



Classificação de grãos de café de acordo com a origem geográfica: uma abordagem estatística.

Análise de regiões produtoras do Havai, Brasil, Etiópia e Quênia

Jéssica Filipa Antunes Vieira

Dissertação para obtenção do Grau de Mestre em

Matemática e Aplicações

Perfil em Probabilidades e Estatística Matemática

Orientadora: Doutora Maria do Rosário de Oliveira Silva
Co-Orientadores: Doutora Carla Isabel Igreja Rodrigues
Doutor Manuel João Cabral Morais

Júri

Presidente: Doutor António Manuel Pacheco Pires
Orientador: Doutora Maria do Rosário de Oliveira Silva
Vogais: Doutor José Domingos Cochicho Ramalho
Doutora Maria Manuela Souto de Miranda

Dezembro 2015

For millions all over the world, sitting in a café with a delicious coffee is one of life's great pleasures.

ANETTE MOLDVAER

Agradecimentos

Agradeço a tudo e todos que me inculiram o gosto pela aprendizagem e assim fizeram com que chegasse até aqui. Desde já agradeço também aos que no presente me desafiam e no futuro me desafiarão nesta procura pelo conhecimento.

Obrigada a quem me deu a oportunidade e acompanhou na realização desta dissertação, nomeadamente à minha orientadora, Professora Doutora M. Rosário Oliveira, à minha co-orientadora, Doutora Carla Rodrigues, e ao meu co-orientador, Professor Doutor Manuel Cabral Moraes.

Por fim agradeço a todos que de alguma forma contribuíram para chegar até aqui. Aos meus amigos, colegas e família, ao Derby, a todos os professores que tive ao longo do meu percurso escolar e às minhas irmãs.

Obrigada a si.

Pai e Mãe, obrigada.

Resumo

A qualidade e o preço do café dependem fortemente da região onde o grão se desenvolve. Consequentemente, neste trabalho estudou-se o problema da classificação da origem geográfica de grãos verdes de café, recorrendo: à composição isotópica do carbono ($\delta^{13}C$), do oxigénio ($\delta^{18}O$), do azoto ($\delta^{15}N$) e do enxofre ($\delta^{34}S$) do grão verde; à razão isotópica do estrôncio ($^{87}Sr/^{86}Sr$); à composição isotópica do carbono ($\delta^{13}C_{caf}$), do oxigénio ($\delta^{18}O_{caf}$) e do azoto ($\delta^{15}N_{caf}$) na molécula de cafeína do grão verde de café; e ainda às percentagens de carbono (C) e de azoto (N). A partir de um conjunto de dados com 138 observações (com valores omissos) nestas dez variáveis, provenientes de quatro origens geográficas diferentes (Havai, Brasil, Etiópia e Quénia), realizaram-se diversas análises estatísticas com o intuito de resolver o problema.

Na análise destes dados, recorreu-se a correlações clássicas e robustas, análise em componentes principais e deteção de *outliers* com base em componentes principais robustas, entre outros métodos. Por fim, realizou-se uma análise de agrupamentos e classificação.

Este trabalho permitiu identificar as composições isotópicas do oxigénio ($\delta^{18}O_{caf}$) e do azoto ($\delta^{15}N_{caf}$) na molécula de cafeína do grão verde de café, bem como a razão isotópica do estrôncio ($^{87}Sr/^{86}Sr$) como as mais promissoras para a diferenciação da origem geográfica de grãos verdes de café. A composição isotópica do enxofre ($\delta^{34}S$) foi identificada como importante na diferenciação da ilha Havai das restantes ilhas do arquipélago do Havai.

De entre os vários classificadores de grãos verdes de café de acordo com a origem geográfica que foram analisados, o que foi selecionado classificou corretamente 91% das observações (valor estimado).

Palavras-chave: Café, Composição isotópica, Classificação, Grão verde de café, Origem geográfica

Abstract

The quality and price of the coffee largely depend on its geographical origin. Consequently, the problem in the present work was the geographical classification of green coffee beans, falling back on: the isotopic composition of carbon ($\delta^{13}C$), oxygen ($\delta^{18}O$), nitrogen ($\delta^{15}N$) and sulfur ($\delta^{34}S$) in the green coffee beans; the isotope ratio of strontium ($^{87}Sr/^{86}Sr$); the isotope compositions of carbon ($\delta^{13}C_{caf}$), oxygen ($\delta^{18}O_{caf}$) and nitrogen ($\delta^{15}N_{caf}$) of the caffeine molecule of the green coffee bean; and the carbon and nitrogen percentages. Taking advantage of 138 observations (with missing values) in these 10 variables, from four distinct geographic origins of coffee (Hawaii, Brazil, Ethiopia and Kenya), several statistical methods were applied to analyse the problem.

The analysis of this data set comprised classical and robust correlations between variables, principal component analysis and outlier detection with principal component analysis were analyzed, among other methods. Lastly, cluster analysis and classification were performed.

With this work, the oxygen ($\delta^{18}O_{caf}$) and the nitrogen ($\delta^{15}N_{caf}$) isotopic compositions of the caffeine molecule extracted from the green coffee bean and the isotope ratio of strontium ($^{87}Sr/^{86}Sr$) were identified as the most promising variables in the geographical differentiation of green coffee beans. Moreover, the isotopic composition of sulfur ($\delta^{34}S$) was identified as important in the differentiation between the Hawaii Island and the rest of the islands of Hawaiian archipelago.

Among the several applied classifiers of geographical differentiation of green coffee beans, the one we selected led to the correct classification of 91% of the considered observations (estimated value).

Keywords: Coffee, Isotope composition, Classification, Green coffee bean, Geographical origin

Conteúdo

Agradecimentos	i
Resumo	iii
Abstract	v
Lista de tabelas	ix
Lista de figuras	xi
Lista de siglas e símbolos	xiii
1 Introdução	1
1.1 O café	1
1.1.1 História	1
1.1.2 Do cultivo ao consumo	2
1.1.3 O café no mundo	3
1.2 Principais objetivos e conteúdos da dissertação	5
2 Discussão do problema real	7
2.1 Classificação de grãos de café de acordo com a origem geográfica	7
2.1.1 Revisão da literatura	7
2.2 Características do grão de café: composição isotópica	9
2.2.1 Composição isotópica do carbono	11
2.2.2 Composição isotópica do azoto	11
2.2.3 Composição isotópica do oxigénio e do hidrogénio	12
2.2.4 Composição isotópica do enxofre	12
2.2.5 Composição isotópica do estrôncio	13
2.3 Descrição dos dados	13
2.3.1 Regiões produtoras de café: Brasil, Etiópia, Havai e Quênia	14
3 Análise preliminar dos dados	16
3.1 O problema dos valores omissos	16
3.2 Análise descritiva	18
3.3 Análise em componentes principais	24
3.3.1 Introdução	24
3.3.2 Resultados da aplicação da ACP aos dados	26

3.4	Deteção de <i>outliers</i>	34
3.4.1	Introdução	34
3.4.2	Resultados da aplicação dos métodos de deteção de <i>outliers</i> ao conjunto de dados	36
4	Agrupamentos	44
4.1	Introdução	44
4.2	Resultados da aplicação aos dados	45
5	Classificação	50
5.1	Seleção de variáveis	52
5.2	Métodos de classificação	53
5.2.1	Árvores de decisão	53
5.2.2	K vizinhos mais próximos	54
5.2.3	Análise discriminante	55
5.2.4	Regressão logística	56
5.2.5	Regressão logística multivariada com incorporação de seleção de variáveis – SMLR e RMLR.	58
5.3	Resultados e discussão	60
5.3.1	Desempenho dos diversos classificadores	62
5.3.2	Árvores de decisão	63
5.3.3	K vizinhos mais próximos	66
5.3.4	Análise discriminante quadrática	66
5.3.5	Regressão logística	67
5.3.6	Regressão logística multivariada com incorporação de seleção de variáveis – SMLR e RMLR	68
5.3.7	Seleção do melhor classificador	70
6	Comentários finais	72
	Referências	75
A	Apêndice	79
A.1	Análise descritiva	79
A.2	Deteção de <i>outliers</i>	81
A.3	Agrupamentos	85

Lista de tabelas

2.1	Resumo da forma como as razões isotópicas são afetadas (fracionadas) no ambiente e como podem ser exploradas para determinação da proveniência de alimentos (adaptado de Kelly <i>et al.</i> , 2005).	10
3.1	Proporção de valores omissos em cada variável.	16
3.2	Conjuntos de variáveis formados.	18
3.3	Número de observações sem valores omissos em cada subconjunto.	18
3.4	Medidas de localização e dispersão no conjunto de dados.	19
3.5	Combinações de métodos de estimação de ACP aplicadas com indicação dos pacotes do <i>software</i> R utilizados.	26
3.6	Variância estimada e proporção cumulativa de variabilidade total explicada pelas componentes principais da ACP clássica, <i>set</i> ₃ (dados padronizados).	27
3.7	<i>Loadings</i> e correlações amostrais entre as CPs e as variáveis, correspondentes à ACP clássica aplicada ao <i>set</i> ₃ (dados padronizados).	28
3.8	Variância estimada e proporção cumulativa de variabilidade total explicada pelas componentes principais da ACP PP aplicada ao <i>set</i> ₉ (dados padronizados).	30
3.9	<i>Loadings</i> e correlações amostrais entre as CPs e as variáveis, correspondentes à ACP PP aplicada ao <i>set</i> ₉ (dados padronizados).	30
3.10	Variância estimada e proporção cumulativa de variabilidade total explicada pelas componentes principais da ACP PP aplicada ao <i>set</i> ₁₀ (dados padronizados).	32
3.11	<i>Loadings</i> e correlações amostrais entre as CPs e as variáveis, correspondentes à ACP PP aplicada ao <i>set</i> ₁₀ (dados padronizados).	32
3.12	Número de CPs retidas para cada conjunto de variáveis e respectivas proporções de variação total explicada.	37
3.13	Percentagens de <i>outliers</i> por país e global.	38
4.1	Métodos de agrupamento, dissimilaridades, funções e pacotes do R aplicados nos subconjuntos listados.	45
4.2	Sumário das partições em grupos selecionadas.	46
4.3	Cruzamento entre as origens geográficas e o agrupamento obtido.	47

4.4 Média e variância de cada componente principal retida, obtida da ACP do <i>set</i> ₃ , em cada grupo formado.	48
5.1 Matriz de confusão.	50
5.2 Classes e subconjuntos de variáveis estudados para cada método de classificação. . . .	61
5.3 Medidas de desempenho dos classificadores selecionados, treino sem valores omissos e sem as observações da Etiópia.	62
5.4 Medidas de desempenho das árvores de decisão selecionadas (sem e com divisão do arquipélago do Havai).	64
5.5 Matriz de confusão obtida após <i>leave-one-out</i> do classificador kNN, com $k = 3$, aplicado aos <i>scores</i> da ACP do <i>set</i> ₁₀	66
5.6 Intervalos de confiança a 95% para a média de cada variável do <i>set</i> ₉	67
5.7 Matriz de confusão obtida após <i>leave-one-out</i> do modelo de classificação de QDA clássica no <i>set</i> ₉	67
5.8 Estimativas dos coeficientes e razões <i>odds</i> da regressão logística obtida a partir do <i>set</i> ₅ , sem valores omissos.	68
5.9 Matriz de confusão obtida após <i>leave-one-out</i> do modelo de classificação de regressão logística aplicado ao <i>set</i> ₅	68
5.10 Matriz de confusão obtida após validação <i>leave-one-out</i> do modelo de RMLR ($\lambda = 0.2$) aplicado ao <i>set</i> ₉ com padronização, sem valores omissos e sem as observações da Etiópia.	69
5.11 Estimativas dos coeficientes e razões <i>odds</i> do classificador RMLR ($\lambda = 0.2$) obtido a partir do <i>set</i> ₉ padronizado, sem valores omissos e sem as observações da Etiópia.	69
5.12 Medidas de desempenho dos melhores classificadores, obtidas por <i>leave-one-out</i> , treino sem valores omissos e sem as observações da Etiópia.	70
A.1 Correlações amostrais robustas (Fast MCD).	80
A.2 Correlações amostrais robustas (Fast MCD) referentes ao Havai.	80

Lista de figuras

1.1	Flores e cerejas de um cafeeiro.	2
1.2	Principais regiões produtoras de café (<i>The Bean Belt</i>) (in www.coffeeforless.com/articles/coffee-bean-growing-belt.html)	3
1.3	Percentagens de produção total de café na colheita de 2014/15, por país exportador (num total de 56 países).	4
1.4	Importação e consumo de café, em kg/ <i>per capita</i> , em 2013 nos países importadores de café (selecionados pela ICO).	5
2.1	Efeitos isotópicos na precipitação (in Rodrigues, 2011).	12
2.2	Mapa da interpolação da composição isotópica do oxigénio-18 na precipitação (in Agarwal <i>et al.</i> , 2010).	13
2.3	Mapas da origem geográfica, com identificação da altitude, das observações do Brasil, Etiópia, Havai e Quénia.	15
3.1	Padrão de valores omissos.	16
3.2	Diagramas de dispersão e correlações amostrais entre as variáveis.	20
3.3	Diagrama de dispersão das variáveis $\delta^{15}N_{caf}$ e $\delta^{18}O_{caf}$, com as observações coloridas de acordo com a origem geográfica.	20
3.4	Diagramas de dispersão e respetivas correlações amostrais referentes ao Havai.	21
3.5	Diagramas de dispersão e respetivas correlações amostrais referentes ao Brasil.	22
3.6	Diagramas em caixa das características do café.	23
3.7	<i>Scree plot</i> da ACP clássica aplicada ao set_3	27
3.8	<i>Scores</i> das componentes principais estimadas da ACP clássica aplicada ao set_3 (dados padronizados).	29
3.9	<i>Scree plot</i> da ACP PP aplicada ao set_9 (dados padronizados).	30
3.10	<i>Scores</i> das componentes principais estimadas da ACP PP aplicada ao set_9 (dados padronizados).	31
3.11	<i>Scree plot</i> da ACP PP aplicada ao set_{10} (dados padronizados).	32
3.12	<i>Scores</i> das componentes principais estimadas da ACP PP aplicada ao set_{10} (dados padronizados).	33

3.13	Gráficos da distribuição de χ_p^2 em função do quadrado da distância de Mahalanobis robusta, nos subconjuntos set_3 , set_7 , set_9 e set_{10} , com identificação da origem geográfica.	37
3.14	Diagramas de dispersão de algumas variáveis padronizadas do set_3 , com identificação das observações classificadas como <i>outliers</i> .	39
3.15	Gráfico de diagnóstico da detecção de <i>outliers</i> aplicada ao set_3 , com identificação das observações classificadas como <i>outliers</i> .	39
3.16	Diagrama de dispersão das variáveis $\delta^{18}O$ e $\delta^{15}N$ padronizadas do set_7 , com identificação das observações classificadas como <i>outliers</i> .	40
3.17	Gráfico de diagnóstico da detecção de <i>outliers</i> aplicada ao set_7 , com identificação das observações classificadas como <i>outliers</i> .	40
3.18	Diagramas de dispersão das padronizadas selecionadas do set_9 , com identificação das observações classificadas como <i>outliers</i> .	41
3.19	Gráfico de diagnóstico da detecção de <i>outliers</i> aplicada ao set_9 , com identificação das observações classificadas como <i>outliers</i> .	41
3.20	Diagramas de dispersão das variáveis padronizadas selecionadas do set_{10} , com identificação das observações classificadas como <i>outliers</i> .	42
3.21	Gráfico de diagnóstico da detecção de <i>outliers</i> aplicada ao set_{10} , com identificação das observações classificadas como <i>outliers</i> .	42
4.1	R^2 em função do número de grupos e referente à análise de agrupamentos dos <i>scores</i> da ACP do set_3 .	46
4.2	Dendrograma dos agrupamentos obtidos dos <i>scores</i> da ACP do set_3 .	47
4.3	Diagramas de dispersão dos <i>scores</i> da ACP do set_3 , com as observações coloridas de acordo com o grupo obtido da análise de agrupamentos.	48
5.1	Árvore de decisão (C5.0) obtida a partir do set_3 , sem valores omissos.	64
5.2	Árvore de decisão (C5.0) obtida a partir do set_6 , sem valores omissos e com divisão do arquipélago do Havai em duas classes.	65
A.1	Mapas da origem geográfica das observações do Brasil, Quênia, Havai e Etiópia, com identificação dos <i>outliers</i> do set_3 .	81
A.2	Mapas da origem geográfica das observações do Brasil, Quênia, Havai e Etiópia, com identificação dos <i>outliers</i> do set_7 .	82
A.3	Mapas da origem geográfica das observações do Brasil, Quênia, Havai e Etiópia, com identificação dos <i>outliers</i> do set_9 .	83
A.4	Mapas da origem geográfica das observações do Brasil, Quênia, Havai e Etiópia, com identificação dos <i>outliers</i> do set_{10} .	84
A.5	Mapas do Brasil, Quênia, Havai e Etiópia, com identificação do grupo obtido da análise de agrupamentos aos <i>scores</i> da ACP do set_3 .	86

Lista de siglas e símbolos

Siglas

Accy overall accuracy — exatidão global

Prec precision — precisão

ACP análise em componentes principais

CP componente principal

FPR false positive rate — taxa de falsos positivos

ICO International Coffee Organization — Organização Internacional do Café

kNN *k nearest neighbors* — k vizinhos mais próximos

LDA linear discriminant analysis — análise discriminante linear

MCD minimum covariance determinant — determinante de covariância mínima

PP Projection Pursuit

QDA quadratic discriminant analysis — análise discriminante quadrática

RMLR ridge multinomial logistic regression

SMLR sparse multinomial logistic regression

Símbolos

$\delta^{13}C$ composição isotópica do carbono no grão verde de café

$\delta^{13}C_{caf}$ composição isotópica do carbono na molécula de cafeína do grão verde de café

$\delta^{15}N$ composição isotópica do azoto no grão verde de café

$\delta^{15}N_{caf}$ composição isotópica do azoto na molécula de cafeína do grão verde de café

$\delta^{18}O$ composição isotópica do oxigénio no grão verde de café

$\delta^{18}O_{caf}$ composição isotópica do oxigénio na molécula de cafeína do grão verde de café

$\delta^{34}S$ composição isotópica do enxofre no grão verde de café

$^{87}\text{Sr}/^{86}\text{Sr}$ razão isotópica do estrôncio no grão verde de café

$C\%$ percentagem de carbono

$N\%$ percentagem de azoto

Sobrescritos

t transposto

tr truncado

Capítulo 1

Introdução

O café é a personagem principal deste trabalho. Assim, neste capítulo dá-se a conhecer alguns aspetos inerentes ao café, desde a sua descoberta até à atualidade.

1.1 O café

1.1.1 História

A história da forma como o café se espalhou por todo o mundo é a história de um mundo em mudança. Engloba a história da religião, da escravidão, do contrabando, do amor e da comunidade (Moldvaer, 2014, p. 10).

Reza a lenda que foi no ano de 800 d.C que um pastor da Etiópia chamado *Kaldi* se apercebeu que as suas cabras ficavam enérgicas quando comiam bagas vermelhas de um arbusto com flores brancas. Intrigado com o que se sucedia procurou o prior do mosteiro e partilhou a sua descoberta. Com o intuito de confirmar o que o pastor dizia, o monge trouxe algumas bagas para o mosteiro onde as começou a utilizar na forma de infusão, assim não só confirmou que a infusão estimulava os sentidos e despertava a mente como também se apercebeu do seu agradável sabor. Esta bebida foi ganhando popularidade tornando-se cada vez mais procurada pelo seu aroma, sabor e efeito energético (SICAL, 2015).

Da Etiópia o café foi levado para a Arábia, onde a planta do café começou a ser cultivada e deram ao café o nome de *Kahwa* (vinho, em árabe). Até ao século XVII os árabes detinham o monopólio do seu cultivo e comercialização, pois apenas exportavam grãos inférteis de modo a que ninguém pudesse cultivar café noutra local. Apenas em 1616 é que os holandeses conseguiram obter grãos férteis que cultivaram em viveiros (ICO, 2015).

O consumo do café foi introduzido na Europa em 1615 através de mercadores venezianos. Seguindo-se a abertura das primeiras Casas de Café, nomeadamente em Veneza e em Oxford, lugares influentes que se tornaram pontos de convívio e discussão política. Foram os holandeses que disseminaram o café pelo mundo, tendo transformado inicialmente as suas colónias em grandes plantações de café e depois transportado para a América, a par de portugueses e franceses (ClubeCafé, 2015). O café foi levado para Belém no Brasil pelo sargento português Francisco de Mello Palheta em 1727.

No séc. XVIII e XIX a cultura do café desenvolveu-se em várias regiões deste país tornando-se muito importante a nível social e económico. O Brasil é atualmente o principal produtor de café do mundo.

Hoje em dia o café é produzido em mais de 50 países, sendo um fator fundamental para as suas economias, e é a segunda bebida mais consumida em todo o mundo (em primeiro lugar está a água). O café é um dos principais produtos de valor e é negociado nas principais bolsas de futuros e mercadorias do mundo. A Organização Internacional do Café, ICO, refere mesmo que “não é possível exagerar a importância do café na economia mundial” (ICO, 2015).

1.1.2 Do cultivo ao consumo

O cafeeiro é a planta na qual crescem frutos que contêm a semente que é o grão de café verde. Esta planta demora entre 3 a 5 anos a desenvolver-se até aparecerem as primeiras flores e frutos – chamados de cerejas (ver Figura 1.1). Apesar do cafeeiro ser um arbusto que pode ser produtivo durante mais de 80 anos, a sua vida útil a nível económico raramente ultrapassa 30 anos (Rodrigues, 2011, p. 3). Existem seis fatores em que se baseia a escolha do local apropriado para as plantações de café, são eles a temperatura, a disponibilidade de água, a



Figura 1.1: Flores e cerejas de um cafeeiro.

intensidade do sol, o vento, o tipo de solo e a topografia do terreno. De notar que a disponibilidade de água e a temperatura são os dois fatores mais importantes. Mais ainda, a temperatura está associada à altitude e esta e as suas flutuações têm um impacto significativo no comportamento dos cafeeiros (Rodrigues, 2011, p. 8).

No que respeita à classificação científica do cafeeiro, este é um arbusto da família *Rubiaceae*, pertence ao género *Coffea*. São conhecidas diversas espécies de *Coffea*, sendo as duas principais espécies cultivadas à escala mundial a *Coffea arabica* e a *Coffea canephora*, conhecidas como café Arábica e café Robusta, respetivamente. A planta do café Robusta caracteriza-se por ser mais resistente a doenças e o seu grão ter um teor mais elevado em cafeína que o café Arábica. Mais ainda, o café Arábica é, geralmente, mais apreciado e considerado de melhor qualidade que o café Robusta, representado mais de 60% da produção mundial do café (Moldvaer, 2014, p. 19; Keidel *et al.*, 2010).

Devido à ocorrência de diferentes estados de maturação das cerejas em cada ramo de um cafeeiro, a sua colheita é muitas vezes feita à mão de modo a colher apenas as que estão maduras (vermelhas ou amarelas, de acordo com a espécie de cafeeiro), quando as cerejas apresentam um teor de açúcares mais elevado. Cada cereja do cafeeiro contém duas sementes, que, para se tornarem grãos verdes de café, devem ser processadas logo após a colheita, de forma a preservar a sua qualidade. Os dois principais métodos utilizados no processamento são a via húmida e a via seca (também referida como método “natural”). A via húmida, apesar de exigir maiores investimentos e cuidados, ajuda a preservar as qualidades intrínsecas dos grãos de cafés e origina menos desperdícios, sendo este o método mais

comum (Delta Cafés, 2015).

Após o processamento, os grãos verdes são separados de acordo com o seu tamanho, densidade ou cor de forma a separar os de melhor qualidade dos restantes, sendo depois enviados para o mercado da especialidade. Os grãos de alta qualidade são enviados em sacos especiais (trata-se do *specialty coffee*) e os restantes são enviados em contentores para serem transportados e comercializados em todo o mundo (Moldvaer, 2014, p. 23). A fase seguinte é a torrefação; é durante esta fase que os grãos verdes revelam o seu perfil aromático e sabor, sendo um indicador fundamental da qualidade do café. Defeitos que geralmente não são observados no grão verde podem ser salientados no café torrado (Delta Cafés, 2015). Após a torra, os grãos estão prontos a dar origem a lotes de acordo com as suas características para serem moídos e consumidos em variadas formas, sendo a bebida a que chamamos café a principal forma de consumo em todo o mundo.

1.1.3 O café no mundo

O café é o único produto alimentar produzido em todo o mundo (à exceção da Europa) associado a vários projetos de sustentabilidade, comércio justo, projetos de cuidados de saúde e de educação implementados nas regiões onde é cultivado (Rodrigues *et al.*, 2013).

As condições ideais para o desenvolvimento dos cafeeiros são encontradas na zona Equatorial, sendo chamada *The Bean Belt*. Esta zona de produção de grão verde de café e as principais regiões produtoras de café estão representadas na Figura 1.2.



Figura 1.2: Principais regiões produtoras de café (*The Bean Belt*) (in www.coffeeforless.com/articles/coffee-bean-growing-belt.html)

Na Figura 1.3 apresentam-se as percentagens de produção total de café da colheita de 2014/15, por país exportador, num total de 56 países e 141 732 milhares de sacos de 60kg de café. Este gráfico foi obtido a partir dos dados da ICO, disponíveis em www.ico.org/new_historical.asp.

Nos últimos 100 anos, o Brasil tem sido o maior produtor de café do mundo, registando 32% da produção total da colheita de 2014/15. O segundo maior produtor de café é o Vietname, onde a produção de café tem crescido rapidamente, representando 19% da produção de 2014/15, sendo a maior parte do café produzido neste país (e nos restantes países do continente asiático) do tipo Robusta

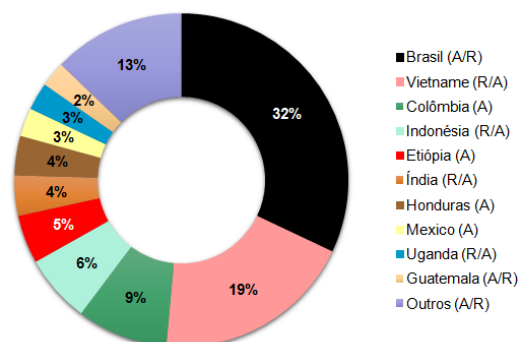


Figura 1.3: Percentagens de produção total de café na colheita de 2014/15, por país exportador (num total de 56 países). Onde as letras A e B significam que os dados são referentes à produção de café do tipo Arábica e Robusta, respetivamente.

(nca, 2015). A Colômbia, um dos países mais conhecidos pela produção de café, ocupa o terceiro lugar da lista dos maiores produtores de café, com 9% da produção de 2014/15, sendo o segundo maior produtor de café do tipo Arábica.

Durante o período de desenvolvimento do grão de café, vários são os fatores que condicionam a sua qualidade como produto final. Desde a variedade do cafeeiro, o tipo de solo, o clima, a quantidade de chuva e de sol, à altitude a que a plantação se encontra. Estas variáveis, combinadas com a forma como as cerejas que contêm o grão são processadas, contribuem para a distinção entre os vários países, regiões produtoras e até mesmo diferentes plantações (nca, 2015). Por exemplo, o café do arquipélago do Havai é considerado um dos melhores do mundo, sendo o café com a denominação *Kona Coffee*, cultivado no distrito de Kona, um dos mais caros do mundo. O café cultivado na África Oriental é um dos mais distintos do mundo, que inclui origens como a Etiópia (nomeadamente da região de Yirgacheffe), o Quênia e a Tanzânia, caracterizadas pelas plantações a elevadas altitudes. Outro dos cafés reconhecido pela sua qualidade é o denominado *Jamaica Blue Mountain*, que cresce a elevadas altitudes nas extensas montanhas da Jamaica.

Segundo dados da Organização Internacional de Café, em relação ao consumo e importação de café, a União Europeia distinguem-se por ser a maior consumidora e importadora de café no mundo, seguindo-se os Estados Unidos da América.

Na Figura 1.4 representa-se a importação e consumo de café em kg/*per capita* de alguns países importadores de café em 2013, estes países correspondem a uma seleção feita pela ICO. Este gráfico foi construído recorrendo aos dados de importação e consumo fornecidos pela ICO em www.ico.org/pt/new_historical_p.asp e aos dados da população por país fornecidos pelo *World Bank Group* em data.worldbank.org/indicator/SP.POP.TOTL.

O país com maior consumo de café, *per capita*, é a Finlândia, onde o café tem um lugar especial na cultura e nas tradições. Entre os países da Figura 1.4, Portugal é o 16.^o maior consumidor de café. A Alemanha, a Bélgica e a Suíça são os principais reexportadores de café. Enquanto que a Alemanha e a Bélgica (graças ao porto de Antuérpia, Crem B.V. e Pierrot, 2015) exportam em maioria grãos verdes de café, no caso da Suíça a maior percentagem vai para o café torrado e solúvel (ICO, 2012). A elevada exportação de café pela Suíça deve-se principalmente à expansão da venda de cápsulas de café dos últimos anos, produzidas nas fábricas da Nestlé sediadas neste país.

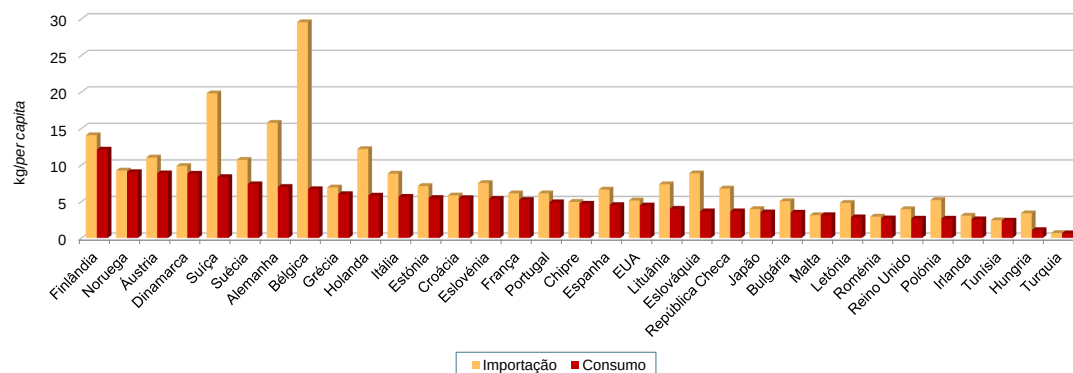


Figura 1.4: Importação e consumo de café, em kg/per capita, em 2013 nos países importadores de café (selecionados pela ICO).

Em Portugal, nos últimos anos o consumo de café teve um aumento significativo, passando de cerca de 4.2kg/per capita em 2009 para 4.9kg/per capita em 2013. Este aumento deve-se principalmente ao consumo de café em casa, a partir das cápsulas de café.

O mercado do café em Portugal é liderado pela Delta Cafés, esta marca possui a maior fábrica de torrefação de café da península ibérica, sediada em Campo Maior, e foi a primeira empresa do ramo certificada em Portugal (pelo sistema de normas NP 29002) (Delta Cafés, 2015). A Delta é uma empresa com mais de 50 anos cujo nome já faz parte da história do café em Portugal. A Delta Cafés SGPS é uma *sub-holding* do Grupo Nabeirogest SGPS SA constituída por 11 empresas que se dedicam à importação, torrefação e comercialização de café. Duas das políticas desta empresa são a inovação e a sustentabilidade, investindo por exemplo na investigação e no comércio sustentável. A Delta importa grão verde de café de várias origens, sendo o maior volume do Brasil, Vietname, Uganda, Índia e Camarões (Delta Cafés, 2011).

1.2 Principais objetivos e conteúdos da dissertação

O clima, topográfica e práticas agrícolas das diversas regiões produtoras de café conferem características singulares que se traduzem na qualidade e no preço do café. A comercialização e globalização do mercado levam a que as empresas importadoras de café necessitem de formas de validar e confirmar a origem do grão importado (Rodrigues *et al.*, 2013). Os produtores de café são profissionais de valor inestimável nas empresas que comercializam café, já que lhes cabe a tarefa de selecionar os melhores lotes e testar a sua qualidade. Mas estes profissionais têm uma missão complexa que exige muitos anos de formação e treino. Por esta razão a indústria do café tem vindo a dedicar atenção aos métodos automáticos de classificar os grãos de café de acordo com a origem geográfica, tendo por base características químicas e físicas do grão de café. Este trabalho propõem-se a dar um contributo neste sentido, sendo o seu principal objetivo a construção de um classificador de grão verde de café de acordo com a origem geográfica.

As amostras de grão verde de café são caracterizadas por várias composições isotópicas, uma vez que os peritos desta área defendem que os isótopos estáveis indicam a presença e magnitude de processos ecológicos essenciais. Muitos destes processos produzem *assinaturas* isotópicas distintas

possibilitando a diferenciação da origem geográfica da planta (Rodrigues, 2011, p. 17), sendo a análise da composição isotópica uma das abordagens mais promissoras para determinar a origem geográfica de alimentos (Kelly *et al.*, 2005).

Este problema é discutido no Capítulo 2, nomeadamente na Secção 2.1. As composições isotópicas em estudo são introduzidas na Secção 2.2. Importa notar que o estudo das composições isotópicas, nomeadamente no composto específico do café, a cafeína, mostrou-se promissor na diferenciação da origem geográfica do café no trabalho pioneiro de Weckerle *et al.* (2002).

