	1.4	98.9
PC 11	0.1	1.1	100.0

(b) Grupo 2 - Sem *deficit hídrico*

	Valores Próprios	Percentagem de Variância	Percentagem de variância acumulada
PC 1	3.6	**32.3**	32.3
PC 2	2.6	**23.8**	56.1
PC 3	2.0	**17.8**	73.8
PC 4	1.0	**8.7**	82.5
PC 5	0.7	**6.3**	**88.8**
PC 6	0.4	3.5	92.3
PC 7	0.3	3.0	95.3
PC 8	0.2	1.9	97.1
PC 9	0.2	1.4	98.6
PC 10	0.1	1.1	99.7
PC 11	0.0	0.3	100.0

(c) Grupo 1 - com *deficit* hídrico

	1ªPC	2ªPC	3ªPC	4ªPC	5ªPC	6ªPC	7ªPC	8ªPC
tAlt	0.31	0.32	0.18	**0.55**	0.30	0.03	0.09	0.27
dcl	0.39	0.03	-0.12	0.20	-0.38	**0.57**	-0.09	**-0.49**
tmin	0.06	0.12	**-0.57**	-0.32	**0.56**	0.33	-0.26	0.05
tmax	**-0.43**	0.16	0.10	-0.08	-0.24	0.23	-0.05	-0.24
tPrec	-0.33	0.40	-0.15	0.06	-0.27	0.13	0.40	0.29
dprec	0.29	-0.36	0.09	-0.35	0.06	0.28	**0.71**	0.12
prec678	0.07	-0.14	**0.67**	-0.08	0.22	0.16	-0.23	-0.08
tEvap	-0.36	0.22	0.27	0.06	0.31	**0.53**	0.08	0.05
Prof	-0.19	**-0.52**	-0.03	0.09	-0.25	0.30	-0.35	**0.59**
Pedreg	**0.42**	0.25	-0.02	0.03	-0.24	0.14	-0.14	0.39
lnAwsc	-0.18	**-0.42**	-0.26	**0.63**	0.21	0.05	0.20	-0.17

(d) Grupo 2 - sem *deficit* hídrico

	1ªPC	2ªPC	3ªPC	4ªPC	5ªPC	6ªPC	7ªPC	8ªPC
tAlt	-0.17	**0.48**	-0.29	-0.06	-0.02	0.14	0.15	**0.58**
dcl	0.15	0.17	0.16	**0.84**	-0.40	-0.15	0.18	0.10
tmin	0.04	**-0.54**	0.22	0.12	-0.10	0.03	-0.29	0.01
tmax	0.36	-0.27	0.22	-0.04	0.29	0.13	**0.73**	0.14
tPrec	**0.49**	0.06	0.19	-0.17	-0.01	0.08	-0.07	0.31
dprec	-0.32	-0.33	-0.04	0.26	**0.50**	-0.21	-0.12	**0.54**
prec678	**-0.47**	-0.11	-0.16	0.14	0.10	-0.02	**0.43**	-0.33
tEvap	0.30	0.35	-0.02	0.09	**0.55**	**-0.54**	-0.06	-0.29
Prof	-0.19	0.29	**0.44**	0.17	0.32	**0.63**	-0.06	-0.17
Pedreg	0.29	-0.07	**-0.44**	**0.36**	0.29	0.41	-0.29	-0.08
lnAwsc	-0.22	0.18	**0.58**	-0.02	0.08	-0.18	-0.18	0.12

Tabela B.10: PCA Convencional. Valores próprios, percentagem de variação total e acumulada e Coordenadas das Variáveis, por grupo com e sem *deficit hídrico*.

	Valores Próprios	Percentagem de Variância	Percentagem de variância acumulada
comp 1	5.3	32.9	32.9
comp 2	3.6	22.6	55.5
comp 3	2.6	16.5	72.0
comp 4	1.1	6.7	78.7
comp 5	0.8	5.3	83.9
comp 6	0.7	4.1	88.1

Tabela B.11: PCA convencional (com variáveis categóricas). Valores próprios e percentagem de variação total.

	1ª PC	2ª PC	3ª PC	4ª PC	5ª PC
tmax	-0.87	-0.12	-0.26	-0.06	0.01
prec	0.84	0.16	0.45	0.00	-0.00
prec678	0.81	0.11	0.48	-0.01	-0.00
evap	-0.88	-0.14	-0.27	-0.05	0.01
Prof	-0.32	0.68	0.23	0.43	0.14
Pedreg	0.65	-0.41	-0.45	0.15	0.22
lnAwsc	-0.10	0.75	0.22	0.08	-0.21
MO	0.65	0.18	0.36	0.13	-0.36
Areno	-0.54	0.15	0.31	0.61	0.27
Lepto	0.04	-0.89	0.28	-0.15	-0.18
Rego	0.21	0.65	-0.63	-0.04	-0.20
Umbri	0.15	0.25	0.47	-0.44	0.64
Pedreg1	-0.68	0.37	0.47	-0.29	-0.17
Prof1	-0.30	-0.77	0.39	0.20	-0.04
Pedreg2	0.62	-0.29	-0.37	0.31	0.11
Prof2	0.34	0.44	-0.59	-0.14	0.13

Tabela B.12: PCA convencional (com variáveis categóricas). Coordenadas das Variáveis.

(a) $k = 2$

Grupo	S	prec	prec678	tmed	evap	awsc	Pedreg	Prof	alt	dcl
1	18.3	736.9	33.0	15.7	47.5	48.0	32.5	46.7	159.6	9.1
2	21.2	1609.1	111.6	14.0	32.7	52.2	48.2	40.0	278.9	16.1

(b) $k = 3$

Grupo	S	prec	prec678	tmed	evap	awsc	Pedreg	Prof	alt	dcl
1	21.2	1609.2	111.6	14.0	32.7	52.3	48.2	40.0	278.6	16.1
2	18.8	726.4	32.5	15.8	48.6	60.0	16.0	55.1	126.5	6.4
3	17.3	756.7	33.9	15.5	45.4	26.4	61.9	31.5	219.4	14.0

(c) $k = 4$

Grupo	S	prec	prec678	tmed	evap	awsc	Pedreg	Prof	alt	dcl
1	18.8	726.4	32.5	15.8	48.6	60.0	16.0	55.1	126.5	6.4
2	22.6	1642.7	112.9	13.9	32.4	74.6	34.5	48.0	275.8	16.0
3	19.7	1570.8	110.1	14.1	33.1	27.5	63.4	31.1	282.3	16.3
4	17.3	755.2	33.7	15.5	45.5	26.5	61.9	31.5	218.5	14.0

Tabela B.13: CA Convencional, método *k-means*. Valores médios das variáveis por grupo.

(a) $k = 2$

	det.	VP max.	VP min.	Mn/mn
Grupo1	0.30	2.01	0.28	7.29
Grupo2	0.03	2.68	0.06	43.29

(b) $k = 3$

	det.	VP max.	VP min.	Mn/mn
Grupo1	0.03	2.67	0.06	43.20
Grupo2	0.31	2.13	0.29	7.32
Grupo3	0.29	1.84	0.27	6.88

(c) $k = 4$

	det.	VP max.	VP min.	Mn/mn
Grupo1	0.31	2.13	0.29	7.32
Grupo2	0.03	2.81	0.04	65.11
Grupo3	0.07	2.67	0.08	32.30
Grupo4	0.30	1.79	0.27	6.55

Tabela B.14: CA Convencional, método *k-means*. Determinante, Valores Próprios máximos e mínimos e fator de restrição máximo.

(a) $k = 2$

Grupo	S	prec	prec678	tmed	evap	awsc	Pedreg	Prof	alt	dcl
1	18.4	735.2	31.7	15.8	47.5	50.0	30.4	48.8	146.2	9.1
2	21.2	1569.1	108.5	14.0	33.1	50.2	49.5	38.9	274.8	16.2
0	17.8	1268.9	87.3	14.5	41.1	46.8	45.8	32.6	354.7	10.3

(b) $k = 3$

Grupo	S	prec	prec678	tmed	evap	awsc	Pedreg	Prof	alt	dcl
1	21.2	1569.5	108.6	14.0	32.9	49.4	50.2	38.4	272.4	16.5
2	18.0	766.3	33.7	15.5	45.6	40.5	52.0	38.6	208.0	13.7
3	18.5	686.0	31.8	16.1	50.4	58.1	6.5	58.6	91.1	2.6
0	21.2	1664.2	107.7	14.0	34.9	75.2	23.6	51.0	304.3	12.3

(c) $k = 4$

Grupo	S	prec	prec678	tmed	evap	awsc	Pedreg	Prof	alt	dcl
1	17.9	757.4	33.5	15.5	45.7	39.4	52.4	38.3	204.8	13.6
2	22.4	1588.0	108.0	14.0	32.8	65.7	39.6	46.5	275.4	16.3
3	18.5	685.9	31.8	16.1	50.5	58.2	6.5	58.7	91.3	2.6
4	19.5	1557.2	111.3	14.1	32.9	24.9	66.0	27.1	270.0	16.4
0	19.5	1447.8	85.0	14.5	38.2	65.1	31.2	42.8	294.9	13.5

Tabela B.15: CA Robusta, método *Trimmed Cluster*. Valores médios das variáveis por grupo. O grupo 0 corresponde aos valores atípicos.

		Min	1º Q	Mediana	Média	3º Q	Max
Total	Resíduos	-12.3	-2.0	0.1	0.0009	2.1	11.4
	S estimado	12.6	18.2	19.4	19.6	21.0	27.0
	S	10.1	16.8	19.3	19.6	22.0	32.8
Grupo 1	Resíduos	-11.7	-1.8	0.04	0.001	1.9	10.3
	S estimado	12.6	17.5	18.4	18.3	18.9	28.2
	S	10.3	16.0	18.1	18.3	20.2	30.4
Grupo 2	Resíduos	-11.7	-2.2	0.1	0.002	2.3	10.1
	S estimado	15.7	19.5	21.1	21.2	22.6	28.1
	S	10.1	18.3	21.1	21.2	24.1	32.8

Tabela B.16: PCR convencional. Estatísticas sumárias dos resíduos e S estimado e medido por conjunto de dados.

Apêndice C

Código R

A seguir apresenta-se o código R desenvolvido para suportar a análise estatística efetuada nesta dissertação, o qual inclui várias funções e *scripts*. O código R apresentado foi desenvolvido e testado na versão 2.15.0 do R, com várias *packages* adicionais à instalação base.

Carregamento das *packages*

```
# PACKAGES PRINCIPAIS
library(FactoMineR) #Analise de Componentes Principais (PCA) Convencional, versão 1.16
library(rrcov)      #Analise de Componentes Principais (PCA) robusta, versão 1.3-01
library(robustbase) #Regressão Linear Múltipla (MLR) convencional, versão 0.7-6
library(robust)             #Regressão Linear Múltipla (MLR) robusta stepwise regression
library(tclust)             #Analise de Agrupamentos (CA) robusta.

# OUTRAS PACKAGES UTILIZADAS
library(nortest)    # Normality Lilliefors (K-S) Test
library(mvnormtest) #Normality Shapiro-Wilk test
library(geoR)               #estima parâmetro para transformação de Box-cox
library(car)               #Transformação box-cox
library(xtable)     #Conversão de tabelas R para latex
library(RODBC)      #ligação a dados em base de dados Microsoft Access
```

Amostragem dos dados. Dados para exercício e dados para validação

```
#Carregamento dos dados a partir da Base de dados
myconn <-odbcConnect("MestradoBaseDados")
dados<-sqlQuery(myconn, paste("select idparc,S_S1 as S, alt,exp,dcl,
```

```
        tmin_cor as tmin,tmax_cor as tmax,tmed_cor as tmed,
        prec_cor as prec,dprec_cor as dprec,prec678_cor as prec678,
        evap_cor as evap,
        Prof_cor as Prof,Pedreg_cor as Pedreg,awsc,x,y
from T_Q_PSF_InvCorrente_InputMod_VarAmb","order by idparc"))
close(myconn)

#SAMPLE DADOS (90%)
nrow(Dados)
ncol(Dados)
dados_treino=Dados[sample(nrow(Dados),replace=FALSE,size=0.9*nrow(Dados)),]
colnames(dados_treino)
nRegistos<-length(dados_treino[,1])
nRegistos
rowIDx<-as.numeric(rownames(dados_treino))
dados_teste<-Dados[-rowIDx,]
nRegistos<-length(dados_teste[,1])
nRegistos
rowIDx_ctr<-as.numeric(rownames(dados_teste))

#EXPORTAR DADOS PARA BASE DE DADOS (amostra e controlo)
myconn <-odbcConnect("MestradoBaseDados")
sqlSave(myconn,dados_treino,tablename = "MS_Inv_DadosTreino",append = FALSE)
sqlSave(myconn,dados_teste,tablename = "MS_Inv_DadosTeste",append = FALSE)
close(myconn)
```