Outras das características deste trabalho são a elevada dimensão da amostra, em comparação com os estudos realizados com as mesmas características (cuja revisão é feita na Secção 2.1.1), e a aplicação de várias técnicas estatísticas para estudar o problema da diferenciação da origem geográfica de grãos verdes de café. O conjunto de dados original é brevemente descrito na Secção 2.3.

No Capítulo 3 é feita uma análise exploratória dos dados considerados para análise, investigando-se as relações existentes entre as variáveis e objetos recorrendo a diversas metodologias estatísticas. Uma vez que os dados em análise são reais, é natural esperar que estes não verifiquem escrupulosamente as hipóteses estatísticas subjacentes aos métodos utilizados, o que justifica o uso de estimadores robustos. Em particular, a Secção 3.3 deste trabalho é dedicada à análise em componentes principais e a Secção 3.4 à deteção de *outliers* utilizando técnicas robustas.

No Capítulo 4 é realizada a análise de agrupamentos, de forma a agrupar o conjunto de dados com base nas relações existentes entre as observações, sem ter em conta a origem geográfica.

No Capítulo 5 procura-se encontrar um classificador das diferentes origens geográficas de grãos verdes de café consideradas. Esta procura é feita recorrendo a várias metodologias de classificação, sendo a regressão logística multinomial que incorpora seleção de variáveis a menos conhecida. Este capítulo termina com a escolha do classificador para o problema em estudo.

São apresentadas no Capítulo 6 as principais conclusões deste estudo realizado à composição isotópica do grão verde de café com vista à diferenciação geográfica.

As diversas análises estatísticas ao conjunto de dados foram efetuadas recorrendo às potencialidades do *software R* (R Core Team, 2015).

Capítulo 2

Discussão do problema real

O principal objetivo deste capítulo é dar a conhecer o problema em estudo, designadamente a diferenciação da origem geográfica do grão verde de café, bem como as características dos dados utilizados para a exploração deste problema.

2.1 Classificação de grãos de café de acordo com a origem geográfica

A qualidade do café começa na planta do café, onde se desenvolvem as cerejas que dão origem ao grão verde de café. Se a qualidade não estiver presente no grão verde, o aroma do café torrado vai ser afetado, não fornecendo ao consumidor a melhor experiência na degustação (Folmer, 2014). A exigência por parte dos consumidores tem aumentado significativamente, não só devido à qualidade exigida e diversidade de preferências, como também ao interesse manifestado pelos métodos de produção dos bens de consumo.

Graças ao clima, topografia e práticas agrícolas distintas (como por exemplo a agricultura biológica), as diversas regiões produtoras de café produzem café com características singulares apreciadas pelos milhões de consumidores em todo o mundo. Estes fatores levam a que o preço e a qualidade do café dependam substancialmente da sua origem geográfica.

Assim, as empresas importadoras de café interessam-se cada vez mais em desenvolver ferramentas analíticas, com base em análises químicas/físicas, para confirmar que o café importado não sofreu adulteração ao longo da cadeia comercial (Rodrigues *et al.*, 2013).

2.1.1 Revisão da literatura

Nas últimas décadas o número de estudos em torno do café tem aumentado significativamente. Ainda assim, o estudo da diferenciação geográfica do grão verde, em particular, é muito recente. Até agora a diferenciação geográfica dos países/regiões produtores de café ainda não foi passível de ser efetuada de forma sistemática por diversos motivos, como os custos inerentes às técnicas que se têm mostrado

mais promissoras (espectrometria de massa, MS) e consequentemente a reduzida quantidade de amostras tidas em conta. Destacam-se apenas diferenciações em macro-regiões (a nível continental) (Liu *et al.*, 2013; Santato *et al.*, 2012; Rodrigues *et al.*, 2011b; Serra *et al.*, 2005; Weckerle *et al.*, 2002) e micro-regiões, como regiões do Havai (McLeod *et al.*, 2013; Rodrigues *et al.*, 2011a).

A análise da composição isotópica do grão verde de café é uma ferramenta promissora para diferenciação da origem geográfica, na medida em que a *assinatura* isotópica deixada no grão deve ser o resultado da variedade da planta, práticas agrícolas e ainda da relação entre a planta e o ambiente onde esta se desenvolve. Desta forma, esperam-se variações na composição isotópica dos grãos de café originários de diferentes origens geográficas com o seu próprio clima e geologia (Rodrigues *et al.*, 2013)

Em relação às composições isotópicas que se têm mostrado mais promissoras, com vista à diferenciação geográfica de grãos verdes de café, salienta-se a composição isotópica do oxigénio (Santato *et al.*, 2012; Rodrigues *et al.*, 2011b), em especial na cafeína presente no grão (Rodrigues *et al.*, 2011a; Weckerle *et al.*, 2002). Outras das composições isotópicas consideradas relevantes para o problema são a composição isotópica do boro, do azoto, do carbono e do hidrogénio (Liu *et al.*, 2013; McLeod *et al.*, 2013; Santato *et al.*, 2012; Serra *et al.*, 2005; Weckerle *et al.*, 2002). A razão isotópica do estrôncio é também apontada como promissora (Liu *et al.*, 2013; Rodrigues *et al.*, 2011a; Rodrigues *et al.*, 2011b), especialmente na diferenciação entre micro-regiões (Kelly *et al.*, 2005).

As técnicas estatísticas utilizadas na maioria destes estudos são a análise em componentes principais e a análise de estatísticas descritivas como a média e o desvio padrão.

A principal limitação dos estudos sobre diferenciação da origem geográfica do grão verde de café, através da análise da composição isotópica, é o reduzido número de amostras consideradas, o que dificulta a análise estatística. Por exemplo, em Liu *et al.* (2013) apenas se consideram 21 amostras de 14 diferentes origens geográficas, que são analisadas através de estatísticas descritivas e da análise em componentes principais; em Weckerle *et al.* (2002) consideram-se 45 amostras e recorre-se a análise discriminante linear e a árvores de decisão para classificar as amostras nas diferentes origens geográficas.

Em Rodrigues (2011, p. 14–16) é feita uma revisão das técnicas analíticas associadas aos estudos sobre diferenciação da origem geográfica de grãos de café, que pelo seu carácter técnico não se encaixa no âmbito deste trabalho. Para informação ainda mais detalhada sobre técnicas analíticas em estudos de diferenciação da origem geográfica, recomenda-se a leitura de Luykx e Ruth (2008). Sugere-se ainda Kelly *et al.* (2005), para uma revisão da literatura abordando vários aspetos da diferenciação geográfica de produtos alimentares.

O presente trabalho tem por ponto de partida os estudos e respetiva recolha de dados apresentados em Rodrigues *et al.* (2011a), Rodrigues *et al.* (2011b) e Rodrigues *et al.* (2009). Estes estudos distinguem-se da maioria dos restantes, com as mesmas características, pelo número de amostras consideradas, nomeadamente 68 de 20 origens geográficas diferentes, em Rodrigues *et al.* (2011b) e em Rodrigues *et al.* (2009), e 47 das várias regiões produtoras de café do Havai em Rodrigues *et al.* (2011a).

Com o intuito de compreender melhor as composições isotópicas e o seu papel na diferenciação geográfica, na próxima secção apresenta-se uma breve introdução sobre este assunto.

2.2 Características do grão de café: composição isotópica

Com o objetivo de desenvolver métodos que permitam a diferenciação geográfica de alimentos, como o azeite, o vinho e o café, tem sido aplicada análise de composições isotópicas e análise multi-elemento em várias áreas (Kelly *et al.*, 2005). Para além desta aplicação, a deteção de adulterações e a identificação do método de produção de produtos alimentares (cultivo biológico vs. sistema de cultivo convencional) são outras das principais aplicações da determinação da composição isotópica (Inácio *et al.*, 2015).

Os isótopos são átomos de um elemento químico com o mesmo número atómico (mesmo número de prótons), mas com diferente número de massa,¹ sendo o número de neutrões o que difere os diferentes isótopos de um elemento químico. Neste trabalho os elementos químicos abordados são o carbono, azoto, enxofre, oxigénio e estrôncio (*C*, *N*, *S*, *O* e *Sr*, respetivamente).

A razão isotópica de um elemento é a razão entre dois dos seus isótopos estáveis, usualmente um dos isótopos considerados é muito raro e o outro muito comum (geralmente o mais pesado é o mais raro e o mais leve o mais comum na natureza). Os isótopos integram processos ecológicos importantes no espaço e no tempo, deixando a sua *assinatura* isotópica e dando assim a indicação da presença e magnitude destes processos. A presença ou ausência de processos ecológicos é indicada através dos valores de razões isotópicas, que em certos casos podem ser utilizados como um registo da resposta dos sistemas às alterações das condições ambientais ou como um indicador para alterações climáticas (Rodrigues *et al.*, 2011a; Rodrigues *et al.*, 2009).

Serão abordadas neste trabalho as razões isotópicas dos seguintes elementos (com as respetivas abundâncias relativas na natureza de cada isótopo) (Dawson e Siegwolf, 2007, p. 6):

- carbono: $^{13}\text{C}/^{12}\text{C}$ (1.108‰ e 98.982‰);
- azoto: $^{15}\text{N}/^{14}\text{N}$ (0.366‰ e 99.630‰);
- oxigénio: $^{18}\text{O}/^{16}\text{O}$ (0.1995‰ e 99.763‰);
- enxofre: $^{34}\text{S}/^{32}\text{S}$ (4.210‰ e 95.02‰);
- estrôncio: $^{87}\text{Sr}/^{86}\text{Sr}$ (7.020‰ e 9.860‰).

Devido à abundância muito reduzida dos isótopos raros, a análise de razões isotópicas é realizada através da comparação entre a razão isotópica no material em estudo e a razão isotópica de referência do elemento (padrão internacional). Esta comparação é realizada recorrendo à notação *delta* (δ), onde a razão isotópica da amostra em análise (SA) é expressa relativamente a um padrão aceite internacionalmente (STD):

$$\delta^{XX}E = \left(\frac{R_{SA}}{R_{STD}} - 1 \right) \times 1000, \quad (2.1)$$

¹O número de massa de um átomo corresponde ao número de prótons mais o número de neutrões do núcleo desse átomo.

onde E é o elemento de interesse (C , N , O ou S), XX é a massa atômica do isótopo mais pesado da razão isotópica e R é a razão absoluta do elemento (e.g., $^{13}C/^{12}C$). Como a abundância dos isótopos é muito reduzida o valor de δ é multiplicado por 1000, por isso o seu valor é expresso em permilagem (‰). Por definição, o padrão aceite (STD) tem valor δ de 0‰, logo uma substância com δ positivo tem uma razão entre o isótopo pesado e o leve, R_{SA} , maior que o padrão, R_{STD} , e *vice-versa*. Substâncias com valor δ positivo (negativo) dizem-se frequentemente mais “pesadas” (“leves”) ou “enriquecidas” (“empobrecidas”) em isótopo mais pesado relativamente ao padrão (Rodrigues, 2011, pp. 18–19).

De notar que no caso do estrôncio é prática comum não considerar a razão isotópica entre o isótopo estável mais pesado e o mais leve, mas sim a razão entre as abundâncias dos isótopos que são medidos, $^{87}Sr/^{86}Sr$, não sendo utilizada a notação δ . Pois é possível obter, pela técnica analítica de alta precisão ICP-MS, valores de razão entre os dois isótopos do estrôncio (Dawson e Siegwolf, 2007, p. 6).

Ocorrem na natureza diferentes abundâncias entre isótopos leves e pesados porque os átomos mais pesados têm menor frequência de vibração que os mais leves. Assim, os átomos ou moléculas mais pesados reagem mais lentamente que os mais leves, pois formam ligações químicas mais fortes, sendo necessária mais energia para as quebrar. Isto resulta numa distribuição desigual dos isótopos em materiais orgânicos e em materiais inorgânicos pelo processo a que se dá o nome de fracionamento isotópico. As alterações da composição isotópica de um dado elemento químico podem ocorrer como resultado de um único processo (por exemplo, a evapotranspiração afeta o valor de $^{18}O/^{16}O$ da água nos tecidos das plantas) ou, por vezes, como o resultado de um conjunto de processos (por exemplo, biossíntese da celulose) (*vide* Rodrigues, 2011, p. 19). Muitos fenómenos naturais, classificados como físico-químicos, podem conduzir ao fracionamento isotópico. A Tabela 2.1 fornece um resumo dos principais processos ou fatores que afetam as razões isotópicas de determinados elementos químicos e como estas podem ser utilizadas para determinar a origem de certos materiais de origem vegetal ou animal.

Tabela 2.1: Resumo da forma como as razões isotópicas são afetadas (fracionadas) no ambiente e como podem ser exploradas para determinação da proveniência de alimentos (adaptado de Kelly *et al.*, 2005).

Razão isotópica	Fracionamento	Informação
$^{13}C/^{12}C$	Plantas com metabolismo fotossintético em C3 e C4	Dieta (indicador geográfico)
$^{15}N/^{14}N$	Nível trófico, práticas agrícolas	Dieta (indicador geográfico)
$^{18}O/^{16}O$	Evapotranspiração, condensação, precipitação	Geografia
$^{34}S/^{32}S$	Bacteriano	Geografia (marinho)
$^{87}Sr/^{86}Sr$	Idade da rocha e razão Rb/Sr	Geologia subjacente, geografia

Numa primeira aproximação, as razões isotópicas do carbono e do azoto fornecem informação acerca do tipo de metabolismo da planta ou da dieta de um animal, enquanto as razões isotópicas do oxigénio, do enxofre e do estrôncio fornecem informação acerca da origem geográfica, por estarem relacionadas com a geologia do local onde se desenvolve o material a estudar (Kelly *et al.*, 2005). Por exemplo, a razão isotópica do carbono é um indicador do tipo de metabolismo fotossintético das plantas, assim como através do estudo da composição isotópica de um determinado tecido animal, podemos obter informação acerca da sua dieta.

Para o cálculo dos valores δ das razões isotópicas, cada amostra foi finamente moída em laboratório. De cada amostra fazem parte um determinado número de grãos de café. Detalhes sobre a preparação e processos envolvidos na medição destes valores podem ser encontrados em Rodrigues (2011). Entre as técnicas envolvidas no processo estão as técnicas de espectrometria de massa de razões isotópicas (IRMS), aplicada para determinação da composição isotópica do oxigénio, carbono, azoto e enxofre, e a técnica de espectrometria de massa acoplada a plasma indutivo (ICP-MS), utilizada na medição de $^{87}\text{Sr}/^{86}\text{Sr}$.

Um caso particular da análise da composição isotópica é a análise da composição isotópica de compostos específicos, tais como a cafeína do grão verde de café. Comparativamente à análise da composição isotópica do grão verde de café, a análise da composição isotópica da molécula de cafeína do grão verde apresenta a vantagem de permitir uma interpretação da composição isotópica da cafeína menos complexa no caso do elemento O, que na molécula da cafeína tem origem na água utilizada pela planta (Rodrigues, 2011, pp. 108–109). No caso do grão como um todo estão associados mais parâmetros de *input*, sendo a composição isotópica resultante de vários processos que envolvem fracionamento bioquímico associados a diferentes vias metabólicas (*vide* Rodrigues, 2011, pp. 108–109).

2.2.1 Composição isotópica do carbono

O valor de $\delta^{13}\text{C}$ nas plantas é sobretudo função do fracionamento isotópico do dióxido de carbono (CO_2) na fotossíntese (Inácio *et al.*, 2015), existindo grandes diferenças quando se comparam plantas terrestres e marinhas, bem como entre as plantas terrestres com diferentes metabolismos fotossintéticos (por exemplo, C3, C4 e CAM). O valor de $\delta^{13}\text{C}$ das plantas com metabolismo em C3, como é o caso do cafeeiro, pode variar de -20 a -35‰ em resposta à humidade relativa do ar, à quantidade de radiação e à disponibilidade de nutrientes (Rodrigues, 2011, p. 21). Os valores de $\delta^{13}\text{C}$ das plantas com metabolismo fotossintético em C4 podem variar aproximadamente entre -10 a -19‰ (Dawson e Siegwolf, 2007, p. 8). A composição isotópica do CO_2 atmosférico (local) tem também uma forte influência nos valores de $\delta^{13}\text{C}$ das plantas (Inácio *et al.*, 2015).

2.2.2 Composição isotópica do azoto

Os principais fatores que afetam o valor de $\delta^{15}\text{N}$ das plantas são as práticas agrícolas, tais como o uso de fertilizantes de azoto, N , e a fixação biológica de azoto, N_2 , os órgãos da planta e a sua idade (Inácio *et al.*, 2015).

Para Inácio *et al.* (2015), $\delta^{15}\text{N}$ parece ser o parâmetro mais promissor, entre os restantes considerados neste estudo, para diferenciar entre matéria vegetal proveniente de cultivo biológico e de cultivo convencional, podendo as plantas provenientes de cultivo orgânico atingir valores mais elevados em $\delta^{15}\text{N}$ (Inácio *et al.*, 2015). Amundson *et al.* (2003) encontraram ainda relação entre o clima e os valores de $\delta^{15}\text{N}$, ao verificar que os valores de $\delta^{15}\text{N}$ no solo e nas plantas decrescem com o aumento da precipitação média anual e o decréscimo da temperatura média anual.

2.2.3 Composição isotópica do oxigênio e do hidrogênio

As composições isotópicas do oxigênio e do hidrogênio são aplicáveis à caracterização da origem geográfica, podendo variar com a latitude, observando-se o denominado efeito isotópico de altitude (Kelly *et al.*, 2005).

Não há fracionamento isotópico durante a captação de água pelas raízes das plantas, por este motivo a composição isotópica da água em raízes e caules reflete a composição isotópica da água disponibilizada à planta (Rodrigues *et al.*, 2009). As duas principais fontes de água das plantas são os reservatórios de água subterrânea e a precipitação. A composição isotópica da água de reservatórios representa uma média de longo prazo da precipitação na área. Apesar de não existir fracionamento isotópico durante a captação de água, este ocorre durante a transpiração das plantas o que leva ao enriquecimento dos valores de $\delta^{18}O$ da água das folhas e por consequência do material vegetal (Rodrigues, 2011, p. 25).

Entre os efeitos mais importantes que influenciam os valores de δ^2H e de $\delta^{18}O$ das águas pluviais estão a sazonalidade, a altitude — ambas com impacto na temperatura de condensação da precipitação — e a latitude. Outro efeito conhecido que influencia a composição isotópica de águas pluviais é a continentalidade, ou o efeito isotópico originado pela distância da massa de ar que contém água que viajou por terra a partir das zonas costeiras, onde a maioria das tempestades têm a sua origem, o que faz com que a água do interior de um continente tenha valores mais negativos de δ^2H e $\delta^{18}O$. Além disso, as chuvas intensas influenciam os valores destes parâmetros na precipitação, tornando-os mais negativos (Rodrigues, 2011, p. 24–25). As figuras 2.1 e 2.2 ilustram estes efeitos.

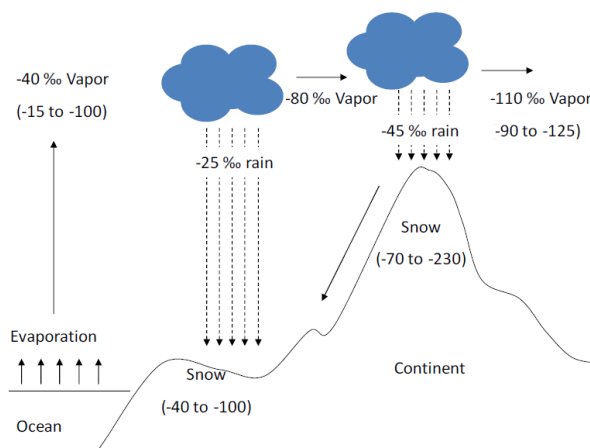


Figura 2.1: Efeitos isotópicos na precipitação (*in* Rodrigues, 2011).

2.2.4 Composição isotópica do enxofre

Os valores de $\delta^{34}S$ das plantas podem ser influenciados pelo teor de compostos de enxofre dos solos, em fertilizantes, através de absorção foliar de poluentes atmosféricos como SO_2 ou pela proximidade ao oceano (Inácio *et al.*, 2015). Organismos que se desenvolvem em regiões de elevada atividade antropogénica são geralmente mais enriquecidos no seu valor de $\delta^{34}S$ (Rodrigues, 2011, p. 28).

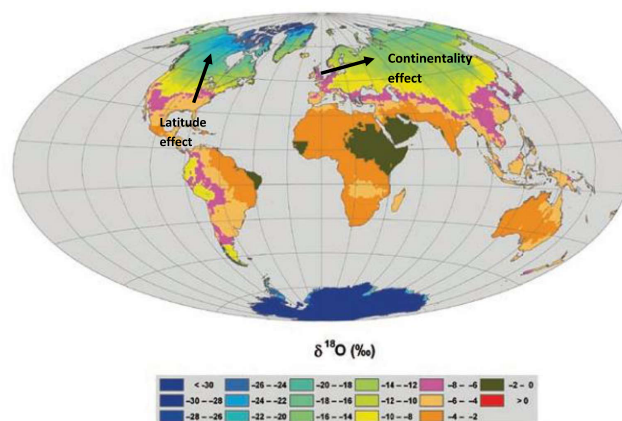


Figura 2.2: Mapa da interpolação da composição isotópica do oxigênio-18 na precipitação (in Aggarwal *et al.*, 2010).

2.2.5 Composição isotópica do estrôncio

A razão isotópica do estrôncio não é, geralmente, expressa na notação δ mas sim como a razão entre os isótopos ^{87}Sr e ^{86}Sr numa amostra.

A composição isotópica do Sr depende da idade da rocha e da razão Rb/Sr (especificamente $^{87}\text{Rb}/^{86}\text{Sr}$)², o isótopo ^{87}Sr é formado através do decaimento radioativo do rubídio-87, ^{87}Rb (Kelly *et al.*, 2005; Rodrigues, 2011, p. 29). Ao contrário da crosta continental, a crosta oceânica é jovem, tendo $^{87}\text{Sr}/^{86}\text{Sr}$ mais reduzido. Nas plantas, a captação de Sr móvel ocorre através da água e este elemento químico representa, portanto, uma fonte de informação isotópica. Os valores de $^{87}\text{Sr}/^{86}\text{Sr}$ refletem o ambiente local e assim esta razão isotópica constitui uma ferramenta promissora em estudos sobre a autenticação de alimentos (Rodrigues, 2011, p. 29).

2.3 Descrição dos dados

Com o objetivo de estudar o problema da diferenciação da origem geográfica de grãos verdes de café foi fornecido, pela Novadelta, Comércio e Indústria de Cafés, S.A. (Campo Maior, Portugal), o conjunto de dados original com 296 amostras de grãos verdes de café com indicação da origem geográfica (Rodrigues *et al.*, 2013; Rodrigues *et al.*, 2011a; Rodrigues *et al.*, 2011b; Rodrigues *et al.*, 2009). O conjunto de dados contempla 34 origens geográficas diferentes dos 4 continentes produtores de café (América, África, Ásia e Oceânia), sendo a maioria do Brasil e do Havai (aproximadamente 19% e 18% das observações, respetivamente), seguidamente da Colômbia, Etiópia, Quênia e Uganda (aproximadamente 5% das observações para cada país). Além da origem geográfica com identificação do país e região produtora, tem-se ainda a informação da espécie (Robusta ou Arábica), o ano de amostragem, coordenadas geográficas (latitude e longitude) e a altitude (obtida a partir das coordenadas geográficas), sendo que estas informações não existem para todas as 296 observações.

Dispõe-se ainda de 10 outras variáveis, oito referentes à composição isotópica do grão verde de café e duas a percentagens de um elemento químico no grão. Entre as composições isotópicas, três destas variáveis referem-se à composição isotópica medida num composto específico do grão, a cafeína. Estas

²O símbolo Rb representa o símbolo químico do rubídio.

10 variáveis são:

- a composição isotópica do carbono, $\delta^{13}C$ (no grão) e $\delta^{13}C_{caf}$ (no composto específico);
- a composição isotópica do azoto, $\delta^{15}N$ (no grão) e $\delta^{15}N_{caf}$ (no composto específico);
- a composição isotópica do oxigénio, $\delta^{18}O$ (no grão) e $\delta^{18}O_{caf}$ (no composto específico);
- a composição isotópica do enxofre, $\delta^{34}S$;
- a razão isotópica do estrôncio, $^{87}Sr/^{86}Sr$;
- a percentagem de carbono, $C\%$, e a de azoto, $N\%$.

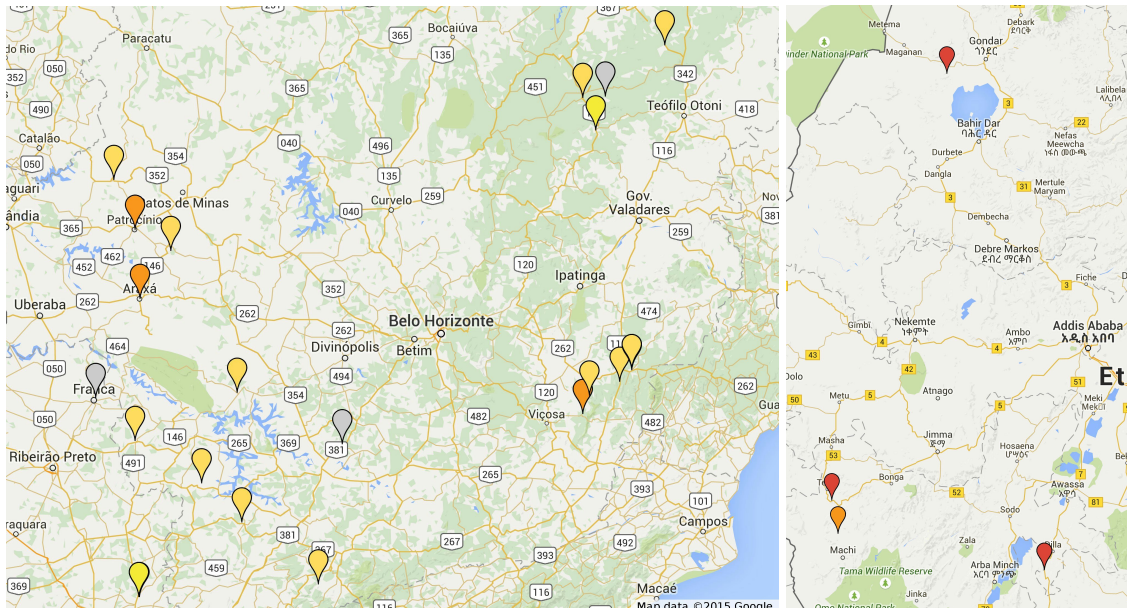
Além destas 10 variáveis dispõe-se ainda dos valores da composição isotópica do oxigénio da precipitação do local de amostragem, obtidos através da calculadora *online* de isótopos na precipitação (OIPC) (Bowen e Revenaugh, 2003).

De um modo geral, a obtenção de valores da composição isotópica envolve processos analíticos e químicos complexos, demorados e dispendiosos. Estas razões justificam o número elevado de valores omissos na amostra, sendo exceção apenas $\delta^{13}C$.

2.3.1 Regiões produtoras de café: Brasil, Etiópia, Havai e Quénia

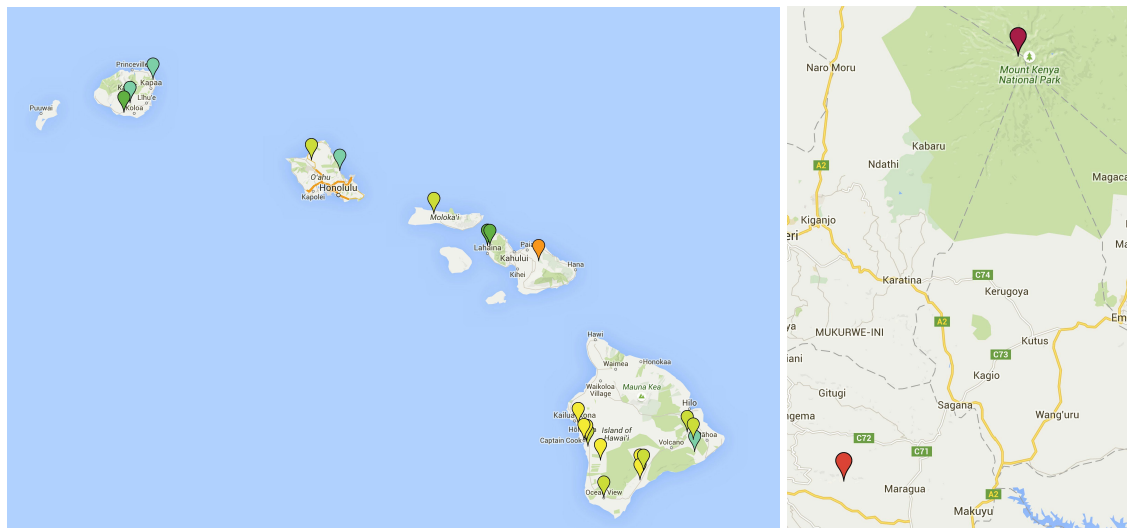
Devido ao elevado número de diferentes regiões e reduzido número de observações em cada uma delas, houve a necessidade de restringir a análise apenas a algumas das regiões. As regiões selecionadas para a análise foram as regiões produtoras do Havai, Brasil, Etiópia e Quénia. Nesta escolha foram tidos em conta não só o número de observações existentes para cada origem, como também a relevância das regiões produtores de café (já referida em 1.1.3). O subconjunto de dados obtido da seleção destas regiões é constituído por 138 observações nas 10 variáveis e informação geográfica descritas no início desta secção. Entre as observações em estudo, 57 são do Brasil, 52 do Havai, 15 da Etiópia e 14 do Quénia, sendo todas referentes a café do tipo Arábica.

As regiões de onde foram recolhidas as amostras estão assinaladas nos mapas da Figura 2.3, com as observações coloridas de acordo com o intervalo de altitude correspondente, em metros (existem observações das quais não se dispõe de coordenadas geográficas e altitude, e algumas, em virtude das coordenadas serem as mesmas, têm os símbolos sobrepostos). As observações do Brasil (Figura 2.3a) são oriundas do estado de Minas Gerais (maioritariamente do Sul) e do Norte de São Paulo, as observações de que não se dispõe de informação acerca das coordenadas geográficas são também destas regiões. Em relação à altitude a maioria está entre 800 e 1500 metros. As observações da Etiópia (Figura 2.3b), que contêm um elevado número de valores omissos, são provenientes de 3 regiões diferentes a elevadas altitudes. No que diz respeito ao Havai (Figura 2.3c), existem observações de todas as regiões produtoras, sendo o conjunto de observações do Havai o mais completo. Verifica-se que, entre todas as regiões em estudo, as regiões de menor altitude são as do Havai, sendo que na ilha Havai as altitudes são, em geral, mais elevadas que nas restantes ilhas. Por fim, as observações do Quénia (Figura 2.3d) são provenientes, na maioria, do Monte Quénia (zona de muito elevada altitude) e apenas duas da zona de Kîrîmîri.



(a) Brasil

(b) Etiópia



(c) Havai

(d) Quênia



Figura 2.3: Mapas com identificação do intervalo de altitude (m) das observações (com altitude e coordenadas geográficas disponíveis) do Brasil, Etiópia, Havai e Quênia, respetivamente. (Map data ©2015 Google, INEGI)

De forma a estudar as características dos dados em estudo, nomeadamente relações existentes entre as variáveis e origens geográficas, no próximo capítulo apresenta-se uma análise exploratória dos mesmos.

Capítulo 3

Análise preliminar dos dados

3.1 O problema dos valores omissos

Uma das principais limitações dos dados em estudo é o elevado número de valores omissos, como se pode verificar na Tabela 3.1, onde se encontram as proporções de valores omissos em cada variável. Como se refere na Secção 2.3, a medição das composições isotópicas do grão verde de café é um processo complexo a vários níveis, o que dificulta a recolha de amostras.

Com o objetivo de ilustrar os valores omissos na amostra em estudo, representa-se na Figura 3.1 uma matriz em que cada linha representa uma variável e cada coluna uma observação. As células a azul representam as medições efetivamente realizadas, enquanto que as células a vermelho representam os valores omissos. Consta-se que as únicas variáveis sem qualquer valor omissos são $\delta^{15}N$ e $\delta^{13}C$, sendo a frequência de valores omissos muito elevada (superior a 24% em metade das variáveis). Mais ainda, em geral o mesmo objeto apresenta valor omissos em várias variáveis.

Tabela 3.1: Proporção de valores omissos em cada variável.

	$C\%$	$N\%$	$\delta^{13}C$	$\delta^{18}O$	$\delta^{15}N$	$\delta^{34}S$	$^{87}Sr/^{86}Sr$	$\delta^{13}C_{caf}$	$\delta^{18}O_{caf}$	$\delta^{15}N_{caf}$
Frequência	0.17	0.17	0	0.14	0	0.35	0.33	0.24	0.24	0.34

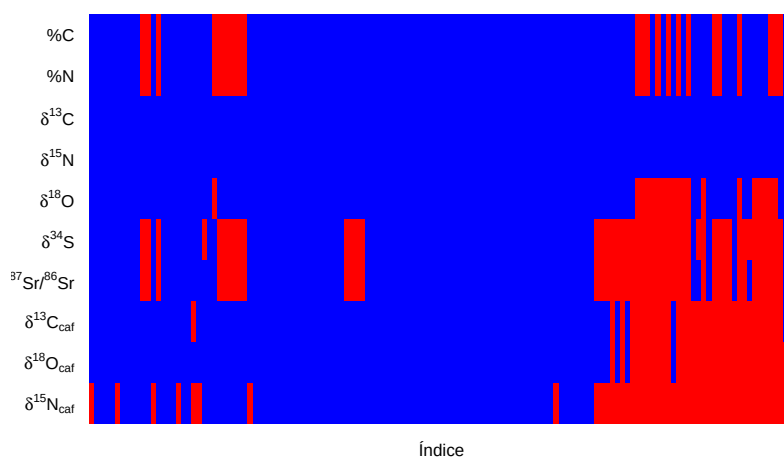


Figura 3.1: Padrão de valores omissos (representados a vermelho).