Transformação dos dados

```
# S e awsc
transf.dados<-transform(dados,S_trf=S^(1/2),lnAwsc=log(awsc))
# Polinomiais
transf.dados<-transform(dados_treino,alt2=alt^2,alt3=alt^3,
        exp_pond2=exp_pond^2,exp_pond3=exp_pond^3,
        dcl2=dcl^2,dcl3=dcl^3,tmin2=tmin^2,tmin3=tmin^3,
        tmax2=tmax^2,tmax3=tmax^3,prec2=prec^2,prec3=prec^3,
        dprec2=dprec^2,dprec3=dprec^3,prec6782=prec678^2,
        prec6783=prec678^3,evap2=evap^2, evap3=evap^3,IH=prec/evap,
        IH2=(prec/evap)^2,IH3=(prec/evap)^3, tProf2=Prof^2,tProf3=Prof^3,
        tPedreg2=Pedreg^2,tPedreg3=Pedreg^3,lnAwsc2=lnAwsc^2,
        lnAwsc3=lnAwsc^3,x2=x^2,x3=x^3,y2=y^2,y3=y^3)

#TRANSFORMAçÕES (Box-Cox)
dataToTransform<-dados_treino[,c("S","alt","exp_pond","dcl",
        "tmin","tmax","tmed","prec","dprec","prec678","evap",
        "Prof","Pedreg","awsc","x","y")]
vetor_lambdas<-c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
        for (i in 1:length(dataToTransform)){
```

```
                vetor_lambdas[i]<-boxcoxfit(dataToTransform[,i])$lambda
        }
matriz_transf<-matrix(0,nrow(dataToTransform),16)
        for (i in 1:length(dataToTransform)){
                        matriz_transf[,i]<-(dataToTransform[,i])^(vetor_lambdas[i])
        }
```

Estandardização dos dados

```
# Método convencional
#Estandardização - Dados treino
dataToNorm<-dados_treino[,c("S","S_trf","alt","exp","dcl",
        "tmin","tmax","tmed","prec","dprec","prec678","evap",
        "Prof","Pedreg",
        "awsc","lnAwsc","x","y")]
DadosNorm_treino<-data.frame(scale(dataToNorm,center=TRUE,scale=TRUE),
                dados_treino[,23:45])
        nlinhas<-nrow(DadosNorm_treino)
        ncolunas<-ncol(DadosNorm_treino)
#        colnames(DadosNorm_treino)<-c(colnames(dados_treino[,c(-1,-2)]))
colnames(DadosNorm_treino)

# Método Robusto
DadosNormRob<-scale(dataToNorm,center=apply(dataToNorm,2,median),
                  scale=apply(dataToNorm,2,mad))
        nlinhas<-nrow(DadosNormRob)
        ncolunas<-ncol(DadosNormRob)
        colnames(DadosNormRob)<-c(colnames(dataToNorm))
```

Analise Exploratória de Dados - Estatísticas Sumárias

```
        #Dados<-subset(DadosCluster,ind==1)
        sumarioEstatistico<-matrix(0,length(dataToExplore),4)
        for (i in 1:length(dataToExplore)){
                mediana<-quantile(dataToExplore[,i],0.50, names=F)
                media<-mean(dataToExplore[,i])
                desvioP<-sd(dataToExplore[,i])
                MADrob<-mad(dataToExplore[,i])
                        sumarioEstatistico[i,1]<-media
                        sumarioEstatistico[i,2]<-mediana
                        sumarioEstatistico[i,3]<-desvioP
                        sumarioEstatistico[i,4]<-MADrob
        }
```

```
sumarioEstatistico
nomecolunas<-c("Media", "Mediana","Desvio Padrão","MAD")
colnames(sumarioEstatistico)<-nomecolunas
rownames(sumarioEstatistico)<-colnames(dataToExplore)
xtable(sumarioEstatistico,digits=2,
caption="Medidas de tendência central e de dispersão para a amostra de
                calibração (com 3022 observações).",
        label="tab:sumStatDataCalibra")
```

Analise Exploratória de Dados - Verificação da gráfica da Normalidade

```
v<-dataToExplore
grupo<-"tot"
# qq-plot, boxplots e histograma das variáveis
par(mfrow=(c(2,2)))
nomes<-colnames(v)
unidades<-unidadesVar
        for (i in 1:length(v)){
                boxplot(v[,i], col='grey',names=nomes[i],main="Box Plot",
                        ylab=paste(nomes[i],unidades[i],sep=" ",collapse = NULL))
                plot(v[,i],main="Data Plot", ylab=paste(nomes[i],
                        unidades[i],sep=" ",collapse = NULL), xlab="Indice")
                hist(v[,i], col='grey',names=nomes[i],main="Histograma",
                        xlab=paste(nomes[i],unidades[i],sep=" ",collapse = NULL),
                        freq=F)#probability=T, ylim=c(0,0.2)
                        lines(density(v[,i]), col="red")
                        curve(dnorm(x,mean(v[,i]),sd(v[,i])),add=T)
                oask <- devAskNewPage(dev.interactive(orNone = TRUE))
                qqnorm(v[,i])
                qqline(v[,i], col='grey')
#        savePlot(filename = paste("QuatroGraf",grupo,nomes[i],sep = "_",
#                         collapse = NULL),type = c("eps"))
        }
          devAskNewPage(oask)
par(mfrow=(c(1,1)))
```

Matriz de Correlação

```
# de Pearson
CorDados<-cor(dataToExplore[,c("S","alt","exp","dcl","tmin",
        "tmax","tmed","prec","dprec","prec678","evap",
        "Prof","Pedreg","awsc","x","y")],method="pearson" )
```

```
xtable(CorDados,
caption="Matriz Correlações (Método de Pearson)}",
                label="tab:CorSpearman")

#Fast - MCD
CorDadosNormRob<-covMcd(dataToExplore[,c("S","alt","exp","dcl","tmin",
        "tmax","prec","dprec","prec678","evap",
        "Prof","Pedreg","awsc","x","y")],
        alpha=0.75,cor=TRUE)$cor
xtable(CorDadosNormRob,
caption="Matriz Correlações robusta (MCD)}",label="tab:CorMCD")
```