Em alguns casos, os algoritmos não lidam bem com valores omissos. Noutros casos, a presença de valores omissos pode resultar em decisões erradas (Nanni *et al.*, 2012). Assim, torna-se necessário recorrer a estratégias para lidar com este problema. Neste trabalho consideraram-se as estratégias que recorrem aos seguintes métodos:

- análise de casos completos;
- análise de casos disponíveis;
- métodos de imputação.

A análise de casos completos consiste em apenas considerar as observações sem qualquer valor omissos (completas). Este método é o mais simples, tendo a grande desvantagem de perder toda a informação acerca das observações com pelo menos um valor omissos.

A análise de casos disponíveis consiste em utilizar diferentes subconjuntos dos dados para diferentes análises, em que destes subconjuntos apenas fazem parte as variáveis necessárias para a respetiva análise e as observações sem qualquer valor omissos nessas variáveis. O uso de diferentes subconjuntos para diferentes análises pode levar a inconsistências e resultados que não podem ser comparados (Faria *et al.*, 2014).

Os métodos de imputação baseiam-se em substituir os valores omissos por valores estimados. As técnicas de imputação simples baseiam-se em substituir os valores omissos pela média/mediana ou valores preditos através de regressão. Uma das técnicas mais utilizadas para lidar com valores omissos é a imputação múltipla. Esta técnica consiste em substituir cada valor omissos por um conjunto de valores preditos, obtidos a partir da distribuição preditiva *a posteriori* dos valores omissos dados os valores observados. Ao contrário da imputação simples, a imputação múltipla reconhece a incerteza associada tanto aos valores omissos como aos parâmetros do modelo de imputação (Faria *et al.*, 2014). Estes métodos têm a vantagem de não descartar informação dos dados. No entanto, caso o número de valores omissos seja muito elevado, a aplicação destes métodos pode tornar os dados artificiais.

Para lidar com valores omissos neste trabalho, os métodos de imputação foram descartados à partida devido ao elevado número de valores omissos e ao seu padrão de ocorrência. A estratégia seguida foi a da análise de casos completos e de casos disponíveis.

A exploração dos dados, considerando a questão dos valores omissos, o interesse das variáveis disponíveis para o nosso estudo e o número de observações *versus* o número de variáveis conduziram à construção de 10 conjuntos de variáveis, que foram explorados em paralelo. Na Tabela 3.2 apresenta-se a lista de variáveis consideradas em cada conjunto de variáveis e na Tabela 3.3 apresenta-se o número de observações completas em cada subconjunto formado e origem geográfica do café (note-se que a última coluna refere-se ao número de observações existente no conjunto de dados original, incluindo observações completas e incompletas).¹

Para melhor representação e distinção das origens geográficas, quando representadas em gráficos e a menos de referência em contrário, estas vão ser distinguidas por 4 cores: a preto serão representadas as observações do Brasil, a **vermelho** as da Etiópia, a **verde** as do Havai e a **azul** as observações originárias do Quênia.

¹Do conjunto de variáveis *set*₁ constam todas as variáveis disponíveis, incluindo informação geográfica. O *set*₂ representa

Tabela 3.2: Conjuntos de variáveis formados.

	$C\%$	$N\%$	$\delta^{13}C$	$\delta^{18}O$	$\delta^{15}N$	$\delta^{34}S$	$\delta^{13}C_{caf}$	$\delta^{18}O_{caf}$	$\delta^{15}N_{caf}$	$^{87}Sr/^{86}Sr$
<i>set</i> ₃	×	×	×	×	×	×	×	×	×	×
<i>set</i> ₄	×	×	×	×	×	×	×	×		×
<i>set</i> ₅				×	×		×	×	×	×
<i>set</i> ₆	×	×					×	×	×	
<i>set</i> ₇			×	×	×					
<i>set</i> ₈			×	×	×		×	×		
<i>set</i> ₉							×	×	×	×
<i>set</i> ₁₀						×	×	×	×	×
<i>set</i> ₁₁			×	×	×	×				×
<i>set</i> ₁₂			×	×	×					×

Tabela 3.3: Número de observações sem valores omissos em cada subconjunto, sendo a última coluna o número de observações (N.^o obs.) existentes no conjunto de dados em estudo.

	<i>set</i> ₃	<i>set</i> ₄	<i>set</i> ₅	<i>set</i> ₆	<i>set</i> ₇	<i>set</i> ₈	<i>set</i> ₉	<i>set</i> ₁₀	<i>set</i> ₁₁	<i>set</i> ₁₂	N. ^o obs.
Brasil	19	20	19	19	42	29	20	20	23	25	57
Etiópia	1	3	1	5	12	10	1	1	4	4	15
Havai	46	47	47	47	52	50	47	46	49	50	52
Quênia	10	13	10	10	13	13	10	10	13	13	14
Total	76	83	77	81	119	102	78	77	89	92	138

Definidas as estratégias a utilizar de forma a lidar com os valores omissos existentes no conjunto de dados em estudo, o próximo passo é a exploração dos dados. Assim, nas próximas secções procura-se encontrar padrões entre os dados e as variáveis, assim como as variáveis mais relevantes para o problema da diferenciação da origem geográfica do grão verde de café.

3.2 Análise descritiva

De forma a extrair mais informações do conjunto de dados em estudo, são de seguida apresentadas algumas medidas descritivas e gráficos dos dados. Estas medidas foram obtidas do conjunto de dados tratando os valores omissos pelo método de análise de casos disponíveis, isto é, cada medida em cada variável (par de variáveis, no caso dos gráficos) foi obtida recorrendo a todas as observações sem valor omissos na variável (no subconjunto formado pelas observações no par de variáveis).

Na Tabela 3.4 apresentam-se algumas medidas de localização e dispersão para cada variável. Uma primeira análise permite verificar que os intervalos de variação das variáveis diferem significativamente, bem como as suas variâncias. As variáveis com maior variância são $\delta^{18}O$ e $\delta^{34}S$ (35.4 e 22.7, respetivamente), por oposição as de menor variância são $^{87}Sr/^{86}Sr$ e $N\%$ (1.8×10^{-5} e 0.052, respetivamente). Os valores obtidos em $\delta^{13}C$ e $\delta^{13}C_{caf}$ permitem concluir que, em geral, os grãos de café verde são pobres no isótopo pesado de carbono (variando entre -30.60‰ a -22.10‰) e com elevada percentagem de carbono (entre 41.50% a 51.10%), enquanto que $\delta^{18}O$ e $^{87}Sr/^{86}Sr$ apresentam intervalos de valores positivos (entre 22.10‰ e 53.00‰ e entre 0.70‰ e 0.73‰, respetivamente), significando que os grãos verdes de café estão enriquecidos em isótopos pesados destes elementos químicos.

Realça-se ainda a semelhança entre as medidas descritivas de $\delta^{13}C$ e $\delta^{13}C_{caf}$ e a diferença entre o valor δ da razão isotópica do oxigénio no grão verde ($\delta^{18}O$) e na molécula de cafeína do grão ($\delta^{18}O_{caf}$), pois a variância de $\delta^{18}O_{caf}$ é muito menor que a de $\delta^{18}O$, bem como a média e restantes medidas de localização (por exemplo a mediana difere aproximadamente 25‰). Isto deve-se ao facto de o todas as variáveis do *set*₃, bem como a composição isotópica do oxigénio da precipitação.

oxigênio presente na molécula de cafeína do grão verde refletir a composição isotópica da água utilizada pela planta, pois não há fracionamento durante a captação de água pela planta (Rodrigues, 2011, p. 25), enquanto que $\delta^{18}O$ no grão pode refletir diversos fenômenos que ocorrem no seu interior. Mais ainda, como se pode verificar na Figura 2.2, a água da chuva tem valores de $\delta^{18}O$ predominantemente negativos, levando a que os valores de $\delta^{18}O_{caf}$ apareçam reduzidos. Por fim, em relação aos valores omissos, as variáveis $\delta^{13}C$ e $\delta^{15}N$ são as únicas sem valores omissos e $\delta^{34}S$, $\delta^{15}N_{caf}$ e $^{87}Sr/^{86}Sr$ as com o maior número, como se pode constatar ao consultar a Tabela 3.1 e a Figura 3.1.

Tabela 3.4: Medidas de localização e dispersão no conjunto de dados.

	Min.	1.ºQ	Mediana	Média	3.ºQ	Máx.	Variância	NA's
$C\%$	41.50	45.50	46.10	46.27	47.00	51.10	2.228	23
$N\%$	1.30	2.00	2.20	2.17	2.30	2.80	0.052	23
$\delta^{13}C$	-30.60	-27.60	-26.90	-26.77	-25.90	-22.10	1.791	0
$\delta^{18}O$	22.10	27.05	29.10	30.52	31.45	53.00	35.437	19
$\delta^{15}N$	0.200	1.625	2.600	2.781	3.800	6.500	1.905	0
$\delta^{34}S$	-1.500	6.825	9.350	9.633	11.850	21.400	22.657	48
$^{87}Sr/^{86}Sr$	0.7045	0.7062	0.7071	0.7084	0.7086	0.7262	1.8×10^{-5}	45
$\delta^{13}C_{caf}$	-30.30	-28.10	-27.40	-27.21	-26.50	-22.10	1.586	33
$\delta^{18}O_{caf}$	-2.600	1.700	4.000	4.246	7.000	11.500	9.756	33
$\delta^{15}N_{caf}$	-2.500	-0.450	0.300	0.697	1.800	6.500	3.046	47

Com o objetivo de analisar as relações existentes entre as variáveis, na Figura 3.2 estão representados os diagramas de dispersão entre as variáveis (incluindo as variáveis com informação geográfica) e as correlações amostrais entre elas. Nos gráficos de dispersão diferenciam-se as origens geográficas por cores diferentes. As correlações amostrais mais *fortes* (em valor absoluto superior ou igual a 0.70) estão identificadas a negrito e são discutidas de seguida.

A forte correlação, $|-0.92|$, entre a altitude e $\delta^{18}O_{prec}$ deve-se ao efeito isotópico de altitude (a água da precipitação vai tornando-se mais negativa em $\delta^{18}O$ com o aumento da altitude). A forte correlação entre a longitude e a altitude, 0.77, e entre a longitude e $\delta^{18}O_{prec}$, $|-0.74|$, deve-se ao efeito continentalidade, uma vez que no conjunto de dados em estudo as regiões mais interiores estão a longitudes e altitudes mais elevadas (Quênia e Etiópia). Verifica-se ainda uma forte correlação entre $\delta^{15}N$ e $\delta^{15}N_{caf}$, 0.70, e um pouco menos elevada entre $\delta^{13}C$ e $\delta^{13}C_{caf}$, 0.65, estas correlações eram já esperadas por se tratarem de composições isotópicas do mesmo elemento químico, uma medida no grão verde como um todo e a outra no seu composto específico, a cafeína. No entanto, tal não se verifica na correlação entre $\delta^{18}O$ e $\delta^{18}O_{caf}$, 0.30. Esta reduzida correlação deve-se às observações atípicas que o Brasil apresenta em $\delta^{18}O$.

Analisando os diagramas de dispersão da Figura 3.2, além de algumas observações atípicas do Brasil com $\delta^{18}O$ elevado, realce-se também a forma com as observações referentes a $^{87}Sr/^{86}Sr$ (Brasil) e $\delta^{34}S$ (Havai) se destacam das restantes, nomeadamente a elevada variabilidade de $\delta^{34}S$ associada às observações do Havai.

Repare-se que as variáveis $\delta^{15}N_{caf}$ e $\delta^{18}O_{caf}$ demonstram elevada importância na diferenciação entre as regiões africanas e as restantes regiões em estudo. Portanto representa-se, de forma mais clara, na Figura 3.3 o diagrama de dispersão destas duas variáveis. Neste diagrama nota-se uma razoável separação entre as observações provenientes do Quênia e da Etiópia das restantes, por apresentarem, simultaneamente, valores elevados em $\delta^{18}O_{caf}$ e em $\delta^{15}N_{caf}$, à exceção uma observação da Etiópia que apresenta $\delta^{18}O_{caf}$ muito reduzido.

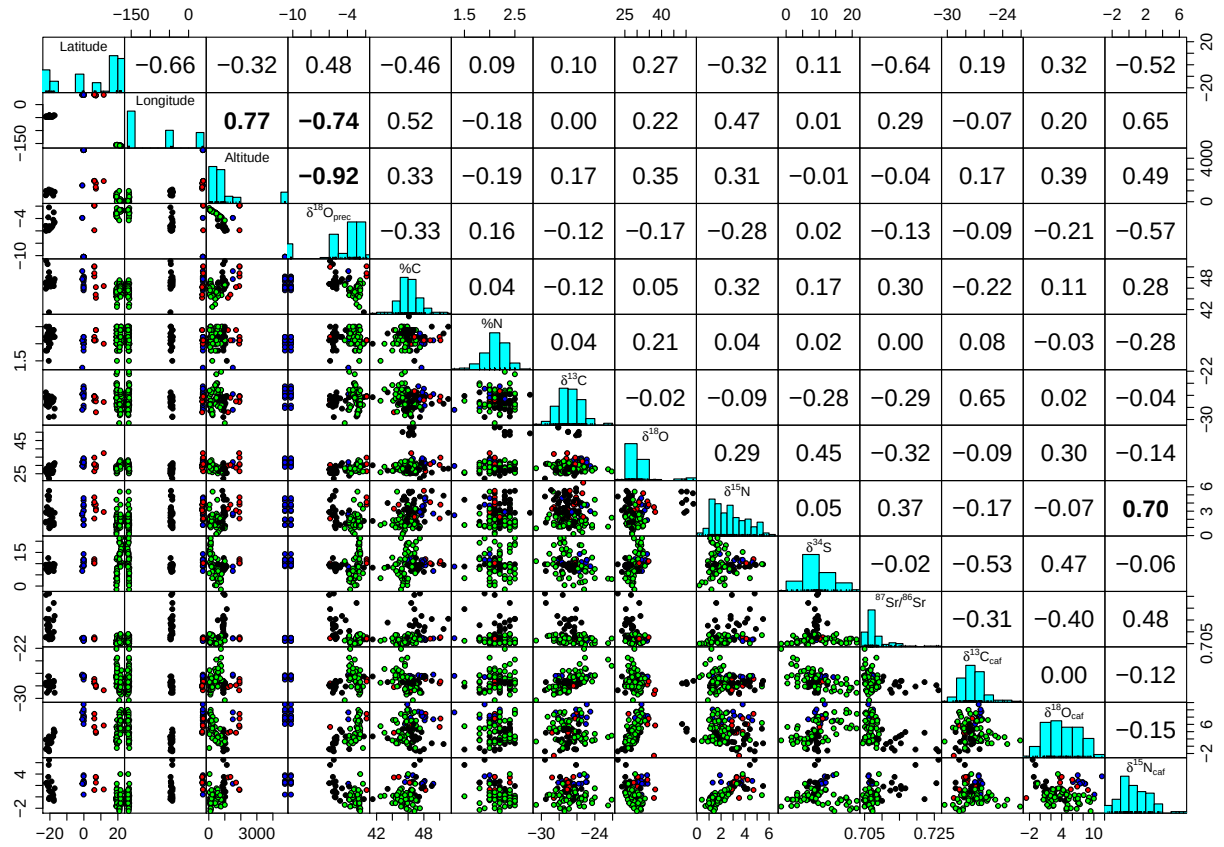


Figura 3.2: Diagramas de dispersão e correlações amostrais entre as variáveis, com as observações coloridas de acordo com a origem geográfica (preto-Brasil, vermelho-Etiópia, verde-Havai e azul-Quênia).

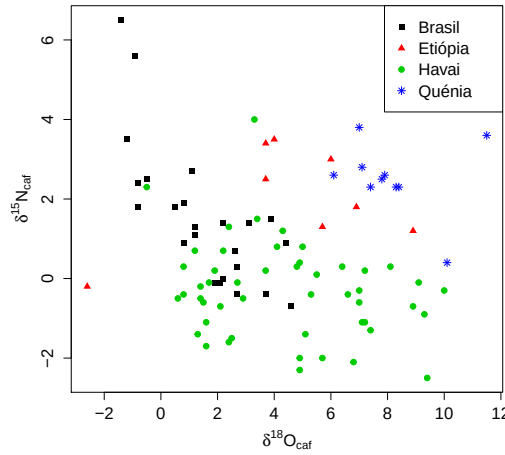


Figura 3.3: Diagrama de dispersão das variáveis $\delta^{15}N_{caf}$ e $\delta^{18}O_{caf}$.

A presença de observações atípicas no conjunto de dados em estudo pode afetar severamente as correlações amostrais, obtidas por métodos clássicas, como foi o caso da correlação entre $\delta^{18}O$ e $\delta^{18}O_{caf}$. Por este motivo foram também obtidas as correlações amostrais recorrendo a um estimador robusto, nomeadamente o determinante de covariância mínima (*Fast MCD*) (Maronna *et al.*, 2006). Estas correlações apresentam-se na Tabela A.1. As correlações robustas demonstram uma associação mais *forte* entre as variáveis que as correlações obtidas através de um estimador clássico. Destacando-se

a correlação entre $\delta^{18}O$ e $\delta^{18}O_{caf}$, 0.71, que se esperava elevada, mas na análise anterior estava claramente afetada pelas observações atípicas do Brasil. Concluindo, as variáveis com relação mais forte com as variáveis geográficas são as referentes às razões isotópicas do oxigénio e do enxofre.

É também importante estudar o comportamento dos dados agrupados de acordo com a sua origem geográfica. Para isso são de seguida analisadas as correlações amostrais entre as variáveis, utilizando as observações dos dois países com maior representação na amostra: Havai (Figura 3.4) e Brasil (Figura 3.5). Devido ao reduzido número de observações das restantes regiões não é possível obter correlações estatisticamente significativas.

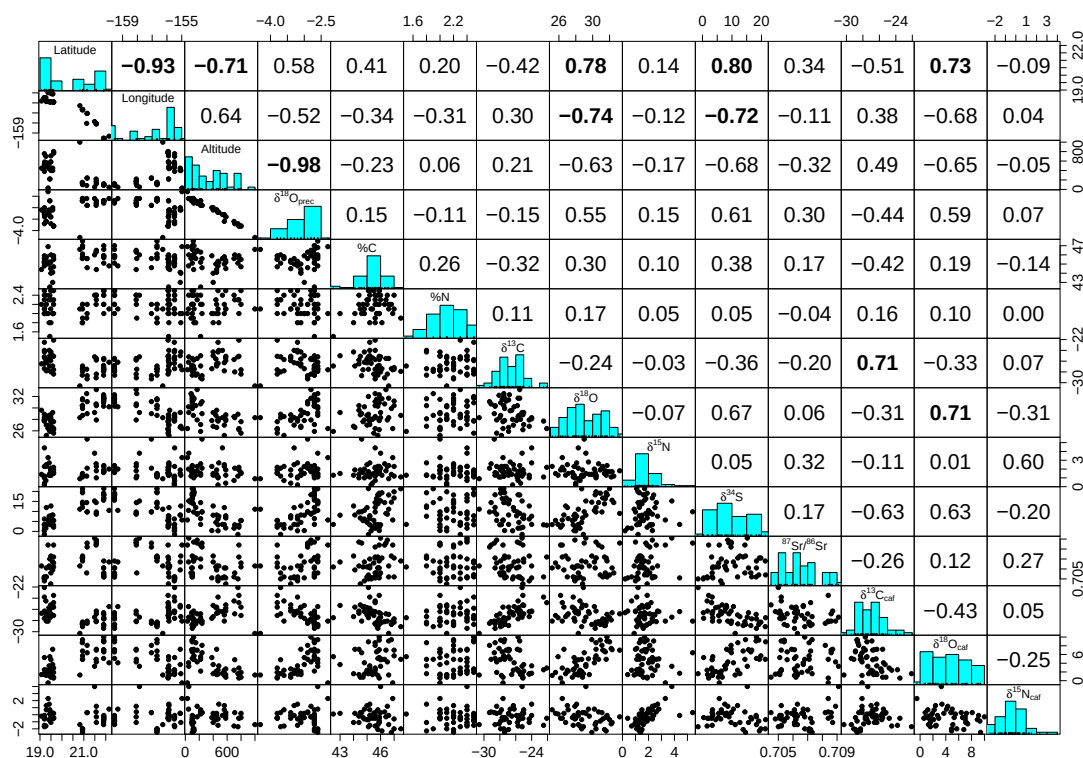


Figura 3.4: Diagramas de dispersão e respetivas correlações amostrais referentes ao Havai.

Das correlações obtidas a partir das observações das regiões produtoras do Havai, presentes na Figura 3.4, destacam-se as fortes correlações entre $\delta^{13}C$ e $\delta^{13}C_{caf}$, 0.71, e entre $\delta^{18}O$ e $\delta^{18}O_{caf}$, 0.71. Estes valores devem-se ao facto de serem composições isotópicas do mesmo elemento químico, uma medida no grão verde e outra na molécula de cafeína do grão verde. A correlação entre $\delta^{18}O$ e $\delta^{34}S$, 0.67, está relacionada com os efeitos da altitude e da distância ao mar, que afetam $\delta^{18}O$ e $\delta^{34}S$, respetivamente. Estes efeitos estão relacionados, pois no caso das ilhas do Havai, devido à atividade vulcânica submarina, as observações perto do mar tornam-se enriquecidas em isótopos pesados de enxofre, sendo estas as de menor altitude e portanto mais enriquecidas em isótopos pesados de oxigénio. As correlações elevadas entre as variáveis geográficas e as restantes refletem também estes fatores, sendo de destacar as correlações entre a altitude e $\delta^{18}O$, -0.63, $\delta^{18}O_{caf}$, -0.65, e $\delta^{34}S$, -0.68. A altitude apresenta ainda uma elevada correlação, em módulo, com $\delta^{18}O_{prec}$, -0.98. Como se referiu anteriormente, esta é uma relação conhecida que se deve ao empobrecimento da composição isotópica da água da precipitação com o aumento da altitude. As fortes correlações entre

as composições isotópicas do oxigênio/enxofre e a latitude e longitude derivam também do que foi referido anteriormente (efeito de altitude e distância ao mar/vulcões). Por fim, a forte correlação entre a latitude e longitude, $|-0.93|$, deve-se à forma como as regiões do Havai se dispõem no globo terrestre.

Na Tabela A.2, apresentam-se as correlações amostrais obtidas através do estimador robusto MCD. À semelhança do que aconteceu anteriormente, as correlações robustas demonstram uma associação forte entre as variáveis referentes às composições isotópicas do oxigênio e do enxofre entre si e entre as variáveis de localização geográfica, neste caso destaca-se ainda a elevada correlação, em valor absoluto, entre $\delta^{13}C_{caf}$ e a altitude, -0.79 , a latitude, -0.72 , $\delta^{18}O_{prec}$, -0.79 , e $\delta^{34}S$, -0.83 . Estas relações podem ser explicadas pela relação conhecida entre o clima e a variação de altitude e de latitude, pois o clima é um dos principais fatores que condiciona o metabolismo primário das plantas (que se reflete em $\delta^{13}C_{caf}$).

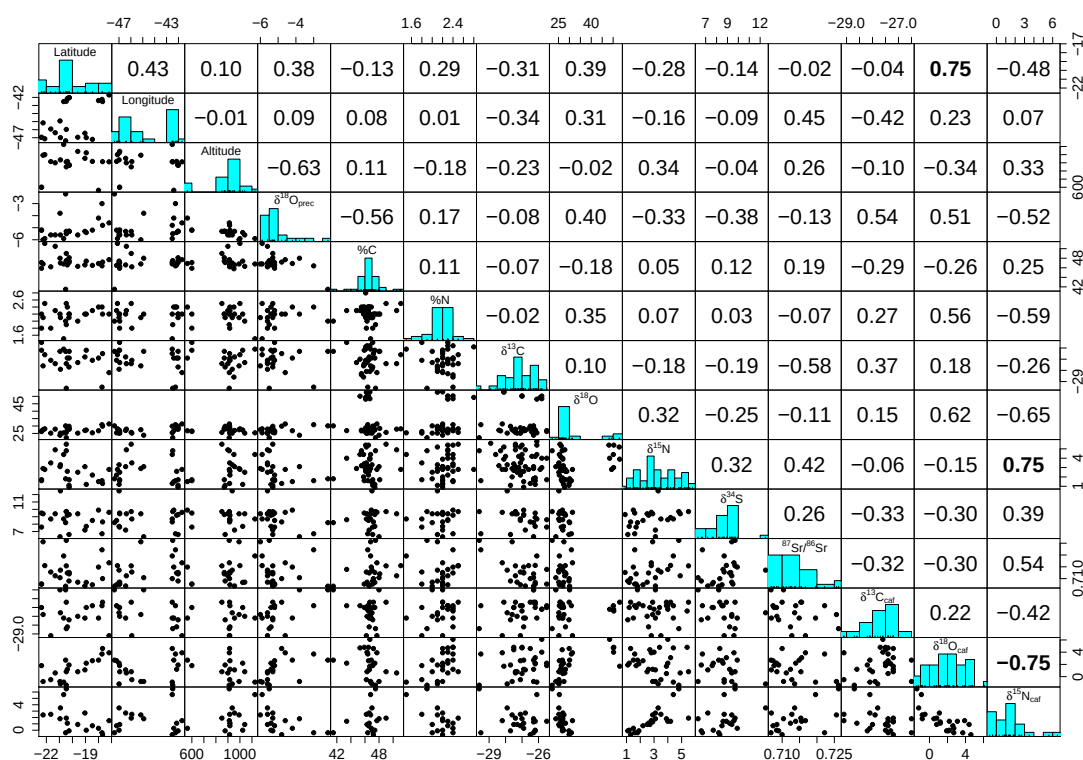


Figura 3.5: Diagramas de dispersão e respectivas correlações amostrais referentes ao Brasil.

No que diz respeito às correlações obtidas através das observações do Brasil (Figura 3.5), apenas se verifica *forte* correlação entre $\delta^{15}N$ e $\delta^{15}N_{caf}$, latitude e $\delta^{18}O_{caf}$, ambas 0.75, e $\delta^{18}O_{caf}$ e $\delta^{15}N_{caf}$, -0.75 . Estes resultados encontram justificações similares às apontadas no caso do Havai. Ao contrário do que era esperado, verifica-se uma reduzida correlação entre $\delta^{13}C$ e $\delta^{13}C_{caf}$, 0.37, que, pela análise do diagrama de dispersão correspondente a estas variáveis, parece dever-se à existência de observações atípicas, que apresentam $\delta^{13}C$ reduzido. Não se apresentam correlações amostrais robustas para este caso pois o número de observações é demasiado pequeno para a generalidade dos algoritmos de estimação de correlações robustas, nomeadamente o algoritmo *Fast MCD* (Maronna *et al.*, 2006).

Na Figura 3.6, representam-se os diagramas em caixa (ou *box plots*) do conjunto de dados em estudo, separados por origem geográfica (Brasil, Etiópia, Havai e Quênia). Estes diagramas permitem confirmar a elevada variabilidade de $\delta^{34}S$ inerente às observações do Havai, assim como a de $\delta^{13}C$ e $\delta^{13}C_{caf}$. Além disso, a variabilidade de $^{87}Sr/^{86}Sr$ é explicada principalmente pelas observações do Brasil, sendo de destacar nesta variável a predominância de valores reduzidos no Havai e elevados no Brasil. A elevada variância de $\delta^{18}O$ deve-se a observações atípicas do Brasil, que apresentam valores elevados nesta variável.

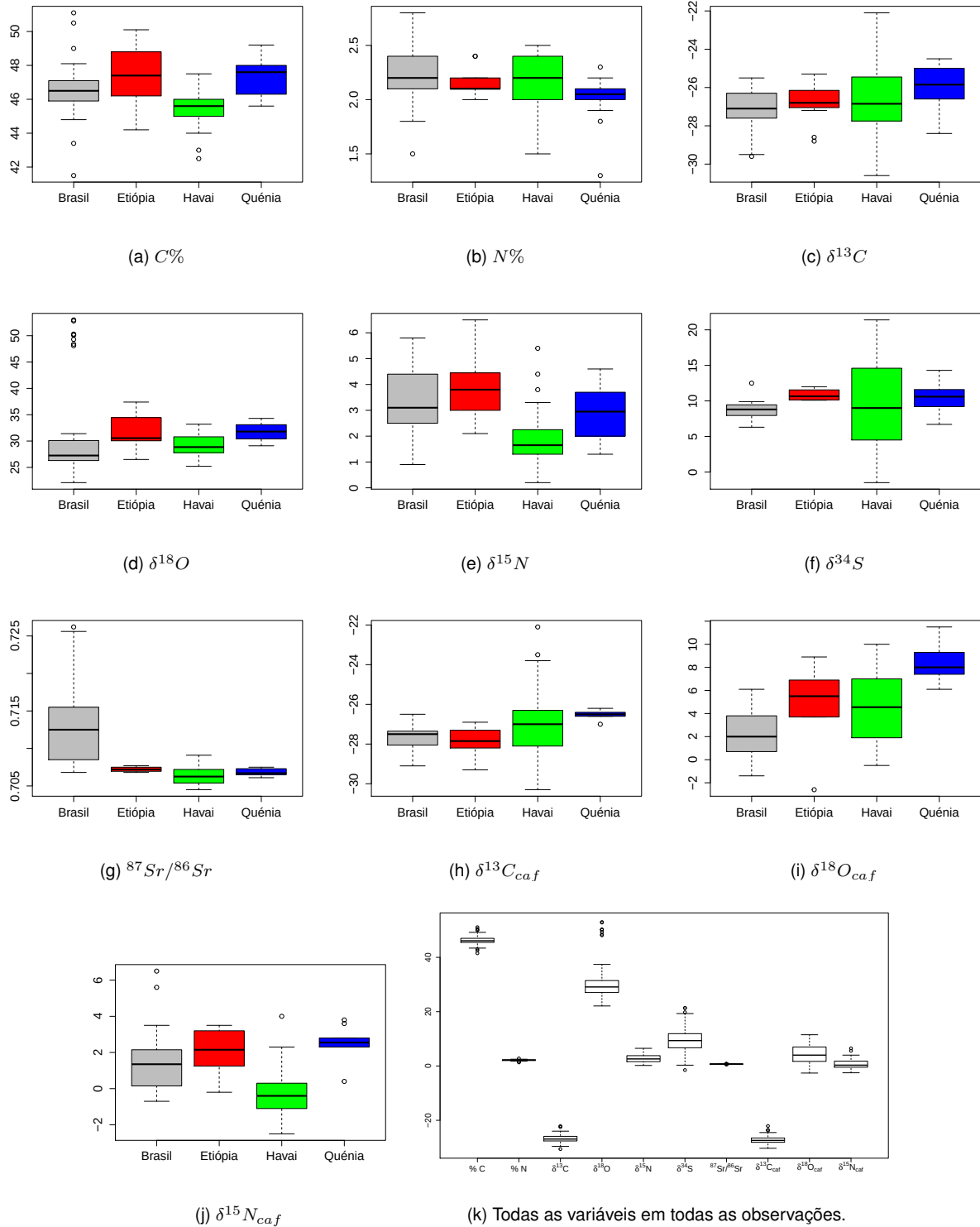


Figura 3.6: Diagramas em caixa das características do café.

Na Figura 3.6k encontram-se os diagramas em caixa das 10 variáveis em estudo do conjunto de dados. Estes diagramas permitem confirmar a diferente gama de variação das variáveis.

Em suma, as análises efetuadas nesta secção sugerem que as variáveis com maior potencial de diferenciação da origem geográfica do grão verde de café são a razão isotópica do estrôncio ($^{87}\text{Sr}/^{86}\text{Sr}$) e as composições isotópicas do oxigénio e do azoto, nomeadamente referentes à molécula de cafeína do grão verde ($\delta^{18}\text{O}_{caf}$ e $\delta^{15}\text{N}_{caf}$).

3.3 Análise em componentes principais

3.3.1 Introdução

Os principais objetivos da análise em componentes principais, doravante designada por ACP, são a redução da dimensionalidade dos dados e a sua interpretação, permitindo, frequentemente, encontrar relações não conhecidas nos dados (Johnson e Wichern, 2007, p. 430). Esta técnica baseia-se em encontrar combinações lineares das variáveis que expliquem ao máximo a variação existente nos dados.