Matrizes de Gráficos de Dispersão - Identificação visual de agrupamentos de dados

```
pairs(dataToExplore[,c("S","alt","dcl","exp","x")],
        col = "grey")
        savePlot(filename = "Inv_Pairs_p1",type = c("eps"))
pairs(dataToExplore[,c("S","Prof","Pedreg","awsc")],
        col = "grey")
        savePlot(filename = "Inv_Pairs_p2",type = c("eps"))
pairs(dataToExplore[,c("S","prec","dprec","prec678","evap","alt","y")],
        col = "grey")
        savePlot(filename = "Inv_Pairs_p3",type = c("eps"))
pairs(dataToExplore[,c("S","evap","tmin","tmax","tmed","alt","y")],
        col = "grey")
        savePlot(filename = "Inv_Pairs_p4",type = c("eps"))
```

Análise de Componentes Principais (ACP) - Método Convencional

```
#Package FactoMineR
dataToPCA<-DadosNorm_treino[,c("tAlt","dcl","tmin","tmax","tEvap","tPrec",
                               "dprec","prec678","Prof","Pedreg","tAwsc",
                               "x","y","MO","Antropico","Areno","Lepto","Rego",
                               "Umbri","Prof1","Prof2","Prof3","Pedreg1","Pedreg2",
                               "Pedreg3","textura")]

nCP<-length(DadosNorm_treino[1,c("tPrec","prec678","dprec","lnAwsc",
                                 "Pedreg","Prof","dcl","tAlt")])
acpFactoMineR<-PCA(DadosNorm_treino[,c("tPrec","prec678","dprec","lnAwsc",
                                 "Pedreg","Prof","dcl","tAlt","ProfSolo")],
        scale.unit = TRUE,ncp=nCP,ind.sup = NULL,quanti.sup=9,quali.sup=9,
        row.w = NULL,col.w = NULL,graph = TRUE, axes = c(1,2))
```

```
savePlot(filename = "PCA_var_fm12.eps",type = c("eps"))

#VALORES PROPRIOS POR CP
acpFactoMineR$eig
#CONTRIBUTOS DAS VARIÁVEIS POR CP
acpFactoMineR$var$contrib[,1:5]
#COORDENADAS DAS VARIÁVEIS POR CP
acpFactoMineR$var$coord[,1:5]

#GRAFICO DE DISPERSÃO DE INDIVIDUOS
plot(acpFactoMineR,axes = c(1, 2),choix = "ind",ellipse = NULL,
        xlim = c(-5,4), ylim = c(-3,3),label=NULL,col.ind = "grey")
        savePlot(filename = "PCAconv_f12ind_tot.eps",type = c("eps"))

#SCREEPLOT: GRAFICO DE VALORES PROPRIOS POR CPS
plot(acpFactoMineR$eig[,1],type="b",
                names.arg = paste("CP",1:nrow(acpFactoMineR$eig), sep = ""),
                xlab="Componente Principal (CP)",ylab="Valores Próprios")
        savePlot(filename = "PCAconv_barplot_tot.eps",type = c("eps"))
```

Análise de Componentes Principais (ACP) - Método Robusto

```
# PACKAGE: rrcov (estimador MCD, n<=m)
# PACKAGE: rrcov (estimador MCD, n<=m)
n<-nrow(dataToRPCA_Norm)
m<-length(dataToRPCA_Norm)
covControloMCD <- CovControlMcd(alpha=1,nsamp = 500,seed = NULL,
                trace= FALSE, use.correction =TRUE)
covMCDrrcov<-CovMcd(dataToRPCA_Norm,
        alpha = 1, nsamp = 500, seed = NULL, trace = FALSE,
                use.correction = TRUE, control=covControloMCD)
pcaRobrrcov<-PcaCov(dataToRPCA_Norm,k = 0, kmax = m,
        cov.control=covControloMCD,na.action = na.fail, scale = FALSE,
        signflip = FALSE, trace=FALSE,corr=TRUE)

pcaRobrrcov_sum<-summary(pcaRobrrcov)
getEigenvalues(pcaRobrrcov)
pcaRobrrcov_lodings<-getLoadings(pcaRobrrcov)
pcaRobrrcov_lodings[,1:5]

#OBSERVAÇÕES ATÍPICAS
ObsAtipicas<-getFlag(covMCDrrcov)
NObsAtipicas<-unique(which(getFlag(covMCDrrcov)==FALSE)) #,prob=0.975)
data_treino_outliers<-data.frame(dados_treino[,-1], ObsAtipicas)

#GRAFICOS:
screeplot(pcaRobrrcov, type="lines",main="Metodo Robusto")
```

```
        savePlot(filename = "PCArob_screeplot.eps",type = c("eps"))
biplot(pcaRobrrcov,col=c("grey","blue"),cex=c(0.5, 0.5),xlim=c(-0.05,0.05),
        ylim=c(-0.07,0.05))
        savePlot(filename = "PCArob_factormap1_2f.eps",type = c("eps"))
plot(covMCDrrcov, which = "dd", classic=TRUE,col="grey",cex=1)#,id.n=10)
        savePlot(filename = "PCArob_distanceConvRob.eps",type = c("eps"))
plot(covMCDrrcov, which = "distance", classic=TRUE,col="grey",cex=1)
        savePlot(filename = "PCArob_distanceConvRob2.eps",type = c("eps"))
plot(covMCDrrcov, which = "qqchi2", classic=TRUE,col="grey",cex=1)
        savePlot(filename = "PCArob_QQplotConvRob.eps",type = c("eps"))


#SCORES
scoresRPCA<-getScores(pcaRobrrcov) #Dados de input para RPCR (RMLR EM CPS)
```