As componentes principais (abreviadas por CPs) dependem apenas da matriz de covariâncias das p variáveis aleatórias, $X_1, X_2, \dots, X_p$. Sejam $\mathbf{X} = (X_1, X_2, \dots, X_p)^t$ o vetor aleatório e $\Sigma_{\mathbf{X}}$ a respetiva matriz de covariâncias, com valores próprios $\lambda_1 \geq \lambda_2 \geq \lambda_p \geq 0$ e vetores próprios $\gamma_1, \gamma_2, \dots, \gamma_p$, onde t denota transposto. Então as componentes principais são definidas pelas combinações lineares em (3.1), maximizando as respetivas variâncias em (3.2) sob as restrições de ortogonalidade em (3.3) e correlações nulas entre elas (Equação (3.4)) (Johnson e Wichern, 2007, p. 431).

$$CP_i = \gamma_i^t \mathbf{X} = \lambda_{i1}X_1 + \lambda_{i2}X_2 + \dots + \lambda_{ip}X_p \quad i = 1, 2, \dots, p \quad (3.1)$$

$$\text{Var}(CP_i) = \text{Var}(\gamma_i^t \mathbf{X}) = \gamma_i^t \Sigma_{\mathbf{X}} \gamma_i = \lambda_i \quad i = 1, 2, \dots, p \quad (3.2)$$

$$\gamma_i^t \gamma_k = 1 \quad i, k = 1, 2, \dots, p, \quad i \neq k \quad (3.3)$$

$$\text{Cov}(CP_i, CP_k) = \text{Cov}(\gamma_i^t \mathbf{X}, \gamma_k^t \mathbf{X}) = \gamma_i^t \Sigma_{\mathbf{X}} \gamma_k = 0 \quad i, k = 1, 2, \dots, p, \quad i \neq k. \quad (3.4)$$

Desta forma, a primeira componente principal será a combinação linear com variância máxima e a última (p -ésima componente principal) a de menor variância. Assim, as p componentes principais reproduzem a variabilidade total dos dados, com variâncias decrescentes. A variabilidade dos dados está essencialmente concentrada nas primeiras k componentes principais ($k < p$), sendo que as restantes $p - k$ componentes refletem, muitas vezes, algum ruído existente nos dados. Com efeito, substituindo as p variáveis iniciais pelas k componentes principais consegue-se uma redução significativa da dimensão dos dados, obtendo-se um conjunto de dados com n observações em k componentes principais, em vez de n observações em p variáveis (Johnson e Wichern, 2007, p. 430). Aos valores dos vetores próprios de uma componente principal, $\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{ip}$, $i = 1, 2, \dots, p$, dá-se o nome de *loadings*. Enquanto que aos valores observados de uma componente principal no conjunto de dados, $CP_i(\mathbf{X})$, dá-se o nome de *scores*.

Uma das grandes desvantagens da ACP com base nas variáveis originais (não padronizadas) é a sensibilidade das componentes principais à gama de variação e unidades de medida de cada variável. De facto, se existir uma grande disparidade entre as variâncias das variáveis, as de maior variância tendem a dominar os *loadings* das primeiras componentes principais (Johnson e Wichern, 2007, p. 439; Jolliffe, 2002, p. 22). Uma solução para este problema é a estimação das componentes principais a partir das variáveis padronizadas (com média nula e variância unitária). Desta forma, as variáveis padronizadas, $X_1^*, X_2^*, \dots, X_p^*$, são definidas por (3.5) (onde μ_i representa a média e σ_{ii} o desvio padrão de X_i) e as componentes principais por (3.6):

$$X_i^* = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \quad i = 1, 2, \dots, p \quad (3.5)$$

$$CP_i = (\gamma_i^*)^t \mathbf{X}^* \quad i = 1, 2, \dots, p. \quad (3.6)$$

Com efeito, γ_i^* corresponde ao i -ésimo vetor próprio da matriz de covariâncias dos dados padronizados, $\mathbf{X}^*$. Uma vez que a matriz de covariâncias de $\mathbf{X}^*$ corresponde à matriz de correlações de $\mathbf{X}$, γ_i^* corresponde ao i -ésimo vetor próprio da matriz de correlações de $\mathbf{X}$ (Jolliffe, 2002, p. 21).

Um dos pontos importantes da ACP é a interpretação das componentes principais, que geralmente é realizada recorrendo à análise dos *loadings* de cada componente principal. No entanto, nem sempre é tarefa fácil. Truncar as componentes principais, ignorando os *loadings* próximos de zero, é uma técnica que ajuda na interpretação; esta técnica é considerada segura já que ignorar estes *loadings* levará a desvios mínimos nas direções das variáveis que eles multiplicam (Cadima e Jolliffe, 1995). Contudo, Cadima e Jolliffe (1995) consideram que esta prática pode induzir em erro, pois a resultante das componentes principais (uma combinação linear) não depende apenas dos coeficientes, mas também das variâncias e do padrão de correlações das variáveis que as definem. Assim, deve-se utilizar informação complementar de forma a validar as interpretações baseadas nos *loadings*, como por exemplo as correlações entre o valor observado das componentes principais estimadas e das variáveis. Para Cadima e Jolliffe (1995) uma alternativa é recorrer aos coeficientes de correlação múltipla, na escolha do subconjunto de variáveis a interpretar, e regressão das componentes principais nos subconjuntos de variáveis. No entanto referem que o método de regressão das componentes principais nas variáveis levanta numerosas dificuldades técnicas.

A estimação dos *loadings* e variâncias das componentes principais, baseadas na estimativa da matriz de covariâncias amostral, pode ser severamente afetada pela existência de observações atípicas. A alternativa consiste no uso de estimativas robustas, que se caracterizam por acomodar potenciais observações discordantes. Os estimadores robustos mais populares dividem-se em dois grupos. Do primeiro grupo fazem parte os estimadores que propõem o uso de uma matriz de covariâncias robusta, como por exemplo o estimador obtido pelo método do determinante de covariâncias mínimo, *Fast MCD* (Maronna *et al.*, 2006). Em alternativa, foram propostos outros métodos robustos que exploram a projeção de dados multivariados em espaços de baixa dimensão, otimizando uma medida adequada de padrões nos dados. No caso particular da ACP, procura-se a direção que maximiza uma medida robusta de dispersão dos dados projetados, sujeita a restrições apropriadas (equações (3.3) e (3.4)). Estes

são denominados métodos de projeção ótima (*Projection Pursuit*, Croux *et al.*, 2007), neste trabalho recorreu-se a ACP que utiliza o algoritmo de Croux e Ruiz-Gazen (2005).

A escolha do número de componentes principais a reter é fulcral e são várias as estratégias para auxiliar o investigador nesta tarefa. Aqui ter-se-á em conta: a proporção de variabilidade total explicada pelas componentes principais retidas (aproximadamente 80%); o *scree plot* (gráfico da variância vs. número de componentes principais), utilizando a regra do cotovelo; e a comparação da variância de cada componente ($\hat{\lambda}_i$) com a média das variâncias de todas as componentes principais ($\hat{\lambda}_i \geq \bar{\lambda}$, $i = 1, \dots, k$), sendo que esta média será representada por uma linha a azul nos *scree plots*.

3.3.2 Resultados da aplicação da ACP aos dados

Em virtude do elevado número de valores omissos, além das combinações referidas na Tabela 3.5, foi também realizada ACP recorrendo à estimação da matriz de covariâncias (utilizando o coeficiente de correlação de Pearson) com todas as observações do conjunto de dados, através da estimação por comparação dos pares de observações completas. Contudo, este método levou a resultados muito semelhantes ao da ACP utilizando apenas observações completas dos conjuntos de variáveis.

Tabela 3.5: Combinações de métodos de estimação de ACP aplicadas com indicação dos pacotes do *software* R utilizados.

Tipo de análise	Pacote utilizado
ACP clássica	princomp do pacote stats
ACP via <i>Projection Pursuit</i>	PCAprój do pacote pcaPP

Devido ao elevado número de subconjuntos considerados para análise (referidos na Secção 3.1) e métodos de estimação das componentes principais, não serão analisadas nesta subsecção todas as combinações testadas (a análise tornar-se-ia redundante), sendo apenas discutidas as que foram consideradas interessantes tendo em vista o problema em estudo. Mais ainda, dado que o interesse reside no estudo das características químicas do grão verde do café e sendo o objetivo deste trabalho a determinação da sua origem geográfica, não foram consideradas as variáveis de localização para estimação das componentes principais, nomeadamente a latitude, longitude e altitude.

A diversidade de magnitude de dispersão entre as variáveis observadas conduziu a optar-se por definir as componentes principais amostrais a partir das variáveis padronizadas, recorrendo à média e ao desvio padrão amostral.

A interpretação das componentes principais obtidas foi baseada nas variáveis observadas cujos *loadings* associados são, em valor absoluto, superiores a 0.25 (i.e., a j -ésima variável é considerada na interpretação da i -ésima componente principal se $|\hat{\lambda}_{ij}| > 0.25$); estes *loadings* serão apresentados a **negrito**. Estas escolhas foram validadas, tal como sugerem Cadima e Jolliffe (1995), calculando a correlação amostral entre os pares associados à componente principal em estudo e aos que se obteriam se os *loadings* cujo valor absoluto é inferior ou igual a 0.25 fossem substituídos por zero.

A interpretação das componentes principais estimadas, obtidas nos diferentes casos, foi discutida com peritos desta área de conhecimento, que ajudaram a estabelecer relações entre a abundância de cada isótopo, de cada elemento químico, e aspetos como a região produtora, práticas agrícolas, etc. Por exemplo, é sabido que os valores de $\delta^{18}\text{O}$ das plantas refletem o valor de $\delta^{18}\text{O}$ da precipitação, que varia

sistematicamente com a latitude e a altitude (como se refere na Secção 2.2). Outro exemplo é a razão isotópica do azoto que é um bom indicador de práticas agrícolas (Rodrigues, 2011, p. 76). Os valores de $\delta^{13}C$ estão associados ao metabolismo primário da planta. Em relação à razão isotópica do estrôncio ($^{87}Sr/^{86}Sr$) e à composição isotópica do enxofre ($\delta^{34}S$), estas variáveis dependem fortemente das características do solo e da rocha mãe. Enquanto que o enxofre (S) é não só um poluente atmosférico, como também um nutriente essencial para as plantas (Rodrigues, 2011, p. 28), o $^{87}Sr/^{86}Sr$ é um indicador do tipo e idade da rocha mãe do solo, na região onde a planta de café se desenvolveu.

Para simplificar a notação, nesta secção deixar-se-á de recorrer ao asterisco (*) para denotar padronização, assumindo-se doravante que as variáveis foram previamente padronizadas.

ACP clássica aplicada ao set_3

Nesta secção explora-se a análise em componentes principais padronizadas utilizando as observações que não registam valores omissos (ditas completas) e que dizem respeito ao conjunto de variáveis do set_3 (todas as variáveis referentes às características químicas do café em estudo).

Através da análise do *scree plot* da Figura 3.7 e da Tabela 3.6 verifica-se que as quatro primeiras das dez componentes principais explicam 78% da variabilidade total dos dados e $\hat{\lambda}_4 > \bar{\lambda} = 1$, pelo que foram retidas as primeiras quatro componentes principais. Os respetivos *loadings* são apresentados na Tabela 3.7a, onde se apresenta também o valor da correlação entre a componentes principal original e a componente principal truncada (CP_i^{tr} , $i = 1, 2, \dots, k$). As correlações entre o valor observado das componentes principais estimadas e as variáveis encontram-se na Tabela 3.7b.

Tabela 3.6: Variância estimada e proporção cumulativa de variabilidade total explicada pelas componentes principais da ACP clássica, set_3 (dados padronizados).

	CP_1	CP_2	CP_3	CP_4	CP_5	CP_6	CP_7	CP_8	CP_9	CP_{10}
Variância	2.81	2.44	1.44	1.07	0.58	0.51	0.47	0.26	0.23	0.20
Prop. cumulativa	0.28	0.53	0.67	0.78	0.83	0.88	0.93	0.96	0.98	1.00

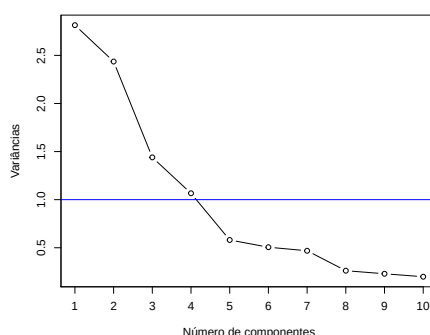


Figura 3.7: *Scree plot* da ACP clássica aplicada ao set_3 (dados padronizados).

Analisando os *loadings* estimados que constam da Tabela 3.7a, verifica-se que para todas as componentes principais os *loadings* dos pares correspondentes à composição isotópica do mesmo elemento, medida no grão e na molécula de cafeína, são semelhantes. Seguidamente será apresentada uma possível interpretação para cada componente principal retida, a interpretação feita com base nos *loadings* é semelhante à interpretação feita com base nas correlações amostrais da Tabela 3.7b. Os

scores da ACP, com identificação da origem geográfica, são representados na Figura 3.8, onde se faz a distinção entre as observações da ilha Havai (Havai_{ilha}) e as observações das restantes ilhas do arquipélago do Havai (Havai_{arq}).

Tabela 3.7: *Loadings* e correlações amostrais entre as CPs e as variáveis, correspondentes à ACP clássica aplicada ao *set*₃ (dados padronizados).

(a) <i>Loadings</i>					(b) Correlações amostrais entre as CPs e as variáveis.				
	<i>CP</i> ₁	<i>CP</i> ₂	<i>CP</i> ₃	<i>CP</i> ₄		<i>CP</i> ₁	<i>CP</i> ₂	<i>CP</i> ₃	<i>CP</i> ₄
<i>C</i> %	0.281	-0.264	-0.197	0.442	<i>C</i> %	0.472	-0.411	-0.237	0.456
<i>N</i> %	-0.174	-0.045	0.149	0.849	<i>N</i> %	-0.291	-0.070	0.178	0.876
$\delta^{13}C$	-0.255	0.320	-0.474	0.034	$\delta^{13}C$	-0.428	0.499	-0.569	0.035
$\delta^{18}O$	-0.293	-0.416	-0.301	0.030	$\delta^{18}O$	-0.492	-0.649	-0.361	0.030
$\delta^{15}N$	0.413	-0.036	-0.371	0.067	$\delta^{15}N$	0.692	-0.057	-0.513	0.069
$\delta^{34}S$	-0.012	-0.371	0.017	-0.114	$\delta^{34}S$	-0.021	-0.841	0.021	-0.118
$^{87}Sr/^{86}Sr$	0.467	0.015	0.105	0.161	$^{87}Sr/^{86}Sr$	0.783	0.024	0.126	0.166
$\delta^{13}C_{caf}$	-0.279	0.406	-0.371	0.163	$\delta^{13}C_{caf}$	-0.468	0.634	-0.445	0.168
$\delta^{18}O_{caf}$	-0.279	-0.442	-0.319	-0.071	$\delta^{18}O_{caf}$	-0.468	-0.689	-0.383	-0.074
$\delta^{15}N_{caf}$	0.442	0.039	-0.436	-0.086	$\delta^{15}N_{caf}$	0.742	0.061	-0.523	-0.089
$Cor(CP_i, CP_j^{tr})$	0.986	0.997	0.925	0.947					

1.^a componente principal: Contraste entre $\delta^{13}C$, $\delta^{18}O$, $\delta^{13}C_{caf}$, $\delta^{18}O_{caf}$ e *C*%, $\delta^{15}N$, $^{87}Sr/^{86}Sr$, $\delta^{15}N_{caf}$.

A primeira componente principal reflete o metabolismo primário da planta e fatores clima e altitude em contraste com o tipo de solo (influenciado pelas práticas agrícolas e idade da rocha mãe). *Scores* elevados nesta componente podem indicar que as respetivas observações são originárias de solos mais antigos, ou sujeitos a práticas agrícolas convencionais, enquanto que observações com *scores* reduzidos (negativos) indiciam que estas são provenientes de zonas mais próximas do mar, a baixas altitudes.

Destaca-se a maioria dos *scores* do Brasil por terem valores elevados e os do Havai por terem valores reduzidos na primeira componente principal (Figura 3.8a), o que sugere diferentes tipos de solo entre estas duas regiões. Esta hipótese é sustentada pelo facto das observações do Havai, por oposição às do Brasil, serem oriundas de regiões próximas do mar (ver Subsecção 2.3.1), adicionalmente, o Brasil apresenta uma grande variabilidade em $^{87}Sr/^{86}Sr$, tendo valores elevados nesta razão isotópica e o Havai os valores mais reduzidos.

2.^a componente principal: Contraste entre *C*%, $\delta^{18}O$, $\delta^{34}S$, $\delta^{18}O_{caf}$ e $\delta^{13}C$, $\delta^{13}C_{caf}$.

A segunda componente principal reflete o efeito continentalidade e clima em contraste com a atividade fotossintética da planta e sua regulação (afetada por fatores climáticos como a temperatura e a humidade relativa do ar, que por sua vez influenciam as taxas de transpiração e fotossíntese das plantas, como refere Rodrigues, 2011, pp. 21–22). *Scores* reduzidos nesta componente principal sugerem que as respetivas observações são oriundas de zonas mais próximas do mar que as restantes, enquanto que *scores* elevados podem indicar que os grãos se desenvolveram em períodos quentes e de maior atividade fotossintética por parte das plantas.

Nos *scores* da segunda componente principal (Figura 3.8a) verifica-se uma divisão das observações do arquipélago do Havai e ainda uma razoável divisão entre as observações do Brasil e do Quénia. A divisão do arquipélago do Havai revela as grandes diferenças de altitude e distância ao mar existentes entre as suas ilhas. O grupo de observações com *scores* reduzidos nesta componente principal é de

zonas mais próximas do mar, enquanto que o grupo com *scores* elevados é de zonas a altitudes mais elevadas (mais distantes do mar, originárias da ilha Havai). Nesta componente principal a divisão do Quênia (*scores* mais reduzidos) e do Brasil (*scores* mais elevados) é um indicador de práticas agrícolas e climas distintos, nomeadamente, pode indicar maior utilização de fertilizantes baseados em enxofre no Quênia, bem como clima seco no período de desenvolvimento do grão de café.

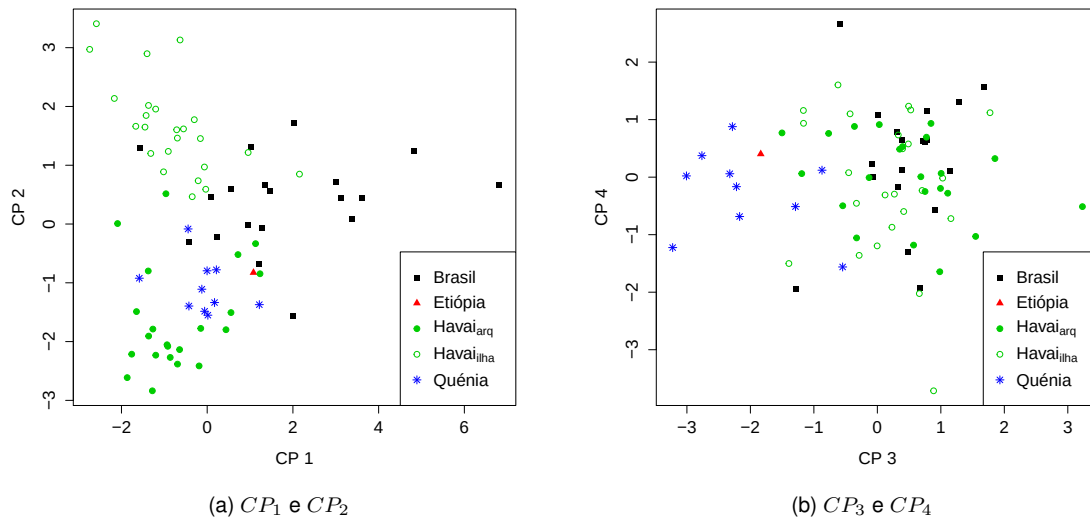


Figura 3.8: *Scores* das componentes principais estimadas da ACP clássica aplicada ao *set3* (dados padronizados).

3.^a componente principal

Pode ser interpretada como uma média ponderada entre as composições isotópicas do carbono, oxigénio e azoto ($\delta^{13}C$, $\delta^{13}C_{caf}$, $\delta^{18}O$, $\delta^{18}O_{caf}$, $\delta^{15}N_{caf}$ e $\delta^{15}N$). Observações com *scores* reduzidos estão enriquecidas no isótopo pesado de carbono, oxigénio e azoto, potencialmente originárias de zonas de clima seco ou solos naturalmente enriquecidos.

No que diz respeito aos *scores* desta componente principal (Figura 3.8b), é de salientar os países africanos (Quênia e Etiópia) que apresentam *scores* reduzidos e portanto o grão verde dos cafeeiros está enriquecido em isótopos pesados.

4.^a componente principal

Reflete as percentagens de carbono e azoto. Observações com *scores* elevados nesta componente principal terão elevadas percentagens deste elementos, nomeadamente de carbono, e vice-versa.

Nos *scores* da quarta componente principal (Figura 3.8b), não há nada a salientar do ponto de vista geográfico. Como esta componente reflete as percentagens de carbono e azoto antevê-se que estas duas variáveis não serão tão importantes como as restantes no que diz respeito à diferenciação.

ACP via *Projection Pursuit* aplicada ao *set9*

Esta secção dedica-se à análise em componentes principais utilizando as observações completas e padronizadas do *set9*. Para a estimação das componentes principais foi utilizado o método *Projection Pursuit* (abreviadamente PP).

Apesar do *scree plot* (Figura 3.9) e de $\hat{\lambda}_3 < \bar{\lambda} = 1$ sugerirem a escolha de apenas duas componentes principais, optou-se por reter três de quatro componentes, obtendo-se assim 90% da variabilidade dos

dados explicada (ver Tabela 3.8 e Figura 3.9). Para esta escolha foi também tida em conta a relevância que a terceira componente principal mostrou ter na diferenciação geográfica.

Tabela 3.8: Variância estimada e proporção cumulativa de variabilidade total explicada pelas componentes principais da ACP PP aplicada ao *set₉* (dados padronizados).

	CP_1	CP_2	CP_3	CP_4
Variância	1.77	1.05	0.78	0.40
Prop. cumulativa	0.44	0.71	0.90	1.00

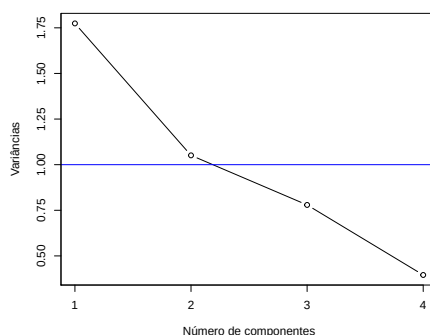


Figura 3.9: Scree plot da ACP PP aplicada ao *set₉* (dados padronizados).

Os *loadings* estimados e as correlações amostrais entre as componentes principais retidas e as variáveis, analisados em seguida, encontram-se nas tabelas 3.9a e 3.9b, respetivamente. Tal como aconteceu na ACP anterior, as correlações amostrais levam a interpretações semelhantes às dos *loadings*. Na Figura 3.10 estão representados os respetivos *scores*, com identificação da origem geográfica de cada observação e distinção entre as observações da ilha Havai (Havai_{ilha}) e as observações das restantes ilhas do arquipélago do Havai (Havai_{arq}).

Tabela 3.9: *Loadings* e correlações amostrais entre as CPs e as variáveis, correspondentes à ACP PP aplicada ao *set₉* (dados padronizados).

(a) <i>Loadings</i>				(b) Correlações amostrais entre as CPs e as variáveis.			
	CP_1	CP_2	CP_3		CP_1	CP_2	CP_3
$^{87}Sr/^{86}Sr$	0.591	-0.001	-0.064	$^{87}Sr/^{86}Sr$	0.787	-0.001	-0.056
$\delta^{13}C_{caf}$	-0.385	0.784	0.374	$\delta^{13}C_{caf}$	-0.512	0.804	0.330
$\delta^{18}O_{caf}$	-0.407	-0.615	0.581	$\delta^{18}O_{caf}$	-0.543	-0.630	0.513
$\delta^{15}N_{caf}$	0.581	0.089	0.720	$\delta^{15}N_{caf}$	0.774	0.092	0.636
$Cor(CP_i, CP_j^{tr})$	1	0.997	0.999				

1.ª componente principal: Contraste entre $\delta^{13}C_{caf}$, $\delta^{18}O_{caf}$ e $^{87}Sr/^{86}Sr$, $\delta^{15}N_{caf}$.

A primeira componente principal reflete o tipo de solo em contraste com o metabolismo da planta e o efeito continentalidade. *Scores* elevados nesta componente principal sugerem que as respetivas observações são originárias de zonas do interior. Enquanto que *scores* reduzidos indica que as observações são originárias de zonas mais próximas do mar e de menor altitude.

No que diz respeito aos *scores*, verifica-se que os das observações do Brasil são, na sua maioria, elevados, enquanto que os do Havai são reduzidos, permitindo deduzir que o tipo de solo das regiões do Brasil é distinto do tipo de solo do Havai, bem como do Quênia. Os *scores* reduzidos das observações do arquipélago do Havai revelam também a menor distância ao mar que estas têm em relação às restantes.

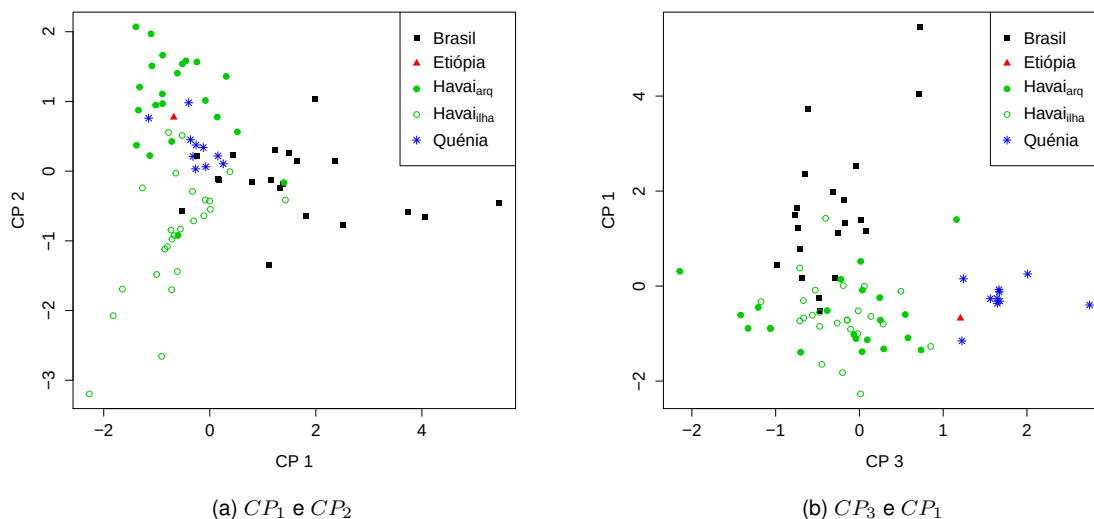


Figura 3.10: Scores das componentes principais estimadas da ACP PP aplicada ao *set9* (dados padronizados).

2.^a componente principal: Contraste entre $\delta^{18}O_{caf}$ e $\delta^{13}C_{caf}$.

Como a composição isotópica do oxigénio na molécula da cafeína do grão verde de café reflete a composição isotópica da água da chuva (Rodrigues, 2011, p. 23), esta componente principal revela o efeito de altitude e de continentalidade em contraste com o metabolismo primário da planta. Scores reduzidos nesta componente principal levam a crer que as observações provêm de zonas litorais e clima quente, enquanto que scores elevados sugerem origem em zonas de altitude.

Os scores das observações do Havai apresentam uma grande variabilidade na segunda componente principal (Figura 3.10a), nomeadamente scores reduzidos no grupo de observações da ilha Havai e elevados nas restantes ilhas do arquipélago do Havai. Esta variabilidade reflete as variações de altitude existentes neste arquipélago, sendo na ilha Havai que se encontram as observações de maior altitude. Mais ainda, estas alterações refletem-se no clima, que por sua vez afeta o metabolismo primário da planta, refletindo-se na composição isotópica do carbono.

3.^a componente principal

Pode ser interpretada como uma média ponderada dos valores da composição isotópica do azoto, oxigénio e carbono na molécula de cafeína do grão verde ($\delta^{15}N_{caf}$, $\delta^{18}O_{caf}$ e $\delta^{13}C_{caf}$), sendo de destacar $\delta^{15}N_{caf}$ com o *loading* mais elevado. Observações com scores elevados nesta componente principal estão enriquecidas em isótopos pesados destes elementos químicos, e *vice-versa*, podendo-se, com esta componente, evidenciar zonas de solos enriquecidos ou sujeitos as práticas agrícolas distintas.

Relativamente aos scores, a terceira componente principal (Figura 3.10b) destaca-se pela clara separação do Quênia e da Etiópia das restantes regiões, tendo estes países scores elevados. Ou seja, as observações destes dois países, quando comparadas com as das restantes regiões, estão enriquecidas em isótopos pesados de oxigénio, azoto e carbono na molécula de cafeína.

ACP via *Projection Pursuit* aplicada ao set_{10}

A ACP agora considerada tira partido das observações completas e padronizadas do set_{10} e recorre ao método *Projection Pursuit* (PCA Proj) para estimação das componentes principais, sendo que o estimador de dispersão escolhido, para maximizar as suas variâncias, foi o desvio médio absoluto (MAD), estimador robusto do desvio padrão.

Foram retidas as primeiras três componentes principais, que explicam 88% da variabilidade total dos dados (ver Tabela 3.10). Tal como no caso anterior, apesar de $\hat{\lambda}_3 < \bar{\lambda} = 1$ e duas componentes principais explicarem 72% da variabilidade dos dados, a regra do cotovelo sugere reter três componentes (Figura 3.11) e a terceira componente principal veio a revelar-se importante do ponto de vista da diferenciação geográfica.

Tabela 3.10: Variância estimada e proporção cumulativa de variabilidade total explicada pelas componentes principais da ACP PP aplicada ao set_{10} (dados padronizados).

	CP_1	CP_2	CP_3	CP_4	CP_5
Variância	1.96	1.71	0.79	0.41	0.22
Prop. cumulativa	0.39	0.72	0.88	0.96	1.00

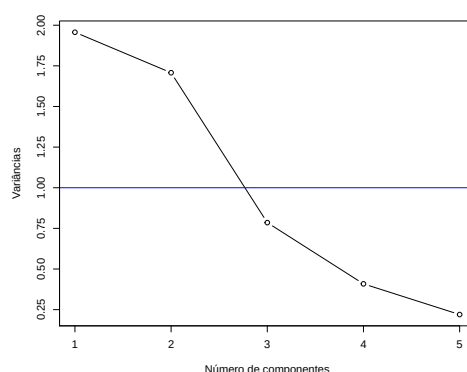


Figura 3.11: *Scree plot* da ACP PP aplicada ao set_{10} (dados padronizados).

Seguidamente analisam-se os *loadings* estimados, assim como os *scores* da ACP, apresentados na Tabela 3.11a e Figura 3.12, respetivamente. Nos gráficos dos *scores* além da identificação da usual origem geográfica das observações, recorre-se também à distinção entre as observações da ilha Havai (Havai_{ilha}) e as observações das restantes ilhas do arquipélago do Havai (Havai_{arq}).

Tabela 3.11: *Loadings* e correlações amostrais entre as CPs e as variáveis, correspondentes à ACP PP aplicada ao set_{10} (dados padronizados).

(a) <i>Loadings</i>				(b) Correlações amostrais entre as CPs e as variáveis.			
	CP_1	CP_2	CP_3		CP_1	CP_2	CP_3
$\delta^{34}S$	0.694	-0.447	-0.211	$\delta^{34}S$	0.970	-0.584	-0.187
$^{87}Sr/^{86}Sr$	-0.293	-0.435	0.196	$^{87}Sr/^{86}Sr$	-0.409	-0.568	0.174
$\delta^{13}C_{caf}$	0.054	0.749	-0.044	$\delta^{13}C_{caf}$	0.076	0.979	-0.039
$\delta^{18}O_{caf}$	0.639	0.201	0.508	$\delta^{18}O_{caf}$	0.894	0.262	0.451
$\delta^{15}N_{caf}$	-0.147	-0.097	0.810	$\delta^{15}N_{caf}$	-0.205	-0.127	0.718
$\hat{Cor}(CP_i, CP_j^{tr})$	0.986	0.976	0.937				

1.ª componente principal: Contraste entre $^{87}Sr/^{86}Sr$ e $\delta^{34}S$, $\delta^{18}O_{caf}$.

Os *loadings* desta componente principal podem ser vistos como um contraste entre o tipo de solo e a altitude e distância ao mar. Observações com *scores* elevados nesta componente principal

podem indicar proximidade ao litoral, enquanto que observações com *scores* reduzidos indiciam que as respectivas observações são provenientes de zonas de altitude/interior.

Os *scores* da primeira componente principal (Figura 3.12a) revelam uma grande variabilidade nas regiões do Havai. Esta variabilidade pode estar relacionada com a ocorrência de atividade vulcânica e variação de altitude a que o grão de café se desenvolve, que influencia os valores de $\delta^{34}S$. As observações do Brasil e da ilha Havai são as que apresentam os *scores* mais reduzidos, sendo as que apresentam altitudes médias (entre todas as observações). Entre as observações do arquipélago do Havai, são as observações da ilha Havai que apresentam distância ao mar mais elevada, apresentando *scores* reduzidos. Por oposição, as observações das restantes regiões do arquipélago do Havai são as que apresentam *scores* mais elevados nesta componente principal, sendo também as que estão mais próximas do mar.