Análise de Agrupamentos (CA) Convencional.

```
set.seed(10)
# DETERMINAR O NÚMERO DE GRUPOS
# within-groups total sum squares
wss <- (nrow(x)-1)*sum(apply(x,2,var)) #n-1 vezes somatório variaças
# Soma dos quadrados para k de 2 a 15 grupos
for (i in 2:15) wss[i]<-sum(kmeans(x,centers=i,nstart=100)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares")
savePlot(filename = "Cluster_kmeans_NGoups_wss",type = c("eps"))

ngrupos<-2
# K-Means Cluster Analysis
fit <- kmeans(x, ngrupos,nstart=100,algorithm ="Lloyd")

# append cluster assignment
x2 <- data.frame(x, fit$cluster)
x1 <- data.frame(dados_treino, fit$cluster)

# get cluster means (em unidades reais??)
fitmed_g<-aggregate(x1[,c("S","prec","prec678","tmed","evap",
        "awsc","Pedreg","Prof","alt","dcl")],
        by=list(fit$cluster),FUN=mean)
fitmed_g
xtable(fitmed_g,caption=paste("Valores médios das variáveis por grupo - k=",
        ngrupos,sep=""),label=paste("tab:InvID10_fitmed_g",ngrupos,sep=""),
        digits=1)

#EXPORTAR DADOS PARA BASE DE DADOS
myconn <-odbcConnect("MestradoBaseDados")
```

```
sqlDrop(myconn, paste("MS_InvTot_kmeans_g",ngrupos,sep=""))
sqlSave(myconn,x1[,c(2:52)],tablename =paste("MS_InvTot_kmeans_g",
                        ngrupos,sep=""),append = FALSE)
close(myconn)

# Cluster Plot against 1 st 2 principal components
clusplot(x2, fit$cluster, color=TRUE, shade=TRUE,labels=2, lines=0)
        savePlot(filename = paste("Cluster_kmeans_g",ngrupos,"_PlotGrups",sep=""),
                type = c("eps"))

#ANALISE DE VALORES PROPRIOS E DETERMINANTES
lista_det<-vector(mode = "logical", length = 5)
lista_eigen_max<-vector(mode = "logical", length = 5)
lista_eigen_min<-vector(mode = "logical", length = 5)
for (i in 1:ngrupos){
g<-subset(x2,fit.cluster==i)
lista_det[i]<-det(cor(g[,c("tPrec","prec678","dprec","lnAwsc",
                                        "Pedreg")]))
lista_eigen_max[i]<-max(eigen(cor(g[,c("tPrec","prec678","dprec","lnAwsc",
                                        "Pedreg")]))$values)
lista_eigen_min[i]<-min(eigen(cor(g[,c("tPrec","prec678","dprec","lnAwsc",
                                        "Pedreg")]))$values)
}
lista_det
lista_eigen_max
lista_eigen_min
Razao_eigen_tot<-max(lista_eigen_max)/min(lista_eigen_min)
Razao_eigen_g1<-lista_eigen_max[1]/lista_eigen_min[1]
Razao_eigen_g2<-lista_eigen_max[2]/lista_eigen_min[2]
Razao_eigen_g3<-lista_eigen_max[3]/lista_eigen_min[3]
Razao_eigen_g4<-lista_eigen_max[4]/lista_eigen_min[4]
lista_Reigen_g<-list(Razao_eigen_g1,Razao_eigen_g2,Razao_eigen_g3,
                                Razao_eigen_g4)
lista_Reigen_g
#d
VP_det<-matrix(0,4,4)
for (i in 1:4){
        VP_det[i,1]<-lista_det[i]
        VP_det[i,2]<-lista_eigen_max[i]
        VP_det[i,3]<-lista_eigen_min[i]
        VP_det[i,4]<-lista_Reigen_g[[i]]
}
colnames(VP_det)<-c("det.","VP max.","VP min.","Mn/mn")
rownames(VP_det)<-c("Grupo1","Grupo2","Grupo3","Grupo4")
xtable(VP_det,caption="Determinante, Valores Próprios maximos e mínimos e
                factor de restrição máximo",label="tab:InvID10_VPedet")

#CAIXA DE BIGODES POR GRUPO
g1<-subset(x1,fit.cluster==1)
g2<-subset(x1,fit.cluster==2)
```

```
g3<-subset(x1,fit.cluster==3)
g4<-subset(x1,fit.cluster==4)
boxplot(list(g1$S,g2$S,g3$S,g4$S),
        col=c("green","cyan3","gray","orange3"),
        names=c("G1","G2","G3","G4"),
        pars = list(boxwex = 0.5,staplewex = 0.2,outwex = 0.5),horizontal = F,
        main = "Caixa de Bigodes Site Index por Grupos",sub="Método K-means",
        ylab = "S (m)",ylim = c(10, 35), axes=F)
legend(3.1, 34, c("G1", "G2","G3","G4"),
        fill = c("green","cyan3","gray","orange3"))
axis(2,at=seq(10,35,by=1),col="grey",labels=T)
savePlot(filename = paste("Cluster_kmeans_g",ngrupos,"_BoxplotGrupos",sep=""),
                type = c("eps"))
```

Análise de Agrupamentos (CA) Robusta.

```
set.seed(10)
#CTL CURVES
factorRest<-44
metodo<-"eigen"
plot (ctlcurves (x, k = 1:8, alpha = seq (0, 0.30, by = 0.05),restr=metodo,
                                restr.fact=factorRest,nstart = 400, iter.max = 160))
savePlot(filename = paste("Cluster_tclust_",metodo,factorRest,"_CTL",sep="",
                collapse = NULL),type = c("eps"))
ngrupos<-3
alfa<-0.05

#MÉTODO tclust TRIMMED
res.a <- tclust (x, k = ngrupos, alpha = alfa, restr.fact = factorRest,
                restr = metodo, equal.weights = TRUE,nstart=1600,iter.max=640)
res.a$cov
plot(res.a, main = "/r")
savePlot(filename=paste("Cluster_tclust_",metodo,factorRest,"_g",ngrupos,
                sep="",collapse = NULL),type = c("eps"))

discr.res.a <- DiscrFact (res.a, threshold = 0.1)
plot (discr.res.a)
savePlot(filename = paste("Cluster_tclust_",metodo,factorRest,
                        "_g",ngrupos,"_discrimFactor",sep="",collapse = NULL),
                        type = c("eps"))
summary(discr.res.a)

discr.res.a$ind
x1 <- data.frame(x, discr.res.a$ind)
x2 <- data.frame(dados_treino, discr.res.a$ind)
```

```
#EXPORTAR DADOS PARA BASE DE DADOS (data_test_grupos)
myconn <-odbcConnect("MestradoBaseDados")
sqlDrop(myconn, paste("MS_InvTot_tclust_g",ngrupos,sep=""))
sqlSave(myconn,x2[,2:52],
        tablename = paste("MS_InvTot_tclust_g",ngrupos,sep=""),append = FALSE)
close(myconn)

#ANALISE DE VALORES PROPRIOS E DETERMINANTES
lista_det<-vector(mode = "logical", length = 5)
lista_eigen_max<-vector(mode = "logical", length = 5)
lista_eigen_min<-vector(mode = "logical", length = 5)
for (i in 1:ngrupos){
g<-subset(x1,discr.res.a.ind==i)
lista_det[i]<-det(cor(g[,c("tPrec","prec678","dprec","lnAwsc",
                                        "Pedreg")]))
lista_eigen_max[i]<-max(eigen(cor(g[,c("tPrec","prec678","dprec","lnAwsc",
                                        "Pedreg")]))$values)
lista_eigen_min[i]<-min(eigen(cor(g[,c("tPrec","prec678","dprec","lnAwsc",
                                        "Pedreg")]))$values)
}
lista_det
lista_eigen_max
lista_eigen_min
Razao_eigen_tot<-max(lista_eigen_max)/min(lista_eigen_min)
Razao_eigen_g1<-lista_eigen_max[1]/lista_eigen_min[1]
Razao_eigen_g2<-lista_eigen_max[2]/lista_eigen_min[2]
Razao_eigen_g3<-lista_eigen_max[3]/lista_eigen_min[3]
Razao_eigen_g4<-lista_eigen_max[4]/lista_eigen_min[4]
Razao_eigen_g5<-lista_eigen_max[5]/lista_eigen_min[5]
lista_Reigen_g<-list(Razao_eigen_g1,Razao_eigen_g2,Razao_eigen_g3,
                                Razao_eigen_g4,Razao_eigen_g5)
lista_Reigen_g
VP_det<-matrix(0,5,4)
for (i in 1:5){
        VP_det[i,1]<-lista_det[i]
        VP_det[i,2]<-lista_eigen_max[i]
        VP_det[i,3]<-lista_eigen_min[i]
        VP_det[i,4]<-lista_Reigen_g[[i]]
}
colnames(VP_det)<-c("det.","VP max.","VP min.","Mn/mn")
rownames(VP_det)<-c("Grupo1","Grupo2","Grupo3","Grupo4","Grupo5")
xtable(VP_det,caption="Determinante, Valores Próprios maximos e mínimos e
        factor de restrição máximo",label=paste("tab:InvID10_VPedet_",metodo,
        factorRest,"_k",ngrupos,"_alfa",alfa, sep=""),digits=2)

#CAIXA DE BIGODES POR GRUPO
g1<-subset(x2,discr.res.a$ind==1)
g2<-subset(x2,discr.res.a$ind==2)
g3<-subset(x2,discr.res.a$ind==3)
g4<-subset(x2,discr.res.a$ind==4)
```

```
g5<-subset(x2,discr.res.a$ind==5)
g0<-subset(x2,discr.res.a$ind==0)
# Expand right side of clipping rect to make room for the legend
par(xpd=T, mar=par()$mar+c(0,0,0,4.5))
#"cyan3","green","orange","gray","white"
boxplot(list(g1$S,g2$S,g3$S,g4$S,g0$S),
        col=c("orange","cyan3","orange","green","white"),
        names=c("G1","G2","G3","G4","Outliers"),
        pars = list(boxwex = 0.5,staplewex = 0.2,outwex = 0.5),horizontal = F,
        main = "Caixa de Bigodes de S por Grupo",
        sub=paste("tclust - restr=",metodo,",","restr.fact=",factorRest,sep=""),
        ylab="S(m)",ylim=c(as.integer(min(dados_treino$S)),
        as.integer(max(dados_treino$S))),axes=F)
        legend(6, 30, c("G1", "G2", "G3","G4","Outliers"),
        fill = c("orange","cyan3","orange","green","white"))
        axis(2,at=seq(as.integer(min(dados_treino$S)),
        as.integer(max(dados_treino$S)),by=1),col="grey",labels=T)
savePlot(filename = paste("Cluster_tclust_",metodo,factorRest,"_g",ngrupos,
                        "_BoxplotGrupos",sep=""),type = c("eps"))
# Restore default clipping rect
par(mar=c(5, 4, 4, 2) + 0.1)

# Exportar dados para base de dados
myconn <-odbcConnect("MestradoBaseDados")
sqlDrop(myconn, paste("MS_Inv_tclust_",metodo,factorRest,"_g",ngrupos,sep=""))
sqlSave(myconn,x2[,c(2:52)],tablename =paste("MS_Inv_tclust_",metodo,factorRest,
                "_g",ngrupos,sep=""),append=FALSE)
close(myconn)

# get cluster means (em unidades reais??)
fitmed_g<-aggregate(x2[,c("S","prec","prec678","tmed","evap",
        "awsc","Pedreg","Prof","alt","dcl")],by=list(discr.res.a$ind),
        FUN=mean)
fitmed_g
xtable(fitmed_g,caption=paste("Valores médios das variáveis por grupo - k=",
                ngrupos,sep=""),label=paste("tab:InvID10_tclust_fitmed_",metodo,
                factorRest,"_g",ngrupos,"_alfa",alfa,sep=""),digits=1)
```

Regressão Linear Múltipla (MLR) convencional

```
#Package MASS lm()
#MODELO INICIAL
model1 <- lm(S ~ dcl + tmin + tmax + tPrec + dprec + prec678 +
tEvap + Prof + Pedreg + lnAwsc + MO + Antropico + Areno + Lepto +
Rego + Umbri + Cambi + Prof1 + Prof2 + Pedreg1 + Pedreg2 + textura +
        x + y + tAlt,data = dados_treino)
```

```
        summary(model1)

# STEPWISE REGRESSION
step <- stepAIC(model1, direction="both")
step$anova # display results

#MODELO FINAL
model1 <- lm(S ~ dcl + tmax + tPrec + dprec + lnAwsc + MO +
    Lepto + Rego + Umbri + Prof2 + Pedreg1  + tAlt
,data = dados_treino)
        summary(model1)
#MODELO FINAL (TABELA ANOVA)
model1 <- aov(S ~ dcl + tmax + tPrec + dprec + lnAwsc + MO +
    Lepto + Rego + Umbri + Prof2 + Pedreg1  + tAlt
,data = DadosNorm_treino)
        summary(model1)
model1$coef
hist(model1$resid)

#TESTE NORMALIDADE DOS RESÍDUOS
residuos<-model1$resid # residuals
lillie.test(residuos)
#ks.test(residuos,"pnorm",mean(residuos),sd(residuos))
shapiro.test(residuos)

#GRÁFICOS DE DIAGNÓSTICO
        par(mfrow=c(2,2))
        plot(model1)
        savePlot(filename = "MLRconv_grafDiag_g2.eps",type = c("eps"))
        par(mfrow=c(1,1))
```