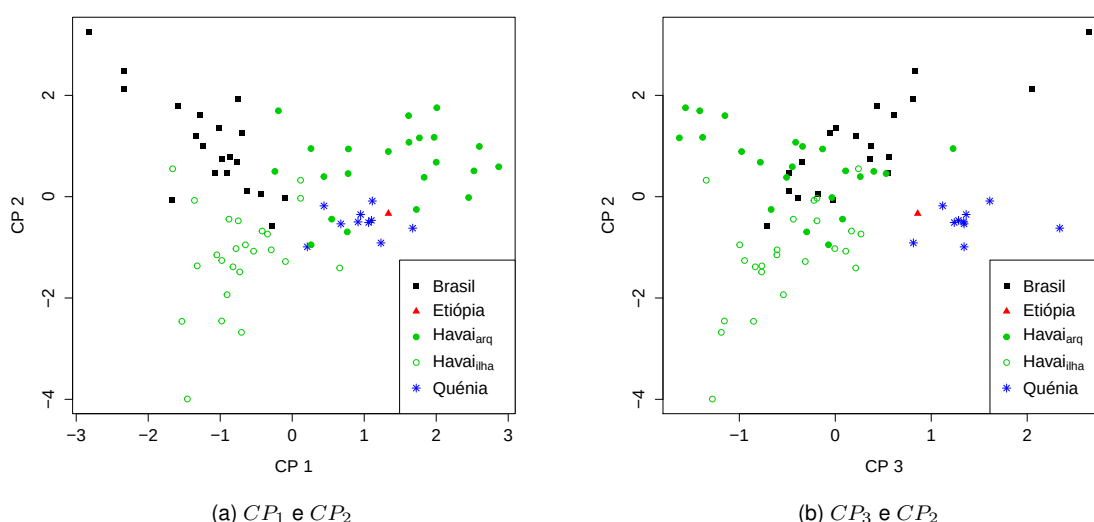


Figura 3.12: *Scores* das componentes principais estimadas da ACP PP aplicada ao *set*₁₀ (dados padronizados).

2.^a componente principal: Contraste entre $\delta^{34}S$, $^{87}Sr/^{86}Sr$ e $\delta^{13}C_{caf}$.

Esta componente principal reflete o metabolismo primário da planta face ao tipo de solo.

Na Figura 3.12 verifica-se que os *scores* da segunda componente principal das observações da ilha Havai e do Quênia (zonas montanhosas) são maioritariamente negativos, enquanto que os *scores* das observações do Brasil e das restantes ilhas do arquipélago do Havai são, de um modo geral, positivos, revelando uma separação entre estes dois grupos de observações.

3.^a componente principal: Média ponderada entre $\delta^{15}N_{caf}$ e $\delta^{18}O_{caf}$.

Esta componente principal pode ser interpretada como uma ponderação entre as práticas agrícolas, o fator continentalidade e o clima. *Scores* elevados nesta componente principal podem indicar que as respectivas observações são provenientes de regiões de reduzida altitude ou clima seco. Por oposição, observações com *scores* reduzidos podem indicar práticas agrícolas mais sustentáveis e clima húmido.

No gráfico dos *scores* da terceira componente principal (Figura 3.12b) verifica-se uma separação quase total dos países africanos (Quênia e Etiópia) das restantes regiões, verificando-se uma clara separação ao cruzar-se os *scores* desta componente principal com os de uma das outras componentes retidas. As observações do Quênia e da Etiópia apresentam *scores* elevados nesta componente

principal, podendo concluir-se que nestas regiões a molécula de cafeína do grão está mais enriquecida em isótopos pesados de oxigénio e azoto que nas restantes, por oposição ao Havai que apresenta os *scores* mais reduzidos.

Concluindo, a análise dos gráficos dos *scores* da Figura 3.12 sugere um padrão interessante que se assemelha à mistura de distribuições normais: entre o conjunto de observações do Havai e do Quénia e as observações do Brasil, na Figura 3.12a; e entre o conjunto de observações da ilha Havai e do Brasil e o conjunto de observações do Quénia e das restantes ilhas do arquipélago do Havai, na Figura 3.12b.

Comentários finais

As ACPs realizadas anteriormente sugerem que o tipo de solo do Brasil é distinto do solo das restantes regiões. Esta evidência pode ser sustentada pelo facto de as observações originárias do Quénia e do Havai provirem de regiões vulcânicas. No caso do Quénia, a maioria é originária do Monte Quénia, um vulcão extinto. No caso do Havai, este arquipélago é conhecido pela intensa atividade vulcânica devida tanto a vulcões ativos como extintos.

As ACPs revelam também grande variabilidade entre as regiões produtoras do Havai que se deve às variações de altitude existentes neste arquipélago e à atividade vulcânica. Mais ainda, sugerem maior enriquecimento em isótopos pesados de oxigénio, carbono e azoto do grão verde de café dos países africanos em relação aos restantes, podendo este fenómeno estar relacionado com maior evapotranspiração dos cafeeiros devido ao clima destas regiões (Rodrigues, 2011, p. 30).

O padrão encontrado nos *scores* das componentes principais do *set*₁₀, leva a considerar que faria sentido realizar uma análise de agrupamentos baseada em misturas de normais, de forma a captar este tipo de comportamentos que as técnicas tradicionais não permitem captar. Esta técnica não será abordada neste trabalho, sugerindo-se a leitura de Bishop (2007) para mais detalhe acerca deste assunto.

3.4 Detecção de *outliers*

3.4.1 Introdução

Quando são aplicados métodos estatísticos clássicos a um conjunto de dados, os seus resultados podem ser afetados pela presença de *outliers*. No entanto, os *outliers* podem representar observações de interesse e a sua identificação ser o objetivo do investigador (Filzmoser *et al.*, 2008).

Os métodos de detecção de *outliers* com base em distâncias propõem-se encontrar observações atípicas presentes nos dados, calculando uma medida que indica quão longe um ponto em particular está do centro dos dados. A medida mais usual é a distância de Mahalanobis robusta, calculada diretamente sobre as observações ou sobre os *scores* das componentes principais (Filzmoser *et al.*, 2008). Esta distância tem a vantagem de ter em conta não só a distância ao centro como também a dispersão inerente aos dados (através da matriz de covariâncias).

Dada uma observação $\mathbf{x}_i = (x_1, \dots, x_p)^t$, $i = 1, 2, \dots, n$, a distância de Mahalanobis (DM) é dada por

$$DM_i = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}, \quad (3.7)$$

onde $\boldsymbol{\mu}$ é a medida de localização e Σ a matriz de covariâncias. Para que a distância de Mahalanobis assuma a sua versão robusta e as medidas de detecção de *outliers* sejam mais fiável, os estimadores destas medidas devem ser robustos (Filzmoser *et al.*, 2008). Para dados com distribuição normal multivariada, o quadrado da distância de Mahalanobis, DM^2 , segue aproximadamente uma distribuição de qui-quadrado com p graus de liberdade, χ_p^2 (relembre-se que p denota o número de variáveis). Desta forma, o valor de corte para classificar uma observação como *outlier* pode ser definido como a raiz quadrada do quantil α da χ_p^2 , $\sqrt{\chi_{(p;\alpha)}^2}$ (Pascoal *et al.*, 2012).

A análise em componentes principais é um método que pode ser utilizado em detecção de *outliers*. As componentes principais têm a sua variância maximizada ao longo de cada componente principal (com restrição de ortogonalidade). Visto que os *outliers* aumentam a variância ao longo das respetivas direções, espera-se que estes sejam mais facilmente detetados no espaço inerente às componentes principais que no do conjunto de dados original (Filzmoser *et al.*, 2008).

Existem duas distâncias importantes no processo de detecção de *outliers* baseada em ACP robusta. Definindo o *score* da i -ésima observação na j -ésima componente principal por $z_{ij} = \hat{\gamma}_j^t (\mathbf{x}_i - \hat{\boldsymbol{\mu}})$, onde $\hat{\boldsymbol{\mu}}$ e $\hat{\gamma}_j$ denotam a estimativa da média e dos *loadings*, respetivamente. Estas distâncias definem-se do seguinte modo:

$$\textbf{Distância score: } DS(\mathbf{x}_i) = \left(\sum_{j=1}^k \frac{z_{ij}^2}{\hat{\lambda}_j} \right)^{\frac{1}{2}} \quad (3.8)$$

$$\textbf{Distância ortogonal: } DO(\mathbf{x}_i) = \|(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) - \mathbf{P}_{p,k} \mathbf{z}_i\|, \quad (3.9)$$

onde $\mathbf{P}_{p,k}$ é a matriz ($p \times k$) dos *loadings* das primeiras k componentes principais robustas retidas e $\mathbf{z}_i = (z_{i1}, \dots, z_{ik})^t$ os *scores* da i -ésima observação. A distância *score* corresponde à distância de Mahalanobis no espaço gerado pelas componentes principais, enquanto que a distância ortogonal corresponde à distância de uma observação ao espaço gerado pela ACP (Pascoal *et al.*, 2012).

Para distinguir entre observações regulares e *outliers* é necessário definir valores de corte para as duas distâncias. Os valores de corte da distância *score*, c_{DS} , e da distância ortogonal, c_{DO} , sugeridos por Pascoal *et al.* (2012) são definidos por:

$$c_{DS} = \sqrt{\chi_{(k;p_s)}^2} \quad (3.10)$$

$$c_{DO} = (\hat{\mu}_o + z_{p_o} \hat{\sigma}_o)^{\frac{3}{2}}, \quad (3.11)$$

onde $\chi_{(k;p_s)}^2$ denota o quantil de ordem p_s da distribuição do qui-quadrado com k graus de liberdade, sendo k o número de componentes principais retidas, ao passo que z_{p_o} denota o quantil de ordem p_o da distribuição normal multivariada, $\hat{\mu}_o$ e $\hat{\sigma}_o$ são, respetivamente, as estimativas de localização e escala das distâncias ortogonais. Os estimadores utilizados foram a mediana e o MAD (*Median Absolute*

Deviation) (Mosteller e Tukey, 1977); as ordens dos quantis foram $p_s = 0.999$ e $p_o = 0.975$. Uma observação com distância ortogonal ou *score* maior que o respetivo valor de corte é classificada como *outlier*.

Neste trabalho, a classificação das observações como regulares ou *outliers* foi feita individualmente. Ou seja, para classificar a i -ésima observação, esta foi retirada da amostra e foram estimadas as componentes principais e o valor de corte associado à distância ortogonal. Repetindo-se este processo para cada observação. Esta estratégia, conhecida como *leave-one-out*, garante-nos que o valor de corte não é enviesado pela presença da observação que estamos a classificar. No entanto, leva a que o valor de corte associado à distância ortogonal seja diferente de observação para observação. Tal já não acontece com o valor de corte da distância *score*, pois este apenas depende do número de componentes principais retidas, que não varia.

3.4.2 Resultados da aplicação dos métodos de deteção de *outliers* ao conjunto de dados

A deteção de *outliers* foi aplicada a todos os subconjuntos considerados na Secção 3.1. Para análise apenas foram considerados quatro destes nove subconjuntos, que correspondem aos mais relevantes e interessantes do ponto de vista do problema real, a saber *set*₃, *set*₇, *set*₉ e *set*₁₀.

Numa primeira análise recorreu-se ao gráfico dos quantis da χ_p^2 em função do quadrado da distância de Mahalanobis robusta (DM^2) calculada para o conjunto de dados original, recorrendo ao estimador MCD. Estes gráficos são apresentados na Figura 3.13. A cinza representa-se a reta de corte que corresponde ao valor de corte $\chi_{(p;0.975)}^2$. Em qualquer dos gráficos, verifica-se que as observações com maior distância dizem respeito ao Brasil. No *set*₇ destaca-se claramente um grupo de observações referentes ao Brasil, contudo nem todas são consideradas *outliers* ao considerar o valor de corte $\chi_{(3;0.975)}^2 = 9.348$ (vide Figura 3.13b); a maioria das observações deste grupo não está presente nos restantes subconjuntos (por apresentarem valores omissos). No *set*₃, *set*₉ e *set*₁₀ (figuras 3.13a, 3.13c e 3.13d, respetivamente) as observações que são consideradas *outliers* têm ID 42 e ID 48, admitindo valor de corte $\chi_{(p;0.975)}^2$.

Esta breve análise sugere a existência de *outliers* no conjunto de dados. Assim, recorreu-se a deteção de *outliers* com base agora em componentes principais robustas. A classificação das observações em *outliers* ou regulares foi realizada recorrendo à estratégia *leave-one-out*, de forma a reduzir o enviesamento dos resultados.

O método utilizado para as ACPs foi o de *Projection Pursuit* recorrendo à função *PcaProj* do pacote *rrcov* do R. O estimador de escala utilizado para detetar a direção de máxima variância foi o MAD e as observações foram padronizadas (as estimativas de localização e escala foram obtidas a partir do estimador MCD). Para o cálculo das distâncias e dos valores de corte foi adaptada a versão dos algoritmos do pacote *rrcov* de Armstrong (2013), por forma a contemplar a possibilidade de utilizar quantis com ordens diferentes de 0.975 para cálculo dos valores de corte. As observações da Etiópia apenas foram utilizadas para efeitos de classificação, não sendo incluídas na estimação das componentes principais.

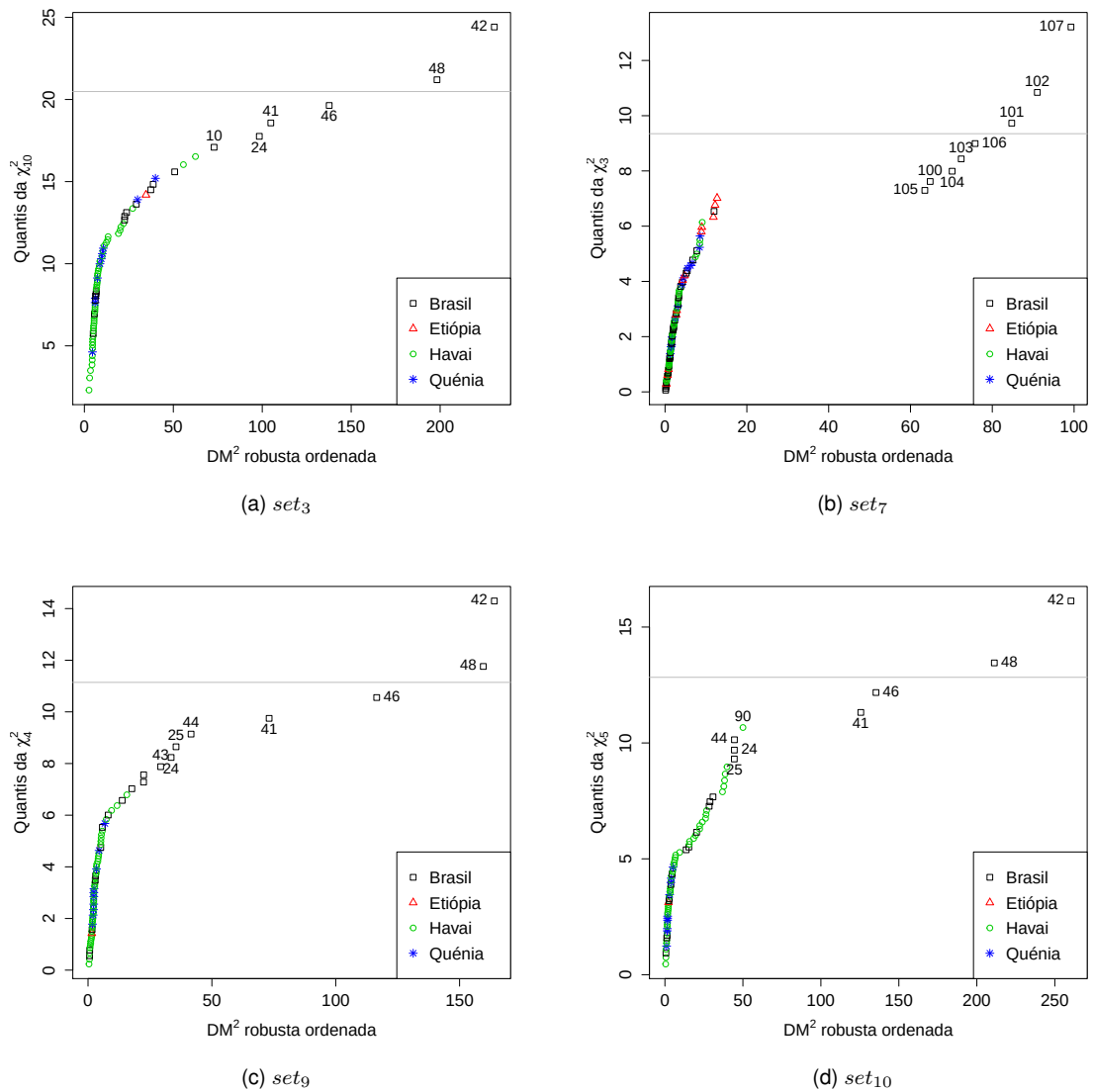


Figura 3.13: Gráficos da distribuição de χ^2_p em função do quadrado da distância de Mahalanobis robusta, nos subconjuntos set_3 , set_7 , set_9 e set_{10} , com identificação da origem geográfica.

As estratégias utilizadas para a escolha do número de componentes principais a reter coincide com as estratégias utilizadas na Secção 3.3. A Tabela 3.12 resume estas escolhas.

Tabela 3.12: Número de CPs retidas para cada conjunto de variáveis e respetivas proporções de variação total explicada.

Conjunto	Método	Total de CPs	CPs retidas	Prop. exp.
set_3	ACP PP	10	4	0.76
set_7	ACP PP	3	2	0.73
set_9	ACP PP	4	2	0.68
set_{10}	ACP PP	5	2	0.71

Na Tabela 3.13 apresentam-se as percentagens de *outliers* detetados em cada subconjunto considerado, por país e global. Verifica-se que, para os subconjuntos de dados considerados para análise, nenhuma observação proveniente da Etiópia foi classificada como *outlier*; no entanto, seria de relembrar o reduzido número de observações da Etiópia (ver Tabela 3.2) e que, por exemplo, do set_3 , set_9 e set_{10} apenas faz parte uma observação desta origem geográfica. No que respeita ao Brasil, registam-se

observações classificadas como *outliers* em todos os quatro subconjuntos de variáveis. De destacar o *set*₉ pela elevada percentagem de observações do Brasil classificadas como *outliers* (10 em 20 observações). À exceção do Brasil, apenas se regista um *outlier* no *set*₉ entre as observações do Quênia.

Tabela 3.13: Percentagens de *outliers* por país e global.

	Brasil	Etiópia	Havai	Quênia	Global
<i>set</i> ₃	0.26	0.00	0.00	0.00	0.07
<i>set</i> ₇	0.19	0.00	0.00	0.00	0.07
<i>set</i> ₉	0.50	0.00	0.00	0.10	0.14
<i>set</i> ₁₀	0.25	0.00	0.00	0.00	0.06

De seguida são analisados mais em detalhe os *outliers* detetados em cada subconjunto. Para esta análise é tida em conta a localização geográfica das observações classificadas como *outliers*, cujos mapas com identificação da localização geográfica se encontram na Secção A.2, com identificação dos *outliers* a vermelho e respetivo ID. Após padronização de cada subconjunto de dados, através do estimador robusto MCD, e classificação dos *outliers*, foram selecionadas as variáveis que mais facilmente os distinguem visualmente. Apesar de não se utilizar uma notação diferente por uma questão de simplificação, na análise de resultados desta secção assume-se que as variáveis estão padronizadas.

Para análise dos resultados recorreu-se ainda ao gráfico de diagnóstico, onde se apresentam as distâncias *scores* e distâncias ortogonais obtidas para cada observação em cada subconjunto. Nestes gráficos será representado por uma reta a vermelho o valor de corte para a distância *score*, que corresponde à raiz quadrada do quantil de ordem 0.999 da $\chi^2_{(k)}$. O valor de corte da distância ortogonal é estimado após a estimação das componentes principais e distâncias ortogonais. Tal como se refere na Subsecção 3.4.1, para cada observação a classificar, esta estimação é efetuada por recurso à abordagem *leave-one-out*. Assim, o valor de corte é diferente para cada observação, não permitindo a sua representação nos gráficos de diagnóstico.

Deteção de *outliers* aplicada ao *set*₃

A deteção de *outliers* no *set*₃ classificou 7% das observações como *outliers*, que corresponde a 5 de 76 observações no total, com IDs 24, 41, 42, 46 e ID 48, sendo todas originários do Brasil. A localização geográfica dos *outliers* apresenta-se, a vermelho, nos mapas da Figura A.1.

Na Figura 3.14 apresentam-se os diagramas de dispersão de algumas das variáveis padronizadas do *set*₃. No gráfico da Figura 3.14a verifica-se que os *outliers* têm valores elevados de $^{87}\text{Sr}/^{86}\text{Sr}$, as observações com ID 41 e ID 42 destacam-se ainda por também apresentarem $\delta^{15}\text{N}_{\text{caf}}$ elevado. Mais ainda, na Figura 3.14b estas observações evidenciam-se pelo valor reduzido de $\delta^{18}\text{O}$, especialmente da observação com ID 42, e valor reduzido em $\%N$ da observação com ID 41.

Para finalizar, na Figura 3.15 representam-se as duas distâncias que nos permitiram classificar as observações como *outliers*. Neste caso, a classificação como *outliers* deveu-se à elevada distância *score* que apresentam, não se identificando nenhuma observação com distância ortogonal superior ao respetivo valor de corte.

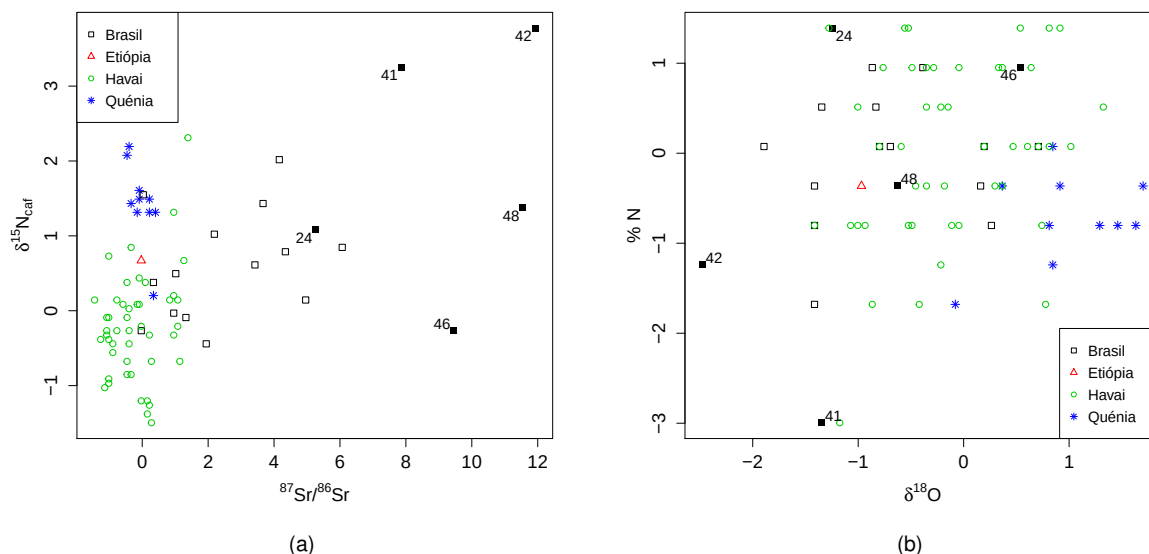


Figura 3.14: Diagramas de dispersão de algumas variáveis padronizadas do *set3*. A cheio e com respectivo identificador representam-se as observações classificadas como *outliers*.

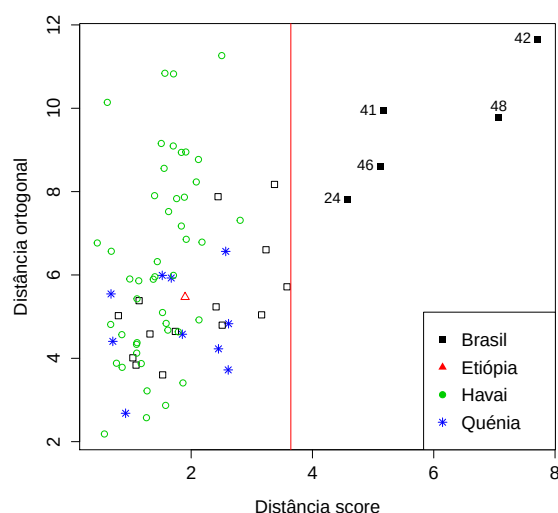


Figura 3.15: Gráfico de diagnóstico da detecção de *outliers* aplicada ao *set3*. A cheio e com o respectivo identificador representam-se as observações classificadas como *outliers*.

Deteção de *outliers* aplicada ao *set7*

A deteção de *outliers* aplicada ao *set7* (subconjunto com o maior número de observações, 119) classificou como *outlier* 8 observações do Brasil, claramente diferenciadas das restantes (observações com ID a partir de 100, que não fazem parte dos restantes subconjuntos aqui analisados). Este facto é claramente evidenciado pela Figura 3.17, onde se constata que estas observações registam valores bastante elevados da distância *score* e distância ortogonal. Este facto é a tradução do padrão evidenciado na Figura 3.16 onde se registam valores elevados de $\delta^{18}O$ para estas observações.

Apesar de não se dispor de coordenadas geográficas para estas observações, sabe-se que são de regiões do estado de São Paulo e Sul de Minas Gerais. Mais ainda, a amostra é da colheita de

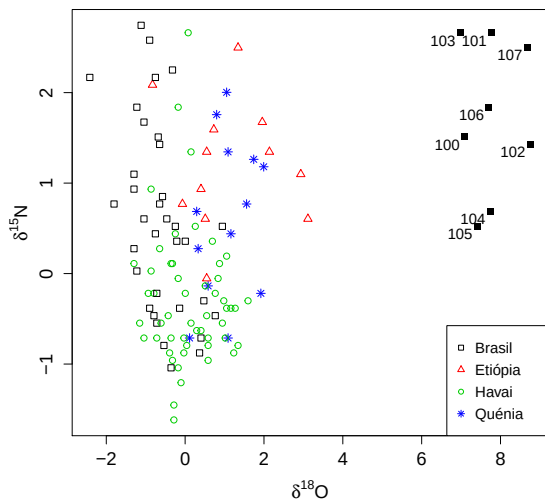


Figura 3.16: Diagrama de dispersão das variáveis $\delta^{18}O$ e $\delta^{15}N$ padronizadas do set_7 . A cheio e com respectivo identificador representam-se as observações classificadas como *outliers*.

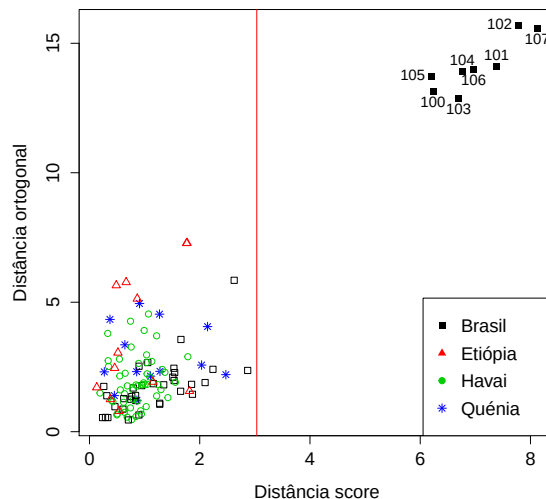


Figura 3.17: Gráfico de diagnóstico da detecção de *outliers* aplicada ao set_7 . A cheio e com o respectivo identificador representam-se as observações classificadas como *outliers*.

2004, sendo estas as únicas observações desse ano. Os valores elevados da composição isotópica do oxigénio poderão dever-se à onda de calor que atingiu as fazendas de São Paulo em 2004 (Marengo *et al.*, 2010).

Na Figura A.2 estão identificadas as origens geográficas das observações do set_7 , das quais se dispõe de coordenadas geográficas. Por não se dispor das coordenadas geográficas das observações classificadas como *outliers* este gráfico é apenas ilustrativo da localização das observações existentes neste subconjunto (que são em maior número que nos restantes subconjuntos).

Deteção de *outliers* aplicada ao set_9

A análise do set_9 conduziu à identificação de metade das observações do Brasil como *outliers* (10 em 20 no total) e uma observação do Quênia (em dez). As observações do Brasil têm IDs 24, 25, 35, 41, 42, 43, 44, 46, 47 e ID 48; a observação do Quênia tem ID 33. As suas localizações geográficas podem ser consultadas na Figura A.3.

Através da análise dos diagramas de dispersão da Figura 3.18, verifica-se que as observações do Brasil classificadas como *outliers* se caracterizam por apresentarem valores elevados de $^{87}Sr/^{86}Sr$. Destacam-se as observações com ID 41 e ID 42 pelos valores elevados de $\delta^{15}N_{caf}$ bem patentes na Figura 3.18a. Alguns destes *outliers* apresentam também $\delta^{18}O_{caf}$ reduzido. Por fim, a observação do Quênia classificada como *outlier* apresenta $\delta^{18}O_{caf}$ elevado (*vide* Figura 3.18b).

As observações do Brasil com IDs 41, 42, 46 e ID 48 são as que apresentam os valores mais elevados na distância ortogonal e *score*, como se constata na Figura 3.19. A observação com ID 33, do Quênia, foi classificada como *outlier* por apresentar distância ortogonal superior ao respetivo valor de corte. Este tipo de observações, com distância ortogonal acima do valor de corte e distância *score* abaixo do respetivo valor de corte, são chamadas de *outliers* ortogonais. Estes *outliers* não se distinguem das observações regulares quando projetados no espaço gerado pelas componentes

principais, no entanto têm elevada distância ortogonal a este espaço (Gharibnezhad *et al.*, 2015).

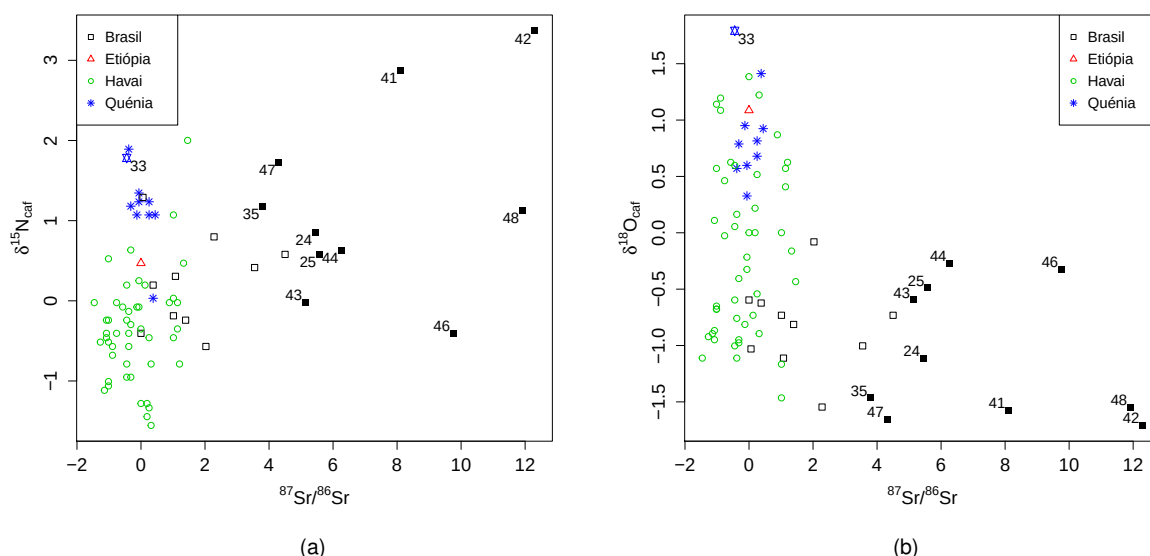


Figura 3.18: Diagramas de dispersão das variáveis padronizadas selecionadas do *set9*. A cheio e com respectivo identificador representam-se as observações classificadas como *outliers*.

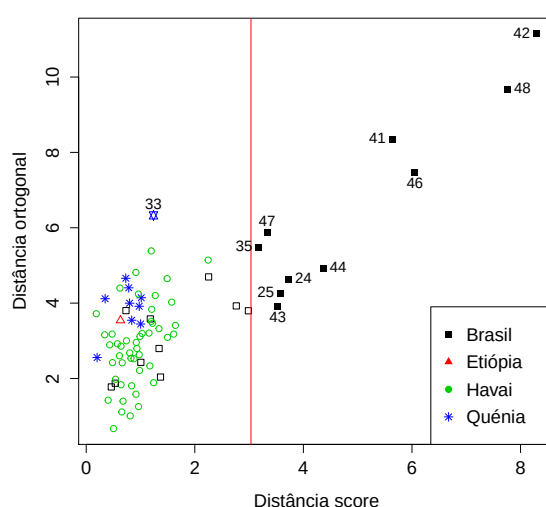


Figura 3.19: Gráfico de diagnóstico da deteção de *outliers* aplicada ao *set9*. A cheio e com o respectivo identificador representam-se as observações classificadas como *outliers*.

Deteção de *outliers* aplicada ao *set10*

A deteção de *outliers* aplicada ao subconjunto *set10* levou a classificar 5 observações como *outliers* (num total de 77), todas originárias do Brasil. A localização mais pormenorizada destas observações pode ser consultada na Figura A.4.