Regressão Linear Múltipla (MLR) robusta

```
#(método MM)
#MODELO INICIAL
lmRob_robust<-lmRob(S ~ tAlt + dcl + tmax + tPrec + dprec + prec678 +
tEvap + Prof + Pedreg + lnAwsc + x + y + MO + Antropico + Areno + Lepto +
Rego + Umbri + Cambi + Prof1 + Prof2 + Pedreg1 + Pedreg2 + textura
                ,data = DadosNormRob_treino)
summary(lmRob_robust)

#STEPWISE REGRESSION
step.lmRob(lmRob_robust, scale=FALSE,
direction = "backward",trace = TRUE, keep = NULL, steps = 1000, fast = FALSE)

#MODELO FINAL:
```

```
model1rob <- lmrob(S ~ -1 + dprec + alt + dcl +
                        lnAwsc + Prof1 + Pedreg1 + MO + Cambi + textura
        , data = DadosNormRob_treino)
        summary(model1rob)

#GRÁFICOS DE DIAGNÓSTICO
        par(mfrow=c(2,2))                    # visualize four graphs at once
        plot(model1rob,which = 2:5)
        savePlot(filename = "MLRrob_grafDiag.eps",type = c("eps"))
        par(mfrow=c(1,1))
```

Regressão em Componentes principais (PCR) convencional

```
set.seed(10)
colnames(dados_treino)
nrow(dados_treino)
Xr<-as.matrix(dados_treino[,c(
"tAlt","dcl","tmin","tmax","tEvap","tPrec","dprec","prec678",
"Prof","Pedreg","lnAwsc","x","y")])
colnames(Xr)
Sest<-as.vector(dados_treino[,"S"])
summary(Sest)
# R PACKAGE chemometrics
#NÚMERO OPTIMO DE COMPONENTES PRINCIPAIS
#Repeated double Cross-Validation for multivariate regression methods,
#like PLS and PCR (mvr_dcv)
pcr_dcv<-mvr_dcv(S ~ Xr, data=dados_treino,ncomp=13,method="svdpc",
                        plot.opt=TRUE)
        savePlot(filename = "PCR_PCs_MSEP.eps",type = c("eps"))#Total
        savePlot(filename = "PCR_PCs_MSEP_g1.eps",type = c("eps"))#Grupo1
        savePlot(filename = "PCR_PCs_MSEP_g2.eps",type = c("eps"))#Grupo2
str(pcr_dcv)
par(mfrow=c(1,1))
#Component plot for repeated DCV: nº ótimo de comp. por DCV
pcr_plot2<-plotcompmvr(pcr_dcv)
        savePlot(filename = "PCR_PCs_relFreqOptPC.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_relFreqOptPC_g1.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_relFreqOptPC_g2.eps",type = c("eps"))
#Plot SEP from repeated DCV:
pcr_plot1<-plotSEPmvr(pcr_dcv,opt=pcr_plot2$opt,Sest,Xr,method="svdpc")
        savePlot(filename = "PCR_PCs_SEP.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_SEP_g1.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_SEP_g2.eps",type = c("eps"))
#MEDIDOS VS ESTIMADOS
#Plot predictions from repeated DCV:
plotpredmvr(pcr_dcv,opt=2,Sest,Xr,method="svdpc")
```

```
        savePlot(filename = "PCR_PCs_medEst.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_medEst_g1.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_medEst_g2.eps",type = c("eps"))
#RESÍDUOS
#plot residuos from repeated DCV:
plotresmvr(pcr_dcv, opt=2,Sest,Xr,method="svdpc")
        savePlot(filename = "PCR_PCs_residuos.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_residuos_g1.eps",type = c("eps"))
        savePlot(filename = "PCR_PCs_residuos_g2.eps",type = c("eps"))
#??
(pcr_dcv)
residuos<-pcr_dcv$resopt
summary(pcr_dcv$resopt)
summary(pcr_dcv$predopt)
summary(dados_treino$S)
```

Regressão em Componentes principais (PCR) robusta

```
#REGRESSÃO EM COMPONENTES PRINCIPAIS ROBUSTA (RPCR)
# 1º ANALISE DE COMPONENTES PRINCIPAIS ROBUSTA (RPCA)
# PACKAGE: rrcov (estimador MCD, n<=m)
n<-nrow(dataToRPCA_Norm)
m<-length(dataToRPCA_Norm)
covControloMCD <- CovControlMcd(alpha=1,nsamp = 500,seed = NULL,
                trace= FALSE, use.correction =TRUE)
covMCDrrcov<-CovMcd(dataToRPCA_Norm,
        alpha = 1, nsamp = 500, seed = NULL, trace = FALSE,
                use.correction = TRUE, control=covControloMCD)
pcaRobrrcov<-PcaCov(dataToRPCA_Norm,k = 0, kmax = m,
        cov.control=covControloMCD,na.action = na.fail, scale = FALSE,
        signflip = FALSE, trace=FALSE,corr=TRUE)

pcaRobrrcov_sum<-summary(pcaRobrrcov)
pcaRobrrcov_sum
getEigenvalues(pcaRobrrcov)
pcaRobrrcov_lodings<-getLoadings(pcaRobrrcov)
pcaRobrrcov_lodings[,1:5]

#OBTER SCORES POR COMPONENTE PRINCIPAL
scoresRPCA<-getScores(pcaRobrrcov)
colnames(scoresRPCA)
nrow(scoresRPCA)

#
dataToRMLR<-data.frame(DadosNorm_treino[,"S"],scoresRPCA)
colnames(dataToRMLR)<-c("S","CP1","CP2","CP3","CP4","CP5","CP6","CP7","CP8")
```

```
colnames(dataToRMLR)
# 2º REGRESSÃO LINEAR MULTIVARIADA ROBUSTA (RMLR) (método MM)
#Modelo inicial, update e comparação de modelos
lmRob_robust<-lmRob(S ~ I(CP1^2) + CP2 + CP3 + CP4 + CP5 + CP6 + CP7
                ,data = dataToRMLR)
summary(lmRob_robust)
step.lmRob(lmRob_robust, scale=FALSE,
direction = "backward",trace = TRUE, keep = NULL, steps = 1000, fast = FALSE)

model1rob <- lmrob(S ~ I(CP1^2) + CP2 + CP3 + CP4 + CP5 + CP6 + CP7,
                         data = dataToRMLR)
        summary(model1rob)

#GRAFICOS DE DIAGNOSTICO
        par(mfrow=c(2,2))                    # visualize four graphs at once
        plot(model1rob,which = 2:5)
        savePlot(filename = "MLRrob_grafDiag_rpcr.eps",type = c("eps"))
        par(mfrow=c(1,1))
```