Na Figura 3.20 apresentam-se os diagramas de dispersão de algumas das variáveis padronizadas do *set10*. À semelhança dos casos anteriores, os *outliers* classificados no *set10* apresentam valores

elevados de $^{87}\text{Sr}/^{86}\text{Sr}$. Destaque-se também o valor elevado de $\delta^{15}\text{N}_{\text{caf}}$ das observações com ID 41 e ID 42 (ver Figura 3.20a). Estas observações e a observação com ID 48 apresentam também valor reduzido de $\delta^{18}\text{O}_{\text{caf}}$ (ver Figura 3.20b). Como se verifica na Figura 3.21, estas observações são as que apresentam, simultaneamente, valores elevados nas distâncias *score* e ortogonal.

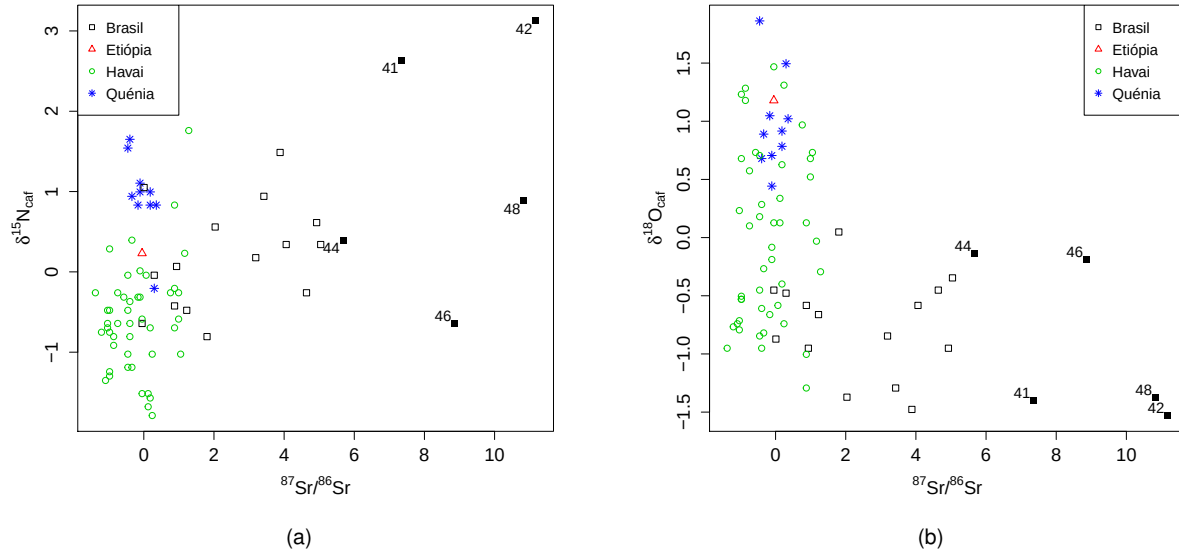


Figura 3.20: Diagramas de dispersão das variáveis padronizadas selecionadas do set_{10} . A cheio e com respetivo identificador representam-se as observações classificadas como *outliers*.

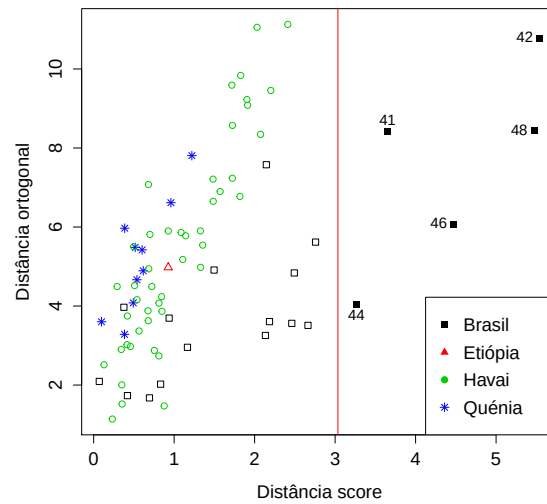


Figura 3.21: Gráfico de diagnóstico da deteção de *outliers* aplicada ao set_{10} . A cheio e com o respetivo identificador representam-se as observações classificadas como *outliers*.

Comentários finais

A deteção de *outliers* baseada em componentes principais foi aplicada a diferentes subconjuntos de observações caracterizadas por distintos grupos de variáveis. Apesar desta multiplicidade de casos, algumas observações foram identificadas como *outliers* em mais de uma análise. Nomeadamente as observações do Brasil com IDs 41, 42, 44, 46 e ID 48 classificadas como *outliers* nos set_3 , set_9 e set_{10} , e

a observação com ID 24 classificada como *outlier* no set_3 e no set_9 . Estas observações caracterizam-se essencialmente por apresentarem valores elevados na razão isotópica do estrôncio.

A existência de um grupo de observações do Brasil (IDs 100, 101, 102, 103, 104, 105, 106 e ID 107) classificadas como *outliers* na análise do set_7 , apenas consideradas neste subconjunto, revelou-se interessante por evidenciar a influência de fatores climáticos na composição isotópica do grão verde de café, nomeadamente do oxigénio. Assim, na mesma origem geográfica, amostras de colheitas diferentes ou localizações com climas diferentes podem revelar-se uma fonte de grande variabilidade na composição isotópica do oxigénio nos dados.

Capítulo 4

Agrupamentos

4.1 Introdução

A análise de agrupamentos, também conhecida por análise de *clusters*, tem como principal objetivo identificar agrupamentos de objetos ou variáveis a partir da informação disponível no conjunto de dados que descreve os objetos e as suas relações (Johnson e Wichern, 2007, p. 671; Tan *et al.*, 2005, p. 487). Pretende-se que os objetos ou variáveis de um determinado grupo sejam semelhantes entre si e diferentes dos objetos ou variáveis dos restantes grupos. A formação destes grupos (ou *clusters*) tem por base medidas de semelhança ou dissemelhança entre os objetos ou variáveis a agrupar. Desta forma, quanto maior a semelhança (ou homogeneidade) dentro de um grupo e quanto maior a diferença entre os grupos, melhor ou mais distintos serão os agrupamentos (Tan *et al.*, 2005, p. 490).

Existem vários métodos de formação de agrupamentos, mas os métodos hierárquicos e os métodos de partição (ou não-hierárquicos) são os mais conhecidos e utilizados. Uma das características destes métodos é a fase em que se escolhe o número de agrupamentos.

No caso dos métodos não-hierárquicos, o número de agrupamentos é fixo à partida ou determinado como parte do processo de agrupamento (Johnson e Wichern, 2007, p. 696). Estes métodos começam de uma partição inicial de objetos em grupos ou de um conjunto de pontos-semente que formarão o núcleo dos agrupamentos, por exemplo *K-Means* e *K-Medoid* com o algoritmo PAM-*Partitioning Around Medoids* (vide Kaufman e Rousseeuw, 1990), respetivamente. Assim, os pontos podem ser escolhidos aleatoriamente entre os objetos ou fornecidos pelo investigador. Este pode utilizar a partição obtida de outro método de formação de agrupamentos, ou construir uma partição de acordo com informação externa sobre os objetos. Os métodos de partição baseiam-se em otimizar uma medida de coesão interna e isolamento externo, fazendo com que os objetos se desloquem entre grupos até que estas deslocações não conduzam a uma melhoria nestas medidas.

Nos métodos hierárquicos o número de agrupamentos é escolhido *a posteriori*. Estes são divididos em 2 tipos: os métodos aglomerativos e os divisivos. Os métodos aglomerativos começam com cada objeto num grupo, existindo tantos grupos como objetos, os objetos mais semelhantes são os primeiros a ser agrupados e assim sucessivamente, até todos os subgrupos formarem apenas um único grupo.

Os métodos divisivos funcionam de forma oposta: começam com todos os objetos num só grupo que é subdividido em dois, de acordo com a maior dissimilaridade entre os objetos, até todos os objetos pertencerem a um grupo distinto, ou seja no final existirão tantos grupos como objetos (Johnson e Wichern, 2007, p. 680). Em geral, se o investigador souber *a priori* o número de grupos que pretende analisar, o algoritmo pode ser interrompido ao atingir-se o número de grupos pretendido. Os métodos hierárquicos têm a vantagem de permitirem construir dendrogramas, que são diagramas que organizam os agrupamentos em forma de árvore, fornecendo informação acerca do seu processo de formação.

Assim como a origem geográfica dos grãos de café verde, existem outras variáveis importantes que distinguem a qualidade do café. Com a análise de agrupamentos pretende-se encontrar relações e semelhanças refletidas nos grupos que possam ser relevantes para o problema, bem como perceber se os dados se organizam naturalmente em função da sua origem geográfica.

4.2 Resultados da aplicação aos dados

Foram aplicados vários métodos de agrupamento baseados em diversas medidas de dissimilaridade aos vários subconjuntos de dados, de forma a encontrar o que melhor se adequava ao problema em estudo. A Tabela 4.1 sumaria as combinações analisadas. Além das dissimilaridades referidas nesta tabela, foi também aplicada a dissimilaridade de Gower; esta tem a vantagem de poder acomodar observações com valores omissos e caracterizadas por variáveis de naturezas distintas (Gower, 1971).

Tabela 4.1: Métodos de agrupamento, dissimilaridades, funções e pacotes do R aplicados nos subconjuntos listados.

Métodos de agrupamento	Dissimilaridades	Função (pacote) do R
Métodos hierárquicos aglomerativos: <i>Ward, Complete Linkage, Average Linkage, Single Linkage e McQuitty</i>	Euclidiana, Manhattan, Minkowski ($r = 3$), Mahalanobis e Supremo	<code>hclust (stats)</code>
Métodos não hierárquicos: K-Medoids (PAM)	Euclidiana, Manhattan e Minkowski ($r = 3$)	<code>pam (cluster)</code>
K-means	Euclidiana	<code>kmeans (stats)</code>
Subconjuntos: <i>set₃, set₄, set₅, set₉, set₁₀</i> (obs. com e sem padronização) e <i>scores</i> da ACP do <i>set₃</i> .		

Após aplicação dos vários métodos de agrupamento foi selecionada uma partição para cada subconjunto de variáveis considerado. A seleção da partição bem como do número de grupos foi efetuada tendo em consideração: o coeficiente aglomerativo; o R^2 (coeficiente de determinação múltiplo) em função do número de grupos; e a *mean average silhouette width* (medida que combina coesão e separação entre os grupos). A validação de cada partição foi realizada recorrendo ao método de adição de erros. Complementarmente, cruzaram-se as diferentes partições de objetos e estas com a origem geográfica dos objetos. Para detalhes sobre estas medidas e métodos de validação de agrupamentos, sugere-se a leitura de Halkidi *et al.* (2001) ou ainda Kaufman e Rousseeuw (1990).

Entre a multiplicidade de partições analisadas, selecionaram-se três que conduziram a resultados muito semelhantes e cujas opções estão sumariadas na Tabela 4.2.

Dadas as semelhanças entre os resultados obtidos para estas três análises, opta-se por apresentar e discutir somente o caso baseado nos *scores* da ACP do *set₃*.

Tabela 4.2: Sumário das partições em grupos selecionadas.

Dados	Padronizados?	Método de agrupamento	Dissemelhança	N.º de grupos retidos
Scores ACP set_3	Sim (*)	Ward	Minkowski ($r = 3$)	4
set_9	Sim	PAM	Euclidiana	4
set_{10}	Sim	PAM	Euclidiana	4

(*) A ACP foi realizada utilizando as variáveis padronizadas.

Agrupamento baseado nos scores da ACP realizada ao set_3

A partição selecionada da análise de agrupamentos baseia-se nos scores da ACP do set_3 , cuja análise foi feita na Subsecção 3.3.2 e se definiu reter 4 componentes principais. Esta partição foi obtida pelo método hierárquico aglomerativo Ward, usando a distância de Minkowski ($r = 3$), e é constituída por 4 grupos. O respetivo coeficiente aglomerativo é 0.97, a *mean average silhouette width* para 4 grupos é de 0.32 (a mais elevada entre 2 a 7 grupos), e o coeficiente de determinação múltiplo igual a $R^2 = 0.57$ (o gráfico do R^2 em função do número de grupos é representado na Figura 4.1). Além destas medidas teve-se em conta que aumentar o número de grupos a reter torna-los-ia de dimensão muito reduzida e sem interesse prático.

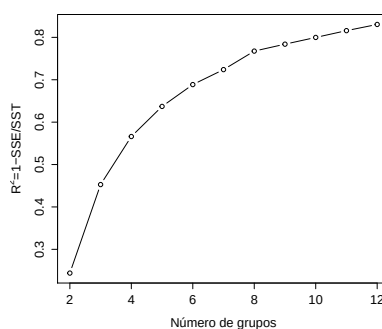


Figura 4.1: R^2 em função do número de grupos e referente à análise de agrupamentos dos scores da ACP do set_3 .

O dendrograma da Figura 4.2 representa o processo de formação dos agrupamentos. Para melhor descrever os agrupamentos, os ramos finais foram coloridos de acordo com a origem geográfica e os índices dos objetos consoante a altitude (em conformidade com a legenda de cores que consta da figura). Em relação à origem geográfica, verifica-se que as regiões ficaram razoavelmente bem separadas. A Tabela 4.1 cruza as origens geográficas com os 4 grupos obtidos. A única observação da Etiópia ficou no grupo do Quénia (G_1) e o Havai ficou separado em dois grupos, um maioritariamente a baixas altitudes (G_2) e o outro a altitudes médias (G_3); esta separação do Havai era esperada, pois a ACP sugeria esta separação na Secção 3.3.

As observações que na Secção 3.4 foram consideradas como *outliers* para o set_3 não correspondem às observações que na análise de agrupamentos foram atribuídas a grupos que não correspondem à sua origem geográfica (utilizar-se-á neste caso o termo *mal alocadas*). Uma justificação para o sucedido é que o foco da deteção de *outliers* é detetar observações com comportamento discordante entre todas as observações do conjunto de dados em estudo, independentemente da origem geográfica. Assim, apesar das observações assinaladas como *outliers* serem as mais díspares do conjunto de dados, estas podem ter graus de semelhança maior com as observações da mesma origem geográfica e como tal virem a ser alocadas no mesmo grupo.

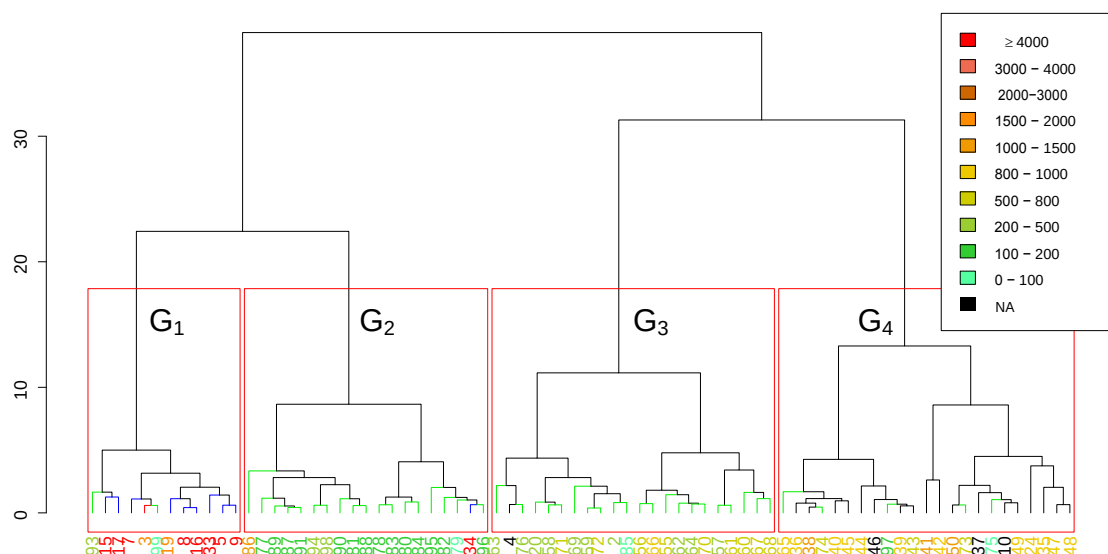


Figura 4.2: Dendrograma dos agrupamentos obtidos dos *scores* da ACP do *set*₃. A cor dos ramos corresponde à origem geográfica (preto - Brasil, **vermelho** - Etiópia, **verde** - Havai, **azul** - Quênia). A cor do índice de cada objeto corresponde ao intervalo de altitude (m) correspondente (ver legenda).

Tabela 4.3: Cruzamento entre as origens geográficas e o agrupamento obtido.

	Brasil	Etiópia	Havai	Quênia
G_1	0	1	2	9
G_2	0	0	18	1
G_3	1	0	21	0
G_4	18	0	5	0

Os mapas com identificação do grupo obtido pela análise de agrupamentos estão representados na Figura A.5; os símbolos em forma de círculo correspondem a observações *mal alocadas*, ao passo que os símbolos em forma de gota correspondem às observações *corretamente alocadas*. As observações do grupo G_2 (a **amarelo**) e do grupo G_3 (a **cinza**) são maioritariamente originárias do arquipélago do Havai. Enquanto que do G_3 fazem parte a maioria das observações da ilha Havai, do G_2 fazem parte as observações das restantes ilhas do arquipélago do Havai.

De forma a descrever melhor os grupos encontrados, apresenta-se a média e variância de cada grupo em cada componente principal na Tabela 4.4, devidamente complementada pelos diagramas de dispersão dos *scores* da ACP do *set*₃, com identificação do grupo alocado, que constam da Figura 4.3 e que estão associados às quatro primeiras componentes principais.

O grupo 1 (a **azul** e constituído maioritariamente por observações do Quênia) caracteriza-se por apresentar valores reduzidos na terceira componente principal e negativos (próximos de 0) na segunda, indiciando que o clima é um dos principais fatores que distingue este grupo dos restantes.

O grupo 2 (a **verde escuro** e formado essencialmente por observações do Havai a reduzidas altitudes) apresenta valores reduzidos na segunda componente principal e predominantemente não negativos na terceira. Assim, este grupo pode caracterizar-se pelo clima mais húmido e originário de zonas mais próximas do mar.

O grupo 3 (a **verde** e composto grosso modo por observações da ilha Havai) caracteriza-se por ter

valores elevados na segunda componente principal e reduzidos na primeira. Com efeito, tanto o G_2 como o G_3 caracterizam-se pela reduzida distância ao mar; por outro lado, o contraste existente entre os *scores* destes dois grupos na segunda componente principal leva a considerar que os principais fatores de distinção entre eles são a altitude e o clima, sendo o G_3 de zonas de maior altitude que o G_2 .

O grupo 4 (a preto e constituído principalmente por observações do Brasil) apresenta valores elevados na primeira componente principal e valores reduzidos mas predominantemente positivos na segunda. Consequentemente, depreende-se que o tipo de solo e práticas agrícolas são fatores diferenciadores deste grupo.

Tabela 4.4: Média e variância de cada componente principal retida, obtida da ACP do set_3 , em cada grupo formado.

	Medida	CP 1	CP 2	CP 3	CP 4
G_1	Média	0.334	-0.993	-1.958	-0.225
	Variância	0.417	0.217	0.821	0.625
G_2	Média	-0.991	-1.853	0.577	-0.013
	Variância	0.548	0.471	1.072	0.495
G_3	Média	-1.140	1.730	0.004	-0.282
	Variância	0.540	0.636	0.484	1.753
G_4	Média	1.735	0.394	0.541	0.398
	Variância	3.065	0.546	0.474	0.971

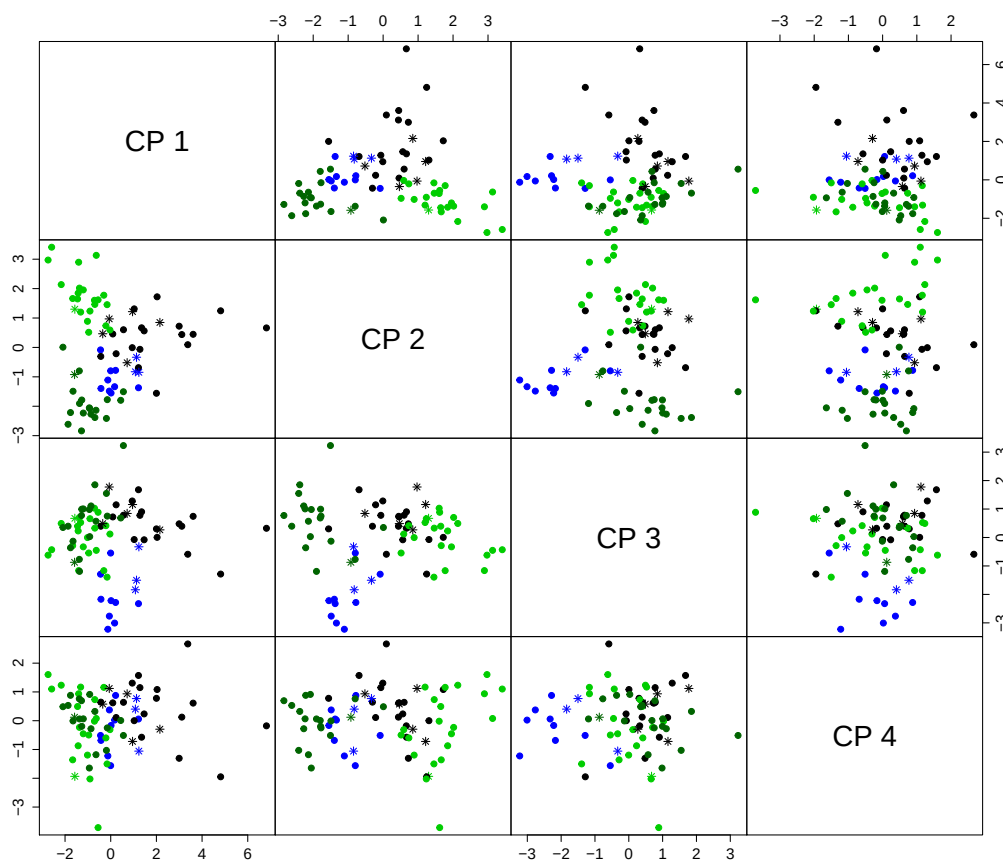


Figura 4.3: Diagramas de dispersão dos *scores* da ACP do set_3 , com as observações coloridas de acordo com o grupo obtido da análise de agrupamentos (azul G_1 , verde escuro G_2 , verde G_3 , preto G_4) e com identificação das observações *mal alocadas* com o símbolo *.

Comentários finais

Neste capítulo os agrupamentos foram construídos de forma não supervisionada, ou seja sem informação da origem geográfica das observações ou outras variáveis de localização, tirando partido apenas das características químicas do grão.

O agrupamento selecionado foi obtido a partir dos *scores* das primeiras quatro componentes principais, consideradas na análise do *set₃* efetuada na Subsecção 3.3.2, e envolveu 76 observações sem valores omissos. O método de formação de grupos utilizado foi o Ward, obtido recorrendo à função *hclust* do pacote *stats* do R e à distância de Minkowski (com $r = 3$). A validação deste agrupamento foi realizada recorrendo a vários métodos, descritos no início desta subsecção. Por fim, o número de grupos selecionado foi quatro.

A análise da partição em quatro grupos permitiu concluir que as características químicas do café em estudo, que neste caso estão refletidas nas componentes principais, têm de facto potencial diferenciador da origem geográfica do grão verde de café. Com efeito, apenas 10 das 76 observações (13%) foram atribuídas a um grupo que não o da sua origem geográfica, designadamente as observações com ID 3 da Etiópia, ID 4 do Brasil, ID 34 do Quênia e as observações com IDs 65, 73, 74, 75, 93, 97 e ID 99 do Havai.

No próximo capítulo aborda-se a classificação dos grãos verdes de café de acordo com a origem geográfica.

Capítulo 5

Classificação

Sendo o principal objetivo deste trabalho a diferenciação da origem geográfica de grãos verdes de café, é imperativo recorrer a métodos de classificação (supervisionada) de forma a encontrar um modelo que satisfaça este propósito. É desejável que este modelo, dada uma nova amostra, seja capaz de classificar segundo a sua origem geográfica com elevada precisão e exatidão. Neste capítulo são apresentados os métodos de classificação considerados e seus principais resultados.

Para a validação dos modelos de classificação recorreu-se à abordagem *leave-one-out*. Com esta abordagem cada conjunto de treino contém $n - 1$ observações e cada conjunto de teste apenas uma observação, onde n representa a dimensão da amostra. Como o número de observações do conjunto de dados em estudo é reduzido, esta abordagem tem a vantagem de utilizar o número máximo possível de dados para treinar o modelo, garantindo que os conjuntos de teste são mutuamente exclusivos e cobrem todo o conjunto de dados (Tan *et al.*, 2005, p. 187). No final, para cada um dos n objetos da amostra dispõe-se da verdadeira classe de cada um e ainda a classe atribuída pelo classificador. Esta reformulação serve de base à construção de indicadores do desempenho do classificador e é, em geral, resumida pela matriz de confusão, que se ilustra na Tabela 5.1, cuja entrada o_{ij} resume o número de observações pertencentes à classe i , C_i , e atribuídas à classe j pelo classificador, C_j , onde $i, j = 1, 2, \dots, c$ e c representa o número de classes em estudo.

Tabela 5.1: Matriz de confusão.

		Classificação			
		C_1	C_2	$\dots$	C_c
Real	C_1	o_{11}	o_{12}	$\dots$	o_{1c}
	C_2	o_{21}	o_{22}	$\dots$	o_{2c}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	C_c	o_{c1}	o_{c2}	$\dots$	o_{cc}

Na vasta literatura existente sobre classificação, são usadas designações distintas para os mesmos indicadores de desempenho de um classificador em diferentes áreas do conhecimento. Os indicadores aqui considerados são: a exatidão global (abreviada por *Accy*);¹ o *recall*;² a precisão; e a taxa de falsos

¹A exatidão será abreviada por *Accy*, abreviatura usualmente utilizada deste termo em inglês (*Overall Accuracy*).

²Uma vez que a tradução do termo *recall* não se encontra no Glossário Inglês-Português de Estatística (Sociedade Portuguesa de Estatística – Associação Brasileira de Estatística), disponível em <http://glossario.spestatistica.pt/> e consultado a 7 de novembro de 2015, iremos utilizar a denominação do mesmo em inglês.

positivos (abreviada por FPR).³

A exatidão global é a percentagem de observações bem classificadas e propõe-se ser uma estimativa da probabilidade total de uma observação ser bem classificada. No caso *leave-one-out*, a exatidão global é dada por

$$Accy = \sum_{i=1}^c \frac{o_{ii}}{n}. \quad (5.1)$$

Este indicador dá informação sobre o desempenho global do classificador, contudo não fornece informação sobre o que acontece em cada classe. Assim, é possível que, no caso de uma das classes ser muito frequente, a exatidão global seja elevada mas a única classe que é bem classificada é a mais numerosa. Para completar esta informação é usual estudar o *recall*, a precisão e a taxa de falsos positivos de cada classe e com base nestes indicadores definir *recall*, precisão e taxa de falsos negativos globais.

O *recall* da classe i , $Recall(i)$, é a percentagem de observações da classe i corretamente classificadas e definida por:

$$Recall(i) = \frac{o_{ii}}{o_{i.}}, \quad (5.2)$$

onde $o_{i.} = \sum_{j=1}^c o_{ij}$ e $i = 1, \dots, c$.

O *recall* global é definido como a média dos valores por classe, isto é,

$$\overline{Recall} = \sum_{i=1}^c \frac{Recall(i)}{c}. \quad (5.3)$$

A precisão da classe i , $Prec(i)$, define-se como a percentagem de observações atribuídas à classe i que efetivamente pertencem a esta classe, ou seja,

$$Prec(i) = \frac{o_{ii}}{o_{.i}}, \quad (5.4)$$

onde $o_{.i} = \sum_{k=1}^c o_{ki}$ e $i = 1, \dots, c$.

Define-se a precisão global, $\overline{Prec}$, como a média de precisão de cada classe:

$$\overline{Prec} = \sum_{i=1}^c \frac{Prec(i)}{c}. \quad (5.5)$$

Por fim, a taxa de falsos positivos da classe i , $FPR(i)$, define-se como a percentagem de observações que não pertencem à classe i mas que foram classificadas nesta classe:

$$FPR(i) = \frac{o_{.i} - o_{ii}}{n - o_{i.}}, \quad (5.6)$$

onde $i = 1, \dots, c$.

Pode definir-se também a taxa de falsos positivos global, $\overline{FPR}$, como a média da taxa de falsos

³A taxa de falsos positivos será abreviada por FPR, abreviatura deste termo em inglês (*False Positive Rate*).

positivos:

$$\overline{FPR} = \sum_{i=1}^c \frac{FPR(i)}{c}. \quad (5.7)$$

Para auxiliar na escolha dos melhores classificadores, recorreu-se ainda a uma medida global que sumaria os quatro indicadores de desempenho. Esta medida foi definida como a distância ao valor ótimo de cada indicador, tendo a designação de distância ao ótimo (*Dist. Otimo*):

$$Dist. Otimo = \sqrt{(Accy - 1)^2 + (\overline{Prec} - 1)^2 + (\overline{Recall} - 1)^2 + \overline{FPR}^2}. \quad (5.8)$$

No que diz respeito aos métodos de classificação, recorreu-se a: árvores de decisão (Quinlan, 1992), nomeadamente ao algoritmo C5.0 (RuleQuest Research, 2013); k vizinhos mais próximos (kNN) (Tan *et al.*, 2005, pp. 223–227); análise discriminante (Johnson e Wichern, 2007, pp. 575–621); e regressão logística (Johnson e Wichern, 2007, pp. 634–644). Foram também aplicadas generalizações da regressão logística, em que na função objetivo a otimizar se incorpora a estimação dos coeficientes de regressão logística e um termo adicional. Este termo adicional pretende penalizar soluções com todos os coeficientes associados às variáveis explicativas elevados e valorizar soluções em que poucas variáveis têm coeficientes elevados e a generalidade delas possui coeficientes baixos. Esta estratégia conduz a uma seleção implícita das variáveis mais relevantes para classificar os objetos. Os dois métodos empregues, pertencentes a esta família, são o SMLR, *Sparse Multinomial Logistic Regression*, e o RMLR, *Ridge Multinomial Logistic Regression* (Krishnapuram *et al.*, 2005; Lal *et al.*, 2006).

Todos estes métodos serão descritos sucintamente na Secção 5.2 e os resultados obtidos da sua aplicação discutidos na Secção 5.3. De seguida apresenta-se uma breve introdução às técnicas automáticas de seleção de variáveis.

5.1 Seleção de variáveis

A procura de um subconjunto de variáveis que possa descrever eficientemente os dados, reduzindo os efeitos de ruído ou a inclusão de variáveis irrelevantes para a análise do problema e conduzindo ainda a bom desempenho de classificação (Chandrashekar e Sahin, 2014), é o foco da seleção de variáveis. Uma correta seleção de variáveis pode conduzir à melhoria de um classificador, não só em termos de velocidade de execução e simplicidade do mesmo como também em capacidade de generalização. Outro dos benefícios associados à redução do número de variáveis é a redução do custo associado à recolha de informação (Bolón-Canedo *et al.*, 2013).

A seleção de variáveis pode ser realizada manualmente, tal como na Secção 3.1 em que foram formados vários subconjuntos com diferentes variáveis, ou recorrendo a métodos automáticos. Estes podem agrupar-se em 3 grandes grupos: filtros, *wrapper* e *embedded*.

Os métodos de seleção de variáveis do tipo filtro atuam como uma fase do pré-processamento dos dados, fornecendo uma ordem da importância das variáveis para a explicação das várias classes, utilizando um índice de relevância. Entre as vantagens destes métodos está a independência do clas-

sificador escolhido, a rapidez e ainda a redução da ocorrência de sobre-estimação dos classificadores ao conjunto de dados em estudo (Bolón-Canedo *et al.*, 2013; Chandrashekar e Sahin, 2014).

Ao contrário dos filtros, os métodos *wrapper* interagem com o classificador, medindo a qualidade do subconjunto de variáveis de acordo com o seu desempenho classificativo (Bolón-Canedo *et al.*, 2013; Chandrashekar e Sahin, 2014). Não obstante, estes métodos não incorporam conhecimento sobre a estrutura específica da classificação, podendo assim ser combinados com qualquer classificador (Lal *et al.*, 2006). Os métodos *wrapper* capturam as dependências entre as variáveis e recorrem a validação cruzada. Esta característica conduz a elevado custo computacional; mais ainda, a propensão a *overfitting* e dependência do classificador são outras das desvantagens deste método (Bolón-Canedo *et al.*, 2013; Chandrashekar e Sahin, 2014).

A seleção de variáveis pode ser definida como uma forma de encontrar o subconjunto de variáveis que leve à maior generalização possível ou, equivalentemente, ao risco mínimo. É neste contexto que surgem os métodos *embedded*, definidos por um problema de minimização em que a seleção de variáveis interage com a fase de treino do classificador. As soluções deste problema de minimização são aproximadas, utilizando, por exemplo, um termo de dispersão na função objetivo para aproximação de soluções do problema, que é o caso de um dos métodos de classificação que será abordado na subsecção 5.2.5, o SMLR (Lal *et al.*, 2006).

À semelhança dos *wrapper*, os métodos *embedded* interagem com o classificador, incorporando a seleção de variáveis no processo de treino, no entanto não necessitam de separar os dados em treino e teste, pelo que são mais eficientes e menos propensos a *overfitting* que os métodos *wrapper*. Por outro lado, os métodos *embedded* são mais propensos a *overfitting* quando comparados com os métodos filtro. Para conjuntos de dados de elevada dimensão os métodos *embedded* têm geralmente melhores desempenhos que os métodos *wrapper* (Chandrashekar e Sahin, 2014).

Em Lal *et al.* (2006) pode encontrar-se mais informação acerca destes métodos de seleção de variáveis, nomeadamente dos métodos *embedded*.

5.2 Métodos de classificação

Nesta secção faz-se uma breve introdução aos métodos de classificação usados neste trabalho: árvores de decisão, kNN, análise discriminante e regressão logística, bem como regressão logística multinomial que incorpora seleção de variáveis *embedded*, nomeadamente SMLR e RMLR.

5.2.1 Árvores de decisão

O princípio básico dos modelos de árvores de decisão é a partição (de forma recursiva) do espaço associado a cada variável, de forma a minimizar uma medida do grau de impureza dos nós (Hand *et al.*, 2001, p. 343). Entre as medidas de impureza mais conhecidas estão a entropia, o índice de Gini e o erro de classificação.

Numa árvore de decisão, a cada nó terminal é atribuída uma classe, os nós não terminais, que

incluem a raiz e outros nós internos, contêm uma condição de teste sobre as variáveis de forma a separar observações com diferentes características (Tan *et al.*, 2005, p. 150). Enumeram-se de seguida algumas das características das árvores de decisão mencionadas por Tan *et al.* (2005, pp. 168–170).

1. As árvores de decisão constituem uma abordagem não paramétrica para construir classificadores.
2. Encontrar uma árvore de decisão ótima é um problema NP completo.
3. As técnicas desenvolvidas para construção de árvores de decisão são computacionalmente pouco dispendiosas, sendo possível construir rapidamente uma árvore mesmo quando o conjunto de treino é de elevada dimensão.
4. As árvores de decisão, especialmente as de tamanho reduzido, são relativamente fáceis de interpretar.
5. Os algoritmos de árvores de decisão são consideravelmente robustos na presença de observações atípicas (especialmente quando são empregues métodos que evitam o *overfitting*).
6. A presença de variáveis redundantes não afetam adversamente o desempenho das árvores de decisão. Contudo, no caso em que o conjunto de dados contém várias variáveis redundantes, algumas podem ser escolhidas durante o processo de construção da árvore, resultando numa árvores maior que o necessário.
7. Como a maior parte dos algoritmos de árvores de decisão utilizam partições construídas de forma recursiva, o número de observações associado a cada nó reduz-se ao longo da árvore. Assim, em alguns dos nós terminais, o número de observações pode ser demasiado reduzido para tomar decisões estatisticamente significativas sobre a classe representante desse nó. Este problema é conhecido como o problema da fragmentação de dados.

5.2.2 K vizinhos mais próximos

O método de classificação dos k vizinhos mais próximos (kNN – *k Nearest Neighbors*) é um método não paramétrico e baseia-se em distâncias entre as observações.

A principal desvantagem do método kNN é que não constrói um modelo de classificação. Na classificação de um novo objeto intervêm todos os objetos do conjunto de treino, o que leva a um custo computacional elevado. Por esta razão, o método kNN é geralmente chamados de preguiçoso (*lazy method*) (Hand *et al.*, 2001, p. 351). Este método tem ainda a desvantagem de ser sensível a observações atípicas, à presença de variáveis irrelevantes, bem como à disparidade de dimensão entre as classes, ou seja, o algoritmo kNN tende a classificar na classe de maior dimensão. Todavia tem a vantagem de ser simples e de fácil compreensão.

Dada uma nova observação, o procedimento do kNN classifica esta observação na classe que está em maioria entre as k observações mais próximas (i.e., as k observações com a menor distância a essa nova observação). Para tal, é necessária a escolha da distância a utilizar e do número de vizinhos, k , a considerar. A distância mais utilizada é a euclidiana, pela qual se optou neste trabalho. A escolha do número de vizinhos a considerar, k , foi feita tendo por base o desempenho do classificador para cada

valor de k , desempenho esse estimado a partir do método de validação *leave-one-out*. Nesta escolha teve-se em conta que um valor de k reduzido leva a maior sensibilidade a observações atípicas e um valor de k elevado pode conduzir à perda de padrões existentes nos dados ou a *overfitting* (Tan *et al.*, 2005).

5.2.3 Análise discriminante

A análise discriminante é uma técnica de classificação que visa encontrar uma regra de classificação que minimize a probabilidade total de má classificação.

Caso o vetor aleatório em estudo, condicional a cada classe, possua distribuição normal multivariada com matrizes de covariâncias comum em todas as classe, prova-se que o classificador ótimo se escreve à custa de uma combinação linear das variáveis em estudo. Esta técnica é conhecida por análise discriminante linear (LDA – *Linear Discriminant Analysis*). Caso as matrizes de covariâncias sejam distintas então a regra de classificação ótima é uma função quadrática das variáveis em estudo. Neste caso, a técnica é conhecida como análise discriminante quadrática (QDA – *Quadratic Discriminant Analysis*).

A dedução das regras de classificação linear e quadrática estão disponíveis na generalidade dos livros de Análise Multivariada que discutem esta metodologia (por exemplo em Johnson e Wichern, 2007, pp. 610–611). Sejam c o número de classes em estudo, $\mathbf{X} = (X_1, \dots, X_p)^t$ o vetor aleatório que caracteriza cada objeto e Y a variável aleatória que indica a classe a que cada objeto pertence. Considere-se ainda que $p_j = P(Y = j)$, com $\sum_{j=1}^c p_j = 1$, ou seja p_j representa a probabilidade *a priori* de um objeto pertencer à classe j . Se pudermos assumir que

- a variável aleatória $(\mathbf{X} | Y = j)$ possui distribuição normal multivariada com valor esperado $\boldsymbol{\mu}_j \in \mathbb{R}^p$ e matriz de covariâncias $\boldsymbol{\Sigma}$, para $j = 1, 2, \dots, c$, isto é,

$$\mathbf{X} | Y = j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \quad j = 1, 2, \dots, c, \quad (5.9)$$

então a regra de classificação que minimiza a probabilidade total de má classificação consiste em atribuir $\mathbf{x}_0$ à classe k com

$$k = \arg \max_{j=1,2,\dots,c} d_j^{LDA}(\mathbf{x}_0), \quad (5.10)$$

onde

$$d_j^{LDA}(\mathbf{x}_0) = \boldsymbol{\mu}_j^t \boldsymbol{\Sigma}^{-1} \mathbf{x}_0 - \frac{1}{2} \boldsymbol{\mu}_j^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \ln(p_j). \quad (5.11)$$

No caso mais geral, em que $\mathbf{X} | Y = j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, então a regra de classificação que minimiza a probabilidade total de má classificação consiste em atribuir $\mathbf{x}_0$ à classe

$$k = \arg \max_{j=1,2,\dots,c} d_j^{QDA}(\mathbf{x}_0), \quad (5.12)$$

onde

$$d_j^{QDA}(\mathbf{x}_0) = -\frac{1}{2} \ln(\det(\boldsymbol{\Sigma}_j)) - \frac{1}{2} (\mathbf{x}_0 - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_j) + \ln(p_j). \quad (5.13)$$

Na prática, $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$ e p_j não são conhecidos e têm de ser estimados. Estas estimações podem ser realizadas da forma clássica ou então recorrendo a estimadores robustos. Por exemplo, na estimação

de μ_j e Σ_j pode recorrer-se ao estimador robusto obtido pelo método do determinante de covariância mínimo, *Fast MCD*, ou ao estimador obtido pelo método do elipsoide de volume mínimo, *Fast MVE* (Maronna *et al.*, 2006, p. 187, 190). As probabilidades *a priori*, p_j , $j = 1, \dots, c$, podem ser estimadas pela frequência relativa de observações em cada classe (Todorov e Filzmoser, 2009). No caso da LDA é necessária a estimação da matriz de covariâncias comum, Σ , uma possibilidade é utilizar a estimativa da matriz de covariâncias agrupada (*pooled*), S_p (Johnson e Wichern, 2007, p. 611):

$$S_p = \frac{\sum_{j=1}^c (n_j - 1) S_j}{\sum_{j=1}^c n_j - c}. \quad (5.14)$$

Ao substituir os verdadeiros valores dos parâmetros nas equações (5.11) e (5.13) pelos seus equivalentes amostrais obtêm-se estimativas das regras ótimas. Espera-se que estas estimativas estejam suficientemente próximas dos equivalentes populacionais, conduzindo a valores de desempenho dos classificadores (5.11) e (5.13) próximos dos ótimos.

As hipóteses de normalidade multivariada dos dados (hipótese: $\mathbf{X} \sim N_p(\mu, \Sigma)$) e dos dados condicionais à classe (hipótese: $\mathbf{X} | Y = j \sim N_p(\mu_j, \Sigma_j)$, $j = 1, 2, \dots, c$) foram testadas recorrendo ao teste de Henze-Zirkler (função `hzTest` do pacote `MVN` do R) e ao de Anderson-Darling (função `adtest` do pacote `robCompositions` do R). Devido à reduzida dimensão da classe da Etiópia a hipótese de normalidade não foi testada para esta classe.

A hipótese de normalidade multivariada das distribuições dos dados condicionais às classes não foi rejeitada pelo teste de Anderson-Darling em nenhum dos subconjuntos (sem valores omissos), ao nível de significância de 5%. No entanto, com o teste de Henze-Zirkler, rejeita-se tal hipótese em especial para a classe do Havai. No que respeita à hipótese de igualdade das matrizes de covariâncias, utilizando o teste de M Box (função `boxM` do pacote `biotool` do R), esta é rejeitada; note-se, no entanto, que este teste pressupõe que os dados seguem uma distribuição normalidade multivariada, hipótese esta de trabalho rejeitada pelos testes de Anderson-Darling e de Henze-Zirkler, ao nível de significância de 5%. Posto isto espera-se obter melhores desempenhos na QDA que na LDA, pois a primeira pressupõe normalidade mas não matrizes de covariâncias iguais.

Importa notar também que, para este conjunto de dados, as classes apresentam dimensão reduzida (como se pode constatar pela Tabela 3.3): a classe do Quénia ronda as 10 observações e a classe do Brasil as 20 observações na maioria dos subconjuntos. Os testes de normalidade em dados de reduzida dimensão têm pouco poder para rejeitar a hipótese nula de normalidade, levando à não rejeição desta hipótese na maioria dos casos (Ghasemi e Zahediasl, 2012).

5.2.4 Regressão logística

A regressão logística analisa a relação entre as várias variáveis explicativas e a variável categórica resposta, estimando a probabilidade de ocorrência de um evento ajustando os dados a uma curva logística (Park, 2013). Neste trabalho a variável resposta é a origem geográfica do grão verde do café.

Existem dois tipos de modelo de regressão logística: a regressão logística binária e a regressão logística multinomial. A regressão logística binária é utilizada quando a variável resposta é binária e as variáveis explicativas são contínuas ou categóricas. Quando a variável resposta compreende mais do que duas categorias pode recorrer-se a regressão logística multinomial (Park, 2013).

Seja Y uma variável resposta binária (0 ou 1) tal que $P(Y = 1 | \mathbf{x}) = \pi(\mathbf{x})$, $P(Y = 0 | \mathbf{x}) = 1 - \pi(\mathbf{x})$ e $\pi(\mathbf{x}) = E(Y | \mathbf{x})$. Mais, designe-se por *odds* a razão entre a probabilidade de um evento ocorrer e a probabilidade desse evento não ocorrer, isto é,

$$\text{odds}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}, \quad (5.15)$$

sendo que o impacto das variáveis explicativas é explicado em termos desta razão (Park, 2013). Na regressão logística, para variáveis binárias, considera-se o logaritmo natural de *odds*, a que chamamos de *logit*:

$$\text{logit}(\pi(\mathbf{x})) = \ln(\text{odds}(\mathbf{x})) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right). \quad (5.16)$$

Ao lidar-se com uma única variável explicativa X , considera-se $\text{logit}(\pi(x)) = \beta_0 + \beta_1 x$ e em termos de probabilidades obtém-se

$$\text{odds}(x) = \frac{\pi(x)}{1 - \pi(x)} = \exp(\beta_0 + \beta_1 x), \quad (5.17)$$

ou equivalentemente,

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \quad (5.18)$$

equação esta que descreve uma curva logística.

Considere-se agora um modelo com p variáveis explicativas. Sejam $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^t$ o vetor constituído pelos valores tomados pelas p variáveis na i -ésima observação e $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^t$ o vetor de coeficientes da regressão logística. Então o modelo admite a seguinte forma:

$$\ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (5.19)$$

Na estimação dos β 's pode recorrer-se ao método da máxima verosimilhança. Após a estimação da função de regressão logística e recorrendo a conjuntos de treino para cada uma das populações ($Y = 1$ e $Y = 0$), a regra de classificação define-se tal como em Johnson e Wichern (2007, pp. 634–638).

Classifica-se uma nova observação $\mathbf{x}_0$ na população 1 se a razão *odds* é maior que 1, ou seja,

$$\frac{\hat{\pi}(\mathbf{x}_0)}{1 - \hat{\pi}(\mathbf{x}_0)} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p\right) > 1. \quad (5.20)$$

Caso se considerem mais que duas classes, recorre-se à técnica 1-versus-todos, descrita de seguida. Defina-se $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(c)})^t$, em que $y^{(j)} = 1$ se $\mathbf{x}$ é uma observação pertencente à classe j , $y^{(j)} = 0$ se $\mathbf{x}$ pertence a qualquer uma das restantes classes, $j \in \{1, 2, \dots, c\}$ e c é o número de classes em estudo. Desta forma pode representar-se o conjunto de treino com n observações por $D = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$. Neste caso a matriz de coeficientes é definida

por $\beta = (\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(c)})$, onde $\beta^{(j)}$ é o vetor de coeficientes da regressão logística associado à classe j . Assim, obtém-se uma função logística para cada classe, sendo uma nova observação classificada na classe cuja função logística leva à maior probabilidade de pertença.

No entanto, devido à condição de normalização

$$\sum_{j=1}^c P(Y^{(j)} = 1 | x, \beta) = 1, \quad (5.21)$$

não é necessário estimar os valores de β para uma das classes. Deste modo e sem perda de generalidade, consideramos $\beta^{(c)} = 0$ e é apenas necessário estimar os vetores de coeficientes $\beta^{(i)}$ para $i = \{1, \dots, c-1\}$, tal como sugerem Krishnapuram *et al.* (2005).

Para finalizar, a probabilidade de uma observação x pertencer à classe j encontra-se em Krishnapuram *et al.* (2005) e é dada por

$$P(Y^{(j)} = 1 | x, \beta) = \frac{\exp(\beta_0^{(j)} + \beta_1^{(j)}x_1 + \dots + \beta_p^{(j)}x_p)}{\sum_{i=1}^c \exp(\beta_0^{(i)} + \beta_1^{(i)}x_1 + \dots + \beta_p^{(i)}x_p)}, \text{ para } j = 1, 2, \dots, c. \quad (5.22)$$

A regra de classificação consiste em atribuir a nova observação x_0 à classe j caso

$$j = \arg \max_{j=1, \dots, c} P(Y^{(j)} = 1 | x_0, \beta). \quad (5.23)$$

5.2.5 Regressão logística multivariada com incorporação de seleção de variáveis – SMLR e RMLR.

Os métodos de classificação apresentados nesta secção fazem parte do conjunto de métodos *embedded*, pois incorporam seleção de variáveis no processo de treino. As abordagens estudadas foram a regressão logística multinomial esparsa (*sparse multinomial logistic regression*, SMLR) proposta em Krishnapuram *et al.* (2005), e a sua versão não-esparsa (*ridge multinomial logistic regression*, RMLR).

No contexto de classificação e ao recorrer-se à regressão logística multinomial descrita na Secção 5.2.4, os coeficientes β são estimados a partir dos dados de treino, maximizando a função verosimilhança (Krishnapuram *et al.*, 2005)

$$\ell(\beta) = \sum_{j=1}^n \ln P(Y_j | x_j, \beta) = \sum_{j=1}^n \left[\sum_{i=1}^c y_j^{(i)} \beta^{(i)t} x_j - \ln \left(\sum_{i=1}^c \exp(\beta^{(i)t} x_j) \right) \right], \quad (5.24)$$

onde n representa o número de observações dos dados de treino, c o número de classes e $\beta = (\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(c)})$ a matriz de coeficientes, que em cada coluna tem o vetor de coeficientes da regressão logística associado a cada classe, $\beta^{(i)}$, $i = 1, 2, \dots, c$. Mais ainda, devido à condição de normalização em (5.21) não é necessário estimar os valores de β para uma das classes. Deste modo, denotar-se-á também por β o vetor dos coeficientes a estimar com dimensão $(p+1) \times (c-1)$ (Krishnapuram *et al.*, 2005).

A maximização da função definida em (5.24) é geralmente efetuada recorrendo ao método dos mínimos quadrados reponderados iterativamente (*Iteratively Reweighted Least Squares* - IRLS). Contudo, a função $\ell(\beta)$ pode tornar-se arbitrariamente grande, sendo necessária a introdução de uma *priori* para β (Krishnapuram *et al.*, 2005). Motivando assim a adoção de um estimador máximo a *posteriori* (MAP) no lugar do estimador de máxima verosimilhança:

$$\hat{\beta}_{MAP} = \arg \max_{\beta} L(\beta) = \arg \max_{\beta} [\ell(\beta) + \ln p(\beta)], \quad (5.25)$$

onde $p(\beta)$ representa a densidade conjunta *a priori* (ou *prior*) do vetor de coeficientes β (Krishnapuram *et al.*, 2005; Narlikar e Hartemink, 2006). O algoritmo IRLS pode ser modificado para acomodar certas *priors*, como por exemplo a *prior* de Gauss

$$p(\beta) \propto \exp(-\lambda \|\beta\|_2^2), \quad (5.26)$$

onde $\|\beta\|_2 = \sqrt{\sum_i \beta_i^2}$ denota a norma l_2 de β . Contudo existem outras *priors* que não são passíveis de acomodar como é o caso da *prior* de Laplace

$$p(\beta) \propto \exp(-\lambda \|\beta\|_1), \quad (5.27)$$

onde $\|\beta\|_1 = \sum_i |\beta_i|$ denota a norma l_1 de β , por esta *prior* não ser diferenciável na origem (Krishnapuram *et al.*, 2005). Para resolver este problema Krishnapuram *et al.* (2005) recorrem a um algoritmo que combina uma abordagem de otimização com um procedimento de atualização componente a componente, em que a função objetivo a maximizar é $L(\beta) = \ell(\beta) + \ln p(\beta)$. Este algoritmo é aplicado à regressão logística multinomial como alternativa ao IRLS.

O método RMLR, que recorre à *prior* de Gauss, tem como função objetivo a função

$$L(\beta) = \ell(\beta) - \frac{\lambda}{2} \|\beta\|_2^2, \quad (5.28)$$

enquanto que para o SMLR, com *prior* de Laplace, a função objetivo é

$$L(\beta) = \ell(\beta) - \frac{\lambda}{2} \|\beta\|_1. \quad (5.29)$$

De notar que, enquanto o SMLR utiliza uma *prior* que promove a dispersão dos coeficientes β , a *prior* do RMLR promove a contração (Krishnapuram *et al.*, 2005). O parâmetro λ denota o recíproco da variância *a priori* e pode ser entendido como um parâmetro de regularização dos coeficientes, quanto maior é o seu valor maior a dispersão (contração) dos coeficientes do SMLR (RMLR) (Narlikar e Hartemink, 2006).

O principal objetivo da inclusão de um parâmetro de regularização é incentivar as estimativas dos coeficientes a serem significativamente grandes ou exatamente zero, levando à remoção automática de variáveis irrelevantes, o que resulta num processo de seleção de variáveis (Krishnapuram *et al.*, 2005). Este processo é especialmente utilizado em casos em que o conjunto de variáveis é de elevada

dimensão e o objetivo é reduzir o tempo e custo de extração de dados. Este tipo de métodos é bastante utilizado em biologia molecular, por exemplo em Narlikar e Hartemink (2006). No que respeita ao nosso estudo, apesar da dimensão do conjunto de variáveis em estudo ser reduzida, o elevado custo associado à recolha de novas amostras leva à necessidade de selecionar as variáveis mais relevantes para o problema de diferenciação da origem geográfica do grão de café verde.

Para o treino e validação dos classificadores recorreu-se à implementação dos algoritmos do SMLR e RMLR desenvolvida por Bosko *et al.* (2007), que consiste num *software package* desenvolvido em Java. Esta implementação permite escolher diversos parâmetros, entre eles estão o *Kernel*, o termo de viés, a *prior* (Laplaciana ou Gaussiana) e o parâmetro de regularização λ . Para que a estimação dos coeficientes da regressão logística seja feita a partir das próprias variáveis, incorporando seleção de variáveis, foi escolhida a opção *Direct* no parâmetro do *Kernel*, isto é, nenhum *kernel* é utilizado. Relativamente ao termo de viés, os valores testados foram 0, 0.5, 1, 2, 5 e 8. Este termo, quando diferente de zero, permite que o classificador encontre uma fronteira de decisão que não passe pela origem (Bosko *et al.*, 2007). Por fim, escolheram-se valores de $\lambda \in \{0.1, 0.2, 0.5, 0.8, 1, 5\}$.

5.3 Resultados e discussão

Como se referiu anteriormente, os modelos de classificação foram sujeitos a validação recorrendo à abordagem *leave-one-out*. Nesta secção o objetivo é analisar os desempenhos dos classificadores, bem como os melhores modelos analisados, de forma a escolher um classificador de grãos verdes de café de acordo com a origem geográfica.

Vários subconjuntos de variáveis e várias partições em classes foram analisadas. Estas podem agrupar-se em 3 conjuntos diferentes.

1. Cada classe corresponde a uma região produtora entre as quatro em estudo, Brasil (BZ), Etiópia (ET), Havai (HW) e Quénia (KE), ou três nos casos em que não se considerou a Etiópia.
2. As classes são as regiões produtoras do Brasil, da Etiópia (quando considerada), do Quénia, da ilha Havai (HW_i) e das restantes ilhas do arquipélago do Havai (HW_r). Neste caso as regiões do arquipélago do Havai foram separadas em 2 grupos, como sugerido no agrupamento encontrado no Capítulo 4: um grupo com as observações da ilha Havai e o outro com as observações das restantes ilhas.
3. Consideram-se apenas duas classes: uma formada pela Etiópia (ET) e outra pelas restantes origens geográficas, Brasil, Havai e Quénia ($\{BZ, HW, KE\}$).

Na Secção 3.4 foi efetuada a deteção de *outliers* no conjunto de dados, não se identificando nenhuma observação da Etiópia como *outlier*. Neste capítulo, a classificação com a partição ET vs. $\{BZ, HW, KE\}$ pode ser encarada como um caso de deteção de *outliers* supervisionada. Com efeito, a classe da Etiópia apresenta um número de observações muito reduzido, quando comparada com a classe $\{BZ, HW, KE\}$, pelo que esta partição foi criada para tentar classificar, com sucesso, as observações da Etiópia. Este caso foi o único a que se recorreu a custos na classificação.

Foram analisadas várias combinações de métodos de classificação, subconjunto de variáveis e conjuntos de classes; a Tabela 5.2 sumaria estas combinações (VO significa Valores Omissos). Adiante-se também que os métodos de regressão logística estudados foram aplicados aos dados padronizados (recorrendo à média e ao desvio padrão), bem como não padronizados, ao contrário dos restantes métodos de classificação, que apenas foram aplicados aos dados não padronizados. Mais ainda, como se pode constatar após a consulta da Tabela 5.2 apenas se tirou partido das observações que contêm valores omissos para treino do classificador C5.0.

Tabela 5.2: Classes e subconjuntos de variáveis estudados para cada método de classificação (VO-Valores omissos).

Classificador	Classes	Com VO	Sem VO	Subconjuntos de variáveis
C5.0	BZ, ET, HW e KE	X	X	<i>set₃</i> a <i>set₁₂</i> , <i>ACP set₃</i> , <i>ACP set₉</i> e <i>ACP set₁₀</i>
	BZ, HW e KE	X	X	<i>set₃</i> a <i>set₁₂</i> , <i>ACP set₃</i> , <i>ACP set₉</i> e <i>ACP set₁₀</i>
	BZ, ET, HW _r , HW _i e KE	X		<i>set₃</i> a <i>set₁₂</i>
	BZ, HW _r , HW _i e KE	X	X	<i>set₃</i> a <i>set₁₂</i>
	ET vs. {BZ, HW, KE}	X		<i>set₃</i> a <i>set₁₂</i>
kNN	BZ, ET, HW e KE		X	<i>set₃</i> a <i>set₁₂</i> e <i>PCA set₃</i>
	BZ, HW e KE		X	<i>set₃</i> a <i>set₁₂</i> , <i>ACP set₃</i> , <i>ACP set₉</i> e <i>ACP set₁₀</i>
A. Discriminante	BZ, HW e KE		X	<i>set₃</i> a <i>set₁₂</i> , <i>ACP set₃</i> , <i>ACP set₉</i> e <i>ACP set₁₀</i>
Reg. Logística	BZ, ET, HW e KE		X	<i>set₃</i> a <i>set₁₂</i>
	BZ, HW e KE		X	<i>set₃</i> a <i>set₁₂</i> , <i>ACP set₉</i> e <i>ACP set₁₀</i>
SMLR	BZ, HW e KE		X	<i>set₃</i> , <i>set₇</i> , <i>set₉</i> e <i>set₁₀</i>
	BZ, HW _r , HW _i e KE		X	<i>set₃</i> , <i>set₇</i> , <i>set₉</i> e <i>set₁₀</i>
	ET vs. {BZ, HW, KE}		X	<i>set₇</i> e <i>set₈</i>
RMLR	BZ, HW e KE		X	<i>set₃</i> , <i>set₇</i> , <i>set₉</i> e <i>set₁₀</i>

ACP no *set₃* com 4 CPs retidas, ACP no *set₉* e ACP no *set₁₀* com 3 CPs retidas.

As funções utilizadas para treino dos classificadores foram:

- árvores de decisão: *C5.0* do pacote *C50* do R;
- kNN: *knn* do pacote *class* do R;
- análise discriminante: *LdaClassic*, *LdaPP*, *QdaClassic* e *QdaRob* do pacote *rrcov* do R;
- regressão logística: *multinom* do pacote *nnet* do R;
- SMLR e RMLR: *software package SMLR* (Bosko *et al.*, 2007).

No caso da QDA robusta, os métodos utilizados para estimação dos parâmetros de escala e localização foram o MVE (*Minimum Volume Ellipsoid*) e o SDE (*Stahel-Donoho Estimates*) (Maronna *et al.*, 2006, pp. 187 e 194). Os restantes argumentos utilizados na QDA e nos restantes métodos de classificação correspondem às opções da rotina por omissão.

De seguida apresentam-se os resultados obtidos ao aplicar os diferentes métodos de classificação aos vários subconjuntos de variáveis e partições em classes. Contudo, não serão apresentados os resultados obtidos para os classificadores em que se incluiu a classe das observações da Etiópia. Esta decisão prendeu-se com o facto de as observações provenientes desta região registarem muitos valores omissos, existindo apenas uma observação sem nenhum valor omissos em todas as variáveis em estudo. Mais ainda, os classificadores obtidos, dos quais foi possível utilizar a maioria das observações da Etiópia, possuem mau desempenho, com valores nos indicadores de desempenho para a Etiópia em torno ou abaixo de 0.5.

5.3.1 Desempenho dos diversos classificadores

Após análise do desempenho de todos os classificadores foram selecionados os melhores de cada método. Os respectivos indicadores de desempenho constam da Tabela 5.3. Estes indicadores foram obtidos por validação *leave-one-out*, usando apenas observações completas.

Os indicadores de desempenho na Tabela 5.3 permitem concluir que se conseguiu classificar geograficamente as 3 origens (Brasil, Havai e Quênia) com bons resultados, pois obteve-se maioritariamente valores acima de 0.8 (abaixo de 0.2) nos indicadores *Recall*, *Precisão* e *Accy* (no indicador FPR), para todos os classificadores. Analisando detalhadamente, verifica-se ainda que o desempenho dos classificadores que utilizam o set_9 ou o set_{10} estão, em geral, entre os melhores de todos os métodos de classificação aplicados.

Tabela 5.3: Medidas de desempenho dos classificadores selecionados, treino sem valores omissos e sem as observações da Etiópia.

Classificador	Subconjunto	Variáveis	Classe	<i>Recall</i>	<i>Precisão</i>	FPR	<i>Accy</i>	Dist. Otim
C5.0	ACP set_{10} 3 CPs		Brasil	0.900	0.750	0.107	0.868	0.233
			Havai	0.826	0.950	0.067		
			Quênia	1.000	0.833	0.030		
			Global	0.909	0.844	0.068		
C5.0	set_3	$C^{\%}, N^{\%}, \delta^{13}C,$ $\delta^{18}O, \delta^{15}N, \delta^{34}S,$ $\delta^{13}C_{caf}, \delta^{18}O_{caf},$ $\delta^{15}N_{caf}, {}^{87}Sr/{}^{86}Sr$	Brasil	0.842	0.800	0.071	0.867	0.274
			Havai	0.891	0.911	0.138		
			Quênia	0.800	0.800	0.031		
			Global	0.844	0.837	0.080		
kNN ($k = 6$)	ACP set_9 3 CPs		Brasil	0.850	0.850	0.053	0.909	0.167
			Havai	0.915	0.935	0.100		
			Quênia	1.000	0.909	0.015		
			Global	0.922	0.898	0.056		
kNN ($k = 3$)	ACP set_{10} 3 CPs		Brasil	0.900	0.783	0.089	0.908	0.154
			Havai	0.891	0.953	0.067		
			Quênia	1.000	1.000	0.000		
			Global	0.930	0.912	0.052		
QDA Clássica	set_9	$\delta^{13}C_{caf},$ $\delta^{18}O_{caf},$ $\delta^{15}N_{caf},$ ${}^{87}Sr/{}^{86}Sr$	Brasil	0.850	0.944	0.018	0.909	0.204
			Havai	0.979	0.885	0.200		
			Quênia	0.700	1.000	0.000		
			Global	0.843	0.943	0.073		
QDA MVE	set_9		Brasil	0.850	0.810	0.070	0.883	0.225
			Havai	0.915	0.896	0.167		
			Quênia	0.800	1.000	0.000		
			Global	0.855	0.902	0.079		
Reg. Logística	set_5	$\delta^{18}O, \delta^{15}N,$ $\delta^{13}C_{caf}, \delta^{18}O_{caf},$ $\delta^{15}N_{caf},$ ${}^{87}Sr/{}^{86}Sr$	Brasil	0.789	0.789	0.070	0.882	0.218
			Havai	0.894	0.913	0.138		
			Quênia	1.000	0.909	0.015		
			Global	0.894	0.871	0.074		
Reg. Logística	set_{10}	$\delta^{34}S, \delta^{13}C_{caf},$ $\delta^{18}O_{caf},$ $\delta^{15}N_{caf},$ ${}^{87}Sr/{}^{86}Sr$	Brasil	0.750	0.833	0.054	0.868	0.252
			Havai	0.913	0.875	0.200		
			Quênia	0.900	0.900	0.015		
			Global	0.854	0.869	0.090		
RMLR ($\lambda = 0.2$)	set_9 Padronizado		Brasil	0.750	0.938	0.018	0.909	0.171
			Havai	0.957	0.900	0.167		
			Quênia	1.000	0.909	0.015		
			Global	0.902	0.916	0.066		
RMLR ($\lambda = 0.1$)	set_9 Padronizado		Brasil	0.750	0.882	0.018	0.896	0.194
			Havai	0.936	0.898	0.167		
			Quênia	1.000	0.909	0.015		
			Global	0.895	0.896	0.072		
SMLR ($\lambda = 0.5$)	set_9 Padronizado		Brasil	0.750	0.882	0.035	0.896	0.194
			Havai	0.936	0.898	0.167		
			Quênia	1.000	0.909	0.015		
			Global	0.895	0.896	0.072		
SMLR ($\lambda = 5$)	set_9 Padronizado		Brasil	0.750	0.882	0.035	0.896	0.201
			Havai	0.957	0.882	0.200		
			Quênia	0.900	1.000	0.000		
			Global	0.869	0.922	0.078		

Ao analisar por origem geográfica, verifica-se que é na QDA clássica (set_9) que o Brasil apresenta melhor desempenho, com 0.944 de *recall*, 0.850 de precisão e 0.018 de *FPR*. O Havai é a origem com os resultados mais consistentes, apresentado em todos os classificadores bons desempenhos, com valores em torno de 0.9 de *recall* e precisão, no entanto é a classe com os piores valores de *FPR*. Em relação aos melhores classificadores para o Havai, destaca-se o kNN aplicado nos *scores* da ACP do set_{10} (com $k = 3$ e três componentes principais), com 0.891 de *recall*, 0.953 de precisão e 0.067 de *FPR*. Por fim o Quênia que, apesar do reduzido número de observações, é uma região produtora que se conseguiu diferenciar facilmente das restantes, sendo o melhor classificador para o Quênia o kNN aplicado nos *scores* da ACP do set_{10} (com $k = 3$ e três componentes principais), onde nenhuma observação foi mal classificada e por isso apresenta o valor ótimo em todos os indicadores de desempenho.