Análise Discriminante Quadrática - Convencional e Robusta

```
#ANÁLISE DISCRIMINANTE QUADRÁTICA

#IMPORTAR DADOS DA BASE DE DADOS (dados treino e k=2)
myconn <-odbcConnect("MestradoBaseDados")
dados_treino_conv<-sqlQuery(myconn,paste("Select * from MS_InvTot_kmeans_g2"))
dados_treino_rob<-sqlQuery(myconn,paste("Select * from
                MS_Inv_tclust_eigen44_g2"))
dados_teste<-sqlQuery(myconn,paste("Select * from MS_Inv_DadosTeste"))
close(myconn)

#DADOS
d1.conv<-dados_treino_conv[,c("S","tAlt","dcl","tmin","tmax","tPrec",
        "dprec","prec678","tEvap","Prof","Pedreg","lnAwsc")]
d1.rob<-dados_treino_rob[,c("S","tAlt","dcl","tmin","tmax","tPrec",
        "dprec","prec678","tEvap","Prof","Pedreg","lnAwsc")]
d2<-dados_teste[,c("S","tAlt","dcl","tmin","tmax","tPrec",
        "dprec","prec678","tEvap","Prof","Pedreg","lnAwsc")]
grupo.conv<-as.factor(dados_treino_conv$fitcluster)
grupo.rob<-as.factor(dados_treino_rob$discrresaind)

#CONVENCIONAL
qda.conv1<-qda(d1.conv,grupo.conv,CV=TRUE)
qda.conv2<-qda(d1.conv,grupo.conv,CV=FALSE)
d2Class.conv<-predict(qda.conv2,d2)$class

#Probabilidades calculadas com base na classificação dos dados (dois grupos)
```

```
qda.conv2$prior
#G1=55% e G2=45%
#valores médios do Site index
qda.conv2$means
#G1=18.2 e G2=21.2

#OUTRAS ANALISES
# Assess the accuracy of the prediction
# percent correct for each category of G
ct <- table(grupo, z2$class)
diag(prop.table(ct, 1))
# total percent correct
sum(diag(prop.table(ct)))

# VISUALIZAÇÃO DE RESULTADOS
# Scatter plot using the 1st two discriminant dimensions
plot(z2) # fit from lda

#ROBUSTA
rqda<-QdaCov(d1.rob,grupo.rob)
d2Class.rob<-predict(rqda,d2)@classification

#prior

#valores médios do Site index
#G1=18.3 e G2=20.5

#DADOS TESTE CLASSIFICADOS
dados_teste_class <- data.frame(dados_teste,d2Class.conv, d2Class.rob)
colnames(dados_teste_class)

#EXPORTAR PARA BASE-DE-DADOS (QDA CONVENCIONAL E ROBUSTA)
myconn <-odbcConnect("MestradoBaseDados")
sqlDrop(myconn,"MS_Inv_QDA")
sqlSave(myconn,dados_teste_class[,-1],tablename ="MS_Inv_QDA",append = FALSE)
close(myconn)
```

Validação dos Modelos

```
#TABELA PARA COMPARARAR EFICIENCIA DOS MODELOS
matriz.residuos<-matrix(0,nrow(dados_treino),5)
for (i in 1:nrow(dados_treino)){
matriz.residuos[i,1]<-dados_treino[i,"S"]
matriz.residuos[i,2]<-dados_treino[i,"S"]-mean(dados_treino[,"S"])
matriz.residuos[i,3]<-pcr_dcv$predopt[i]
matriz.residuos[i,4]<-pcr_dcv$predopt[i]-mean(pcr_dcv$predopt)
matriz.residuos[i,5]<-dados_treino[i,"S"]-mean(pcr_dcv$predopt)
```

```
        }
matriz.residuos
colnames(matriz.residuos)<-c("Qi","Ri","Qiest","Riest","RiObsEst")

#Coeficiente de determinação de Pearson
R2<-function(Ri,Riest){
(sum(Ri*Riest)/(sqrt(sum(Ri^2)*sum(Riest^2))))^2
}
#Coeficiente de Eficiencia
CE<-function(RiObsEst,Ri){
1-(sum(RiObsEst^2)/sum(Ri^2))
}
#Raiz quadrada do erro quadrático médio
RMSE<-function(RiObsEst){
sqrt(sum(RiObsEst^2)/nrow(dados_teste))
}
#Raiz relativa quadrada do erro quadrático médio
RRMSE<-function(RiObsEst){
(sqrt(sum(RiObsEst^2)/nrow(dados_teste)))/mean(dados_teste[,"S"])
}
p<-14 #mudar nº coeficientes no modelo
n<-nrow(dados_teste)
#Critério de informação de Akaike
AIC<-function(RMSE){
nrow(dados_teste)*log(RMSE)+2*p
}
#Critério de informação Baysiano
BIC<-function(RMSE){
n*log(RMSE)+p*log(n)
}
#Coeficiente de determinação ajustado
R2adj<-function(R2){
1-(1-R2)*((n-1)/(n-p-1))
}

matrix.validacao<-matrix(0,7,1)
N<-1        #número de modelos
for (i in 1:N){
matrix.validacao[1,i]<-R2(matriz.residuos[,"Ri"],matriz.residuos[,"Riest"])
matrix.validacao[2,i]<-CE(matriz.residuos[,"RiObsEst"],matriz.residuos[,"Ri"])
matrix.validacao[3,i]<-RMSE(matriz.residuos[,"RiObsEst"])
matrix.validacao[4,i]<-RRMSE(matriz.residuos[,"RiObsEst"])
matrix.validacao[5,i]<-AIC(RMSE(matriz.residuos[,"RiObsEst"]))
matrix.validacao[6,i]<-BIC(RMSE(matriz.residuos[,"RiObsEst"]))
matrix.validacao[7,i]<-R2adj(R2(matriz.residuos[,"Ri"],matriz.residuos[,"Riest"])
)
}
rownames(matrix.validacao)<-c("R2","CE","RMSE","RRMSE","AIC","BIC","R2adj")
matrix.validacao
```