Globalmente, verifica-se que o kNN, apesar de ser um método simples, apresenta bons resultados. Os classificadores com melhor desempenho foram: o kNN nos *scores* da ACP do set_{10} ($k = 3$), com o menor valor na medida global de desempenho, $Dist.Otimo = 0.154$; o RMLR no set_9 com $\lambda = 0.2$; e o kNN nos *scores* da ACP do set_9 ($k = 6$).

Os resultados obtidos para o melhor classificador para cada método de classificação são analisados de seguida. Serão também alvo de análise algumas árvores de decisão associadas a classificadores que, embora não tenham sido todos referidas na Tabela 5.3 e não estarem entre os melhores classificadores, são uma ferramenta útil para a compreensão dos dados.

5.3.2 Árvores de decisão

O método de árvores de decisão foi o mais explorado, no entanto não foi o que conduziu aos melhores desempenhos. Em termos de desempenho dos classificadores, a utilização dos dados sem retirar valores omissos não levou a melhorias; em particular, registaram-se algumas observações que passam as condições de teste de forma aleatória, por terem valor omissos nas variáveis associadas. Além disso, as árvores tornaram-se mais complexas (o número de nós é maior), por exemplo o caso de três classes, {BZ, HW, KE}, no set_6 .

O conjunto de classes com a divisão do arquipélago do Havai em 2 classes também não conduziu a melhorias no desempenho, piorando em alguns casos. Por exemplo, algumas observações do arquipélago do Havai foram mal classificadas entre as duas classes correspondente. Um dos casos em que o desempenho do classificador piora, em relação ao equivalente sem a divisão do arquipélago do Havai em duas classes, é o que utiliza o set_3 sem valores omissos, em que para o conjunto de classes {BZ, HW_r, HW_i, KE} se obteve $Dist.Otima = 0.360$, enquanto que para {BZ, HW, KE} o valor foi $Dist.Otima = 0.274$. Relembre que HW_i representa a classe com as observações da ilha Havai e HW_r a classe com as observações das restantes ilhas do arquipélago do Havai.

A maioria das árvores associadas aos modelos encontrados têm como nó inicial (raiz) a variável $^{87}Sr/^{86}Sr$ ou $\delta^{18}O_{caf}$ (como é o caso das árvores das figuras 5.1 e 5.2, respetivamente), sugerindo a importância destas variáveis na diferenciação geográfica do grão verde de café. Refira-se ainda que

as árvores obtidas nas análises sem valores omissos e sem a Etiópiã dos set_3 , set_9 e set_{10} são muito semelhantes, sendo as dos últimos dois subconjuntos iguais e a do set_3 apenas difere nos intervalos de valores de um dos nós.

A árvore de decisão obtida para o set_3 , sem valores omissos e sem a Etiópiã, encontra-se na Figura 5.1. A árvore associada à divisão do Havai em 2 classes, no set_6 sem valores omissos e sem a Etiópiã, foi representada na Figura 5.2. Estas árvores foram escolhidas para análise tendo em conta não só o desempenho dos classificadores, mas também a pertinência da sua interpretação. Na Tabela 5.4 apresenta-se o desempenho das árvores selecionadas.

Tabela 5.4: Medidas de desempenho das árvores de decisão (C5.0) selecionadas, sem e com divisão das regiões produtoras do arquipélago do Havai em $Havai_i$ (ilha Havai) e $Havai_r$ (restantes regiões do arquipélago do Havai), nos set_3 e set_6 (respetivamente) sem valores omissos.

Classificador	Subconjunto	Variáveis	Classe	Recall	Precisão	FPR	Accy	Dist. Ótimo
C5.0	set_3	$C\%, N\%, \delta^{13}C,$	Brasil	0.842	0.800	0.071		
		$\delta^{18}O, \delta^{15}N, \delta^{34}S,$	Havai	0.891	0.911	0.138		
		$\delta^{13}C_{caf}, \delta^{18}O_{caf},$	Quénia	0.800	0.800	0.031		
		$\delta^{15}N_{caf}, {}^{87}Sr/{}^{86}Sr$	Global	0.844	0.837	0.080	0.867	0.274
C5.0	set_6	$C\%,$	Brasil	0.842	0.800	0.070		
		$N\%,$	Havai _r	0.773	0.810	0.074		
		$\delta^{13}C_{caf},$	Havai _i	0.840	0.840	0.078		
		$\delta^{18}O_{caf},$	Quénia	0.900	0.900	0.015		
		$\delta^{15}N_{caf}$	Global	0.839	0.837	0.059	0.829	0.292

Na árvore de decisão da Figura 5.1, obtida a partir do set_3 sem valores omissos, com as classes do Brasil, arquipélago do Havai e Quénia, a primeira condição de teste (raiz – nó 1) diz respeito à variável ${}^{87}Sr/{}^{86}Sr$, separando logo cerca de 70% das observações do Brasil das restantes, por apresentarem elevado valor nesta variável (maior que 0.7091).

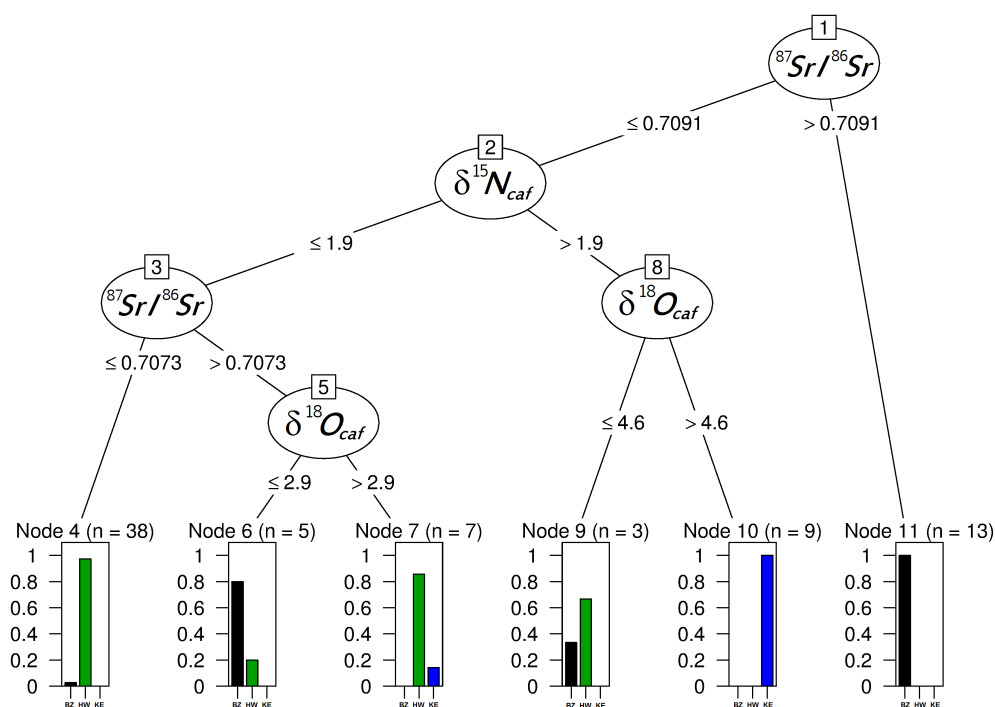


Figura 5.1: Árvore de decisão (C5.0) obtida a partir do set_3 , sem valores omissos (Brasil, BZ, a preto; arquipélago do Havai, HW, a verde; e Quénia, KE, a azul).

Entre as observações com ${}^{87}Sr/{}^{86}Sr$ menor ou igual que 0.7091, a maioria das observações do

Quénia (90%) são bem classificadas ao passar pelos nós 2 e 8, que têm condições de teste sobre $\delta^{15}N_{caf}$ e $\delta^{18}O_{caf}$, respetivamente, por apresentarem valores elevados nestas duas variáveis (maiores que 1.9 e 4.6, respetivamente).

A maioria das observações do Havai (80%) passam a condição de teste do nó 2 e do nó 3 por apresentarem valores reduzidos tanto em $\delta^{15}N_{caf}$ como em $^{87}Sr/^{86}Sr$ (menor ou igual a 1.9 e 0.7073, respetivamente).

Após estes dois nós, para as observações com $^{87}Sr/^{86}Sr$ maior que 0.7073, a árvore apresenta um nó com condição sobre $\delta^{18}O_{caf}$ (nó 5), este nó separa razoavelmente as observações do Brasil das observações do arquipélago do Havai, tendo as primeiras valores nesta variável reduzidos (menor ou igual a 2.9).

A árvore de decisão da Figura 5.2, obtida a partir do set_6 sem valores omissos e com divisão do arquipélago do Havai em duas classes, apresenta condição de teste sobre $\delta^{18}O_{caf}$ na raiz (nó 1). Este nó inicial dividiu as observações do Brasil e a maioria das observações da ilha Havai (87%), das observações do Quénia e da maioria das observações das restantes ilhas do Havai (86%). As primeiras por apresentarem valores em $\delta^{18}O_{caf}$ menores ou iguais a 4.8 e as segundas por apresentarem valores acima de 4.8 nesta variável.

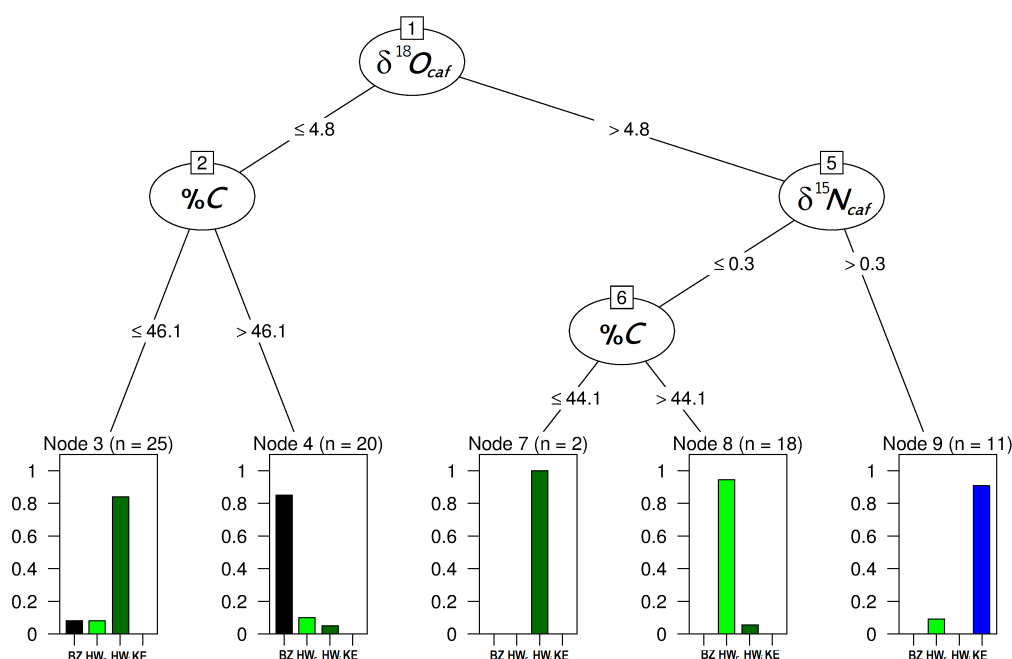


Figura 5.2: Árvore de decisão (C5.0) obtida a partir do set_6 , sem valores omissos e com divisão do arquipélago do Havai em duas classes (ilha Havai, HW_i, a verde escuro; restantes ilhas do arquipélago do Havai, HW_r, a verde claro; Brasil, BZ, a preto; e Quénia, KE, a azul).

O nó 2 apresenta condição de teste sobre a variável %C, separando a maioria das observações do Brasil das observações do Havai, por apresentar valores mais elevados.

Relativamente ao nó 5 e às observações com $\delta^{18}O_{caf}$ maior que 4.8, este nó dividiu as observações do Quénia da maioria das observações do arquipélago do Havai, graças à condição referente a $\delta^{15}N_{caf}$, sendo que as observações do Quénia se distinguiram pelos valores acima de 0.3 em $\delta^{15}N_{caf}$.

Note-se que a árvore com divisão do Havai em duas classes (Figura 5.2) é simples e tem bom

desempenho, quando comparada com as restantes árvores de decisão encontradas.

A análise das árvores de decisão é importante do ponto de vista de exploração dos dados, pois permite encontrar características das diferentes origens geográficas. Algumas já encontradas nas análises realizadas nos capítulos anteriores. De seguida enumeram-se as principais conclusões desta análise.

1. As observações do Brasil caracterizam-se por apresentarem os valores mais elevados em $^{87}Sr/^{86}Sr$ e reduzidos em $\delta^{18}O_{caf}$.
2. As observações do Quênia apresentam, em geral, valores elevados em $\delta^{15}N_{caf}$ e em $\delta^{18}O_{caf}$.
3. As observações do Havai caracterizam-se pelos valores reduzidos em $^{87}Sr/^{86}Sr$ e elevados em $\delta^{18}O_{caf}$. Mais ainda, as observações da ilha Havai distinguem-se das restantes regiões do arquipélago do Havai principalmente por apresentarem valores mais reduzidos em $\delta^{18}O_{caf}$.

5.3.3 K vizinhos mais próximos

A ACP que serve de base para o classificador analisado nesta subsecção é a ACP com as primeiras três componentes principais retidas obtida a partir do set_{10} e que foi analisada na Subsecção 3.3.2. Este classificador, obtido a partir dos *scores* da ACP do set_{10} , possui o melhor desempenho global a julgar pelos resultados da Tabela 5.3.

Como se referiu anteriormente, o kNN não constrói um modelo explícito. Por esta razão, a discussão dos resultados do kNN centra-se na matriz de confusão apresentada na Tabela 5.5.

Tabela 5.5: Matriz de confusão obtida após *leave-one-out* do classificador kNN, com $k = 3$, aplicado aos *scores* da ACP do set_{10} .

		Classificação		
		Brasil	Havai	Quênia
Real	Brasil	18	2	0
	Havai	5	41	0
	Quênia	0	0	10

Pode afirmar-se que o classificador confunde principalmente o Brasil com o Havai (as sete observações mal classificadas correspondem a trocas entre estas duas origens geográficas). Algumas das observações mal classificadas correspondem a observações que no agrupamento encontrado no capítulo 4 foram *mal alocados*, nomeadamente as observações com ID 4, ID 75 e ID 97. As restantes observações mal classificadas têm IDs 37, 72, 76 e ID 86; à exceção da observação com ID 37, tratam-se de observações do Havai que foram classificadas como originárias do Brasil.

5.3.4 Análise discriminante quadrática

O classificador de QDA clássica obtido a partir do set_9 foi o que obteve melhor desempenho entre todos os classificadores de análise discriminante analisados, com $Dist.Otimo = 0.204$. De facto a estimação clássica conduziu a melhores medidas globais de desempenho do classificador do que a versão robusta, associada a $Dist.Otimo = 0.225$. A regra de classificação inerente a esta análise discriminante tira partido das médias aritméticas por classe e matrizes de covariâncias para classificar uma nova observação.

Tabela 5.6: Intervalos de confiança a 95% para a média de cada variável do set_9 .

	$^{87}Sr/^{86}Sr$	$\delta^{13}C_{caf}$	$\delta^{18}O_{caf}$	$\delta^{15}N_{caf}$
Brasil	[0.7111, 0.7166]	[-28.145, -27.455]	[0.529, 2.261]	[0.742, 2.518]
Havai	[0.7061, 0.7069]	[-27.422, -26.459]	[3.485, 5.132]	[-0.661, 0.052]
Quênia	[0.7064, 0.7072]	[-26.681, -26.379]	[7.029, 9.291]	[1.865, 3.175]

Para comparar as médias das variáveis entre cada classe foram calculados intervalos de confiança a 95% para a média, recorrendo ao teste de *t-Student* (Johnson e Wichern, 2007, p. 210–226). Estes intervalos de confiança constam da Tabela 5.6. A média que se destaca pela diferença entre classes é a de $\delta^{18}O_{caf}$; tendo em conta a reduzida variância de $^{87}Sr/^{86}Sr$ (como consta da Tabela 3.4), destaca-se ainda a diferença entre a média de $^{87}Sr/^{86}Sr$ no Brasil e a média das restantes regiões; por fim, a média de $\delta^{15}N_{caf}$ no Havai revelou-se a mais reduzida.

No que diz respeito à matriz de confusão da Tabela 5.7, verifica-se que a maioria das observações mal classificadas encontram-se nas observações atribuídas ao Havai. Cruzando as observações mal classificadas com o agrupamento encontrado no Capítulo 4, verifica-se que 3 das observações mal classificadas foram também *mal alocadas* no agrupamento, nomeadamente a observação com ID 4 do Brasil e ID 34 do Quênia, ambas classificadas na classe do Havai, e a observação com ID 75 do Havai classificada como originária do Brasil.

Tabela 5.7: Matriz de confusão obtida após *leave-one-out* do modelo de classificação de QDA clássica no set_9 .

		Classificação		
		Brasil	Havai	Quênia
Real	Brasil	17	3	0
	Havai	1	46	0
	Quênia	0	3	7

5.3.5 Regressão logística

O classificador estimado a partir de regressão logística multinomial aplicada no set_5 , sem valores omissos e sem as observações da Etiópia, apresentou desempenho global $Dist.Otimo = 0.218$ (como pode constatar pela Tabela 5.3), sendo o de melhor desempenho entre os classificadores obtidos a partir deste método de classificação.

Os coeficientes estimados e razões *odds* para este classificador são apresentados na Tabela 5.8. De relembrar que, devido à condição de normalização (5.21), não se estimaram os coeficientes para uma das classes, nomeadamente a do Brasil. Portanto, para efeitos de classificação considera-se que os coeficientes para a classe do Brasil são 0. As observações com valores elevados nas composições isotópicas em estudo têm uma chance elevada de serem originárias do Quênia. Esta afirmação prende-se com os elevados valores na razão *odds* em praticamente todas as variáveis do set_5 . A variável que é exceção é $\delta^{15}N$, em que valores elevados privilegiam a classificação em Havai.

Relativamente às observações mal classificadas na validação *leave-one-out*, apresenta-se na Tabela 5.9 a matriz de confusão do classificador em estudo. Verifica-se que oito das nove observações mal classificadas devem-se a trocas entre o Brasil e o Havai. As observações mal classificadas têm IDs 4, 36, 40, 49, 68, 75, 80, 86 e ID 99. As observações com ID 4, ID 75 e ID 99 foram *mal alocadas* no agrupamento encontrado no Capítulo 4.

Tabela 5.8: Estimativas dos coeficientes e razões *odds* da regressão logística obtida a partir do *set*₅, sem valores omissos.

Variáveis	Coeficientes		Razões <i>odds</i>	
	Havai	Quênia	Havai	Quênia
$(\hat{\beta}_0)$	847.3	4.551		
$\delta^{18}O$	0.144	93.15	1.155	2.852×10^{40}
$\delta^{15}N$	1.355	-147.4	3.877	0.000
$^{87}Sr/^{86}Sr$	-1192.2	2.666	0.000	14.381
$\delta^{13}C_{caf}$	0.366	173.4	1.442	2.016×10^{75}
$\delta^{18}O_{caf}$	0.449	197.4	1.567	5.564×10^{85}
$\delta^{15}N_{caf}$	-0.537	365.6	0.585	5.991×10^{158}

Tabela 5.9: Matriz de confusão obtida após *leave-one-out* do modelo de classificação de regressão logística aplicado ao *set*₅.

	Classificação		
	Brasil	Havai	Quênia
Real			
Brasil	15	4	0
Havai	4	42	1
Quênia	0	0	10

5.3.6 Regressão logística multivariada com incorporação de seleção de variáveis – SMLR e RMLR

Os métodos de classificação SMLR e RMLR foram aplicados aos subconjuntos *set*₃, *set*₇, *set*₉ e *set*₁₀ quer com dados padronizados, quer com dados não padronizados, tal como se referiu na Secção 5.3. Em relação aos parâmetros escolhidos, testaram-se diferentes valores de viés, nomeadamente 0, 0.5, 1, 2, 5 e 8. No entanto, o desempenho dos classificadores foi idêntico para viés diferente de 0; no caso de viés nulo o desempenho foi pior. Assim, decidiu-se apresentar os resultados obtidos, nomeadamente os coeficientes da regressão logística, com viés igual a 1. Os valores considerados para o parâmetro de regularização λ foram 0.1, 0.2, 0.5, 0.8, 1 e 5.

A consulta da Tabela 5.3 leva a concluir que entre estes classificadores o RMLR no *set*₉ padronizado, sem valores omissos, sem a Etiópia e com $\lambda = 0.2$, é o que apresenta melhor desempenho global, com *Dist.Otimo* = 0.171.

A seleção de variáveis inerente aos métodos que estamos a analisar (SMLR e RMLR) apenas conduziu à exclusão de variáveis nos casos em que os dados não foram padronizados. As variáveis excluídas em cada subconjunto, considerando o conjunto de classes {BZ, HW, KE} foram (independentemente dos parâmetros):

- *set*₃: $N\%$ e $\delta^{15}N$;
- *set*₇: nenhuma;
- *set*₉: $^{87}Sr/^{86}Sr$;
- *set*₁₀: $^{87}Sr/^{86}Sr$.

Para a classificação do conjunto de classes com a divisão do Havai, {BZ, HW_r, HW_i, KE}, apenas se recorreu ao método SMLR com dados não padronizados, como se pode constatar pela Tabela 5.2. Nos classificadores com este conjunto de classes também ocorreu a exclusão de variáveis, nomeadamente:

- *set*₃: $N\%$, $\delta^{15}N$ e $^{87}Sr/^{86}Sr$;
- *set*₉: $^{87}Sr/^{86}Sr$;
- *set*₁₀: $^{87}Sr/^{86}Sr$.

Deste modo, verifica-se que a variável $^{87}Sr/^{86}Sr$ é o principal alvo de exclusão pela seleção de variáveis. No entanto, como se constatou anteriormente, esta variável é uma das principais

diferenciadoras das observações do Brasil. Esta exclusão pode ser justificado pela reduzida variância desta variável (ver Tabela 3.4).

Apesar da exclusão de variáveis nos classificadores com os dados não padronizados, estes obtiveram pior desempenho que os classificadores com os dados padronizados, cujo desempenho se encontra sumariado na Tabela 5.3 para os quatro melhores casos.

Apresenta-se na Tabela 5.10 a matriz de confusão obtida após validação *leave-one-out* do classificador RMLR no *set₉*, com padronização e $\lambda = 0.2$. Esta matriz permite verificar que o valor de *recall* reduzido na classe do Brasil se deveu à classificação errada de cinco observações desta origem na classe do Havai. Estas observações correspondem aos IDs 4, 36, 38, 40 e 49. Mais ainda, as observações com ID 75 e ID 99, ambas do Havai, foram classificadas nas classes do Brasil e do Quênia, respetivamente. Destas sete observações mal classificadas, três foram também *mal alocadas* no agrupamento obtido no Capítulo 4, designadamente as observações com IDs 4, 75 e 99.

Tabela 5.10: Matriz de confusão obtida após validação *leave-one-out* do modelo de RMLR ($\lambda = 0.2$) aplicado ao *set₉* com padronização, sem valores omissos e sem as observações da Etiópia.

	Classificação		
	Brasil	Havai	Quênia
Brasil	15	5	0
Havai	1	45	1
Quênia	0	0	10

A causa do valor reduzido de *Recall* do Brasil pode atribuir-se ao enviesamento dos estimadores devido à existência de *outliers* desta origem geográfica no conjunto de dados, como se verificou na Secção 3.4. A padronização dos dados a partir de um estimador robusto ou a eliminação dos *outliers* parece ser uma solução para este problema, solução esta não explorada neste trabalho por limitações de tempo.

Para finalizar, apresentam-se na Tabela 5.11 os coeficientes e razões *odds* estimadas para o modelo de classificação em estudo, sendo que um asterisco, *, distingue as variáveis padronizadas das originais.

Tabela 5.11: Estimativas dos coeficientes e razões *odds* do classificador RMLR ($\lambda = 0.2$) obtido a partir do *set₉* padronizado, sem valores omissos e sem as observações da Etiópia.

Variáveis	Coeficientes		Razão odds	
	Brasil	Havai	Brasil	Havai
(β_0)	1.462	2.544		
$^{87}Sr/^{86}Sr^*$	2.064	-1.710	7.877	0.181
$\delta^{13}C_{caf}^*$	-1.301	-0.916	0.272	0.400
$\delta^{18}O_{caf}^*$	-2.937	-1.721	0.053	0.179
$\delta^{15}N_{caf}^*$	-1.837	-2.218	0.159	0.109

À semelhança do que aconteceu na Subsecção 5.2.4, não se estimaram os coeficientes para uma das classes, nomeadamente a classe do Quênia, considerando-se para efeitos de classificação que os coeficientes são zero e as razões *odds* iguais a um.

Da análise das razões *odds* destaca-se a variável $^{87}Sr/^{86}Sr^*$ por ser a principal responsável pela chance de classificar as observações na classe do Brasil. Adicionalmente, a probabilidade de classificar uma observação na classe do Quênia aumenta com o aumento de $\delta^{13}C_{caf}^*$, de $\delta^{18}O_{caf}^*$ e de $\delta^{15}N_{caf}^*$.

5.3.7 Seleção do melhor classificador

Realizado o treino e validação dos classificadores, o próximo passo seria testar os melhores com um novo conjunto de dados. Contudo, não foi possível efetuar tal teste devido ao elevado custo inerente à recolha de novas observações.

Na escolha de um classificador de grãos verdes de café segundo a origem geográfica (designadamente as origens Brasil, Havai e Quénia), deve ter-se em consideração não só o seu desempenho como também outros fatores. Entre eles estão o custo associado à classificação de uma nova observação, por um lado em termos de custo de recolha de observações para cada variável, por outro lado em termos de custo computacional, a interpretação e a simplicidade da regra de classificação.

Com o objetivo de realizar uma comparação final, apresenta-se agora, na Tabela 5.12, o desempenho dos classificadores que se entende terem conduzido aos melhores resultados. Designadamente, o RMLR no set_9 padronizado, com $\lambda = 0.2$, e o kNN a partir dos *scores* da ACP do set_{10} , com $k = 3$, ambos sem valores omissos.

Tabela 5.12: Medidas de desempenho dos melhores classificadores, obtidas por *leave-one-out*, treino sem valores omissos e sem as observações da Etiópia.

Classificador	Subconjunto	Variáveis	Classe	Recall	Precisão	FPR	Accy	Dist. Ótimo
RMLR ($\lambda = 0.2$)	set_9 Padronizado	$\delta^{13}C_{caf}$,	Brasil	0.750	0.938	0.018	0.909	0.171
		$\delta^{15}N_{caf}$,	Havai	0.957	0.900	0.167		
		$\delta^{18}O_{caf}$,	Quénia	1.000	0.909	0.015		
		$^{87}Sr/^{86}Sr$	Global	0.902	0.916	0.066		
kNN ($k = 3$)	ACP set_{10} 3 CPs		Brasil	0.900	0.783	0.089	0.908	0.154
			Havai	0.891	0.953	0.067		
			Quénia	1.000	1.000	0.000		
			Global	0.930	0.912	0.052		

Eis algumas das características destes dois classificadores aqui selecionados e analisados previamente nas subsecções 5.3.6 e 5.3.3.

- kNN nos *scores* da ACP do set_{10} : o kNN não constrói um modelo de classificação e baseia-se em distâncias entre as observações, elevando o custo computacional e a dificuldade de interpretação e validação. Para a classificação de uma nova observação, exige a medição de todas as variáveis do set_{10} (cinco variáveis) apesar de apenas se usarem três componentes principais. A principal vantagem deste classificador é a sua simplicidade.
- RMLR no set_9 : obteve o pior desempenho na classe do Brasil, com 0.750 de *recall*. Uma das vantagens deste classificador é a facilidade de classificar uma nova observação, a partir da regra de classificação. Outra das vantagens face ao kNN é o número de variáveis associadas, pois do set_9 apenas fazem parte quatro variáveis, enquanto que no set_{10} fazem parte cinco.

No seguimento da afirmação anterior, a escolha entre um classificador que tira partido do conjunto de variáveis do set_9 e um que tira partido do set_{10} é pertinente, apesar da diferença ser apenas de uma variável ($\delta^{34}S$), ou não fossem as técnicas analíticas diferentes. A adição de uma variável eleva significativamente o número de procedimentos e os custos associados à medição dos valores das variáveis.

O número de observações mal classificadas foi sete em ambos os classificadores. Destas observações, a maioria deve-se a confusão dos classificadores entre as classes do Brasil e do Havai.

É necessário referir que o decréscimo de 9% no valor da precisão da classe do Quénia, no RMLR, se deve apenas a uma observação mal classificada e se dispõe apenas de 10 observações do Quénia sem valores omissos no set_9 .

Com efeito, o classificador escolhido para o conjunto de classes constituído pelo Brasil, Havai e Quénia, $\{BZ, HW, KE\}$, é o RMLR estimado com as variáveis do set_9 , dados padronizados e sem valores omissos.

Para concluir apresenta-se de seguida a regra de classificação associada a este classificador.

Regra de classificação: Dada uma nova observação, com valores $\mathbf{x}_0 = (x_{01}, x_{02}, x_{03}, x_{04})^t$, nas variáveis $^{87}Sr/^{86}Sr$, $\delta^{13}C_{caf}$, $\delta^{18}O_{caf}$ e $\delta^{15}N_{caf}$, respetivamente, padronizam-se as observações da seguinte forma:

$$\mathbf{x}_0^* = \left(\frac{x_{01} - 0.70845}{0.00446}, \frac{x_{02} + 27.11039}{1.39770}, \frac{x_{03} - 4.05195}{3.16708}, \frac{x_{04} - 0.56494}{1.77777} \right)^t. \quad (5.30)$$

Define-se

$$\hat{\beta}_1^t \mathbf{x}_0^* = 1.462 + 2.064 x_{01}^* - 1.301 x_{02}^* - 2.937 x_{03}^* - 1.837 x_{04}^* \quad (5.31)$$

$$\hat{\beta}_2^t \mathbf{x}_0^* = 2.544 - 1.710 x_{01}^* - 0.916 x_{02}^* - 1.721 x_{03}^* - 2.218 x_{04}^* \quad (5.32)$$

e calcula-se

$$\hat{p}_1 = \hat{P}(\text{classe ser o Brasil} \mid \mathbf{x}_0^*, \hat{\beta}) = \frac{\exp(\hat{\beta}_1^t \mathbf{x}_0^*)}{\exp(\hat{\beta}_1^t \mathbf{x}_0^*) + \exp(\hat{\beta}_2^t \mathbf{x}_0^*) + 1} \quad (5.33)$$

$$\hat{p}_2 = \hat{P}(\text{classe ser o Havai} \mid \mathbf{x}_0^*, \hat{\beta}) = \frac{\exp(\hat{\beta}_2^t \mathbf{x}_0^*)}{\exp(\hat{\beta}_1^t \mathbf{x}_0^*) + \exp(\hat{\beta}_2^t \mathbf{x}_0^*) + 1} \quad (5.34)$$

$$\hat{p}_3 = \hat{P}(\text{classe ser o Quénia} \mid \mathbf{x}_0^*, \hat{\beta}) = 1 - (\hat{p}_1 + \hat{p}_2) \quad (5.35)$$

Por fim, atribui-se $\mathbf{x}_0$ à classe que maximiza $\hat{p}_1$, $\hat{p}_2$ e $\hat{p}_3$.

Capítulo 6

Comentários finais

O problema estudado ao longo deste trabalho foi a diferenciação da origem geográfica do café, recorrendo a análise da composição isotópica do grão verde. A necessidade desta diferenciação provém da crescente procura, por parte dos importadores de café, em desenvolver metodologias que permitam detetar adulterações do café ao longo da cadeia comercial, nomeadamente a nível de origem geográfica. Por sua vez, este interesse é justificado não só pela grande variabilidade do preço do café consoante a região onde o grão verde se desenvolveu, como também pelo crescente interesse por parte dos consumidores em aspetos relacionados com a autenticidade, qualidade, métodos de produção e origem dos alimentos, não sendo o café uma exceção.

Assim, para o estudo deste problema recorreu-se à análise estatística de um conjunto de dados provenientes de quatro regiões produtoras de café (Brasil, Etiópia, Havai e Quênia) caracterizadas por 10 variáveis (percentagens de oxigénio e de azoto; composições isotópicas do carbono, do azoto, do oxigénio e do enxofre no grão verde de café; a razão isotópica do estrôncio; e as composições isotópicas do carbono, do azoto e do oxigénio no composto específico do grão verde de café, a cafeína). Os dados em estudo caracterizam-se por um número elevado de valores omissos, uma vez que a generalidade dos valores observáveis são resultado de métodos químicos complexos e de custo elevado, o que impossibilita a recolha completa das observações. Para superar esta limitação foram considerados vários subconjuntos de variáveis de acordo com a análise e relevância das mesmas. Em geral, o número de observações em estudo em cada subconjunto ronda as 76 observações. Este é um valor modesto do ponto de vista estatístico, mas constitui uma amostra de dimensão elevada em comparação com outros trabalhos nesta área de aplicação.

As percentagens de azoto e oxigénio revelaram-se à partida pouco interessantes do ponto de vista do problema em estudo. Por outro lado, as composições isotópicas revelaram o seu carácter diferenciador da origem geográfica do grão verde de café, tal como era expectável entre os peritos da área.

As composições isotópicas do oxigénio ($\delta^{18}O_{caf}$) e do azoto ($\delta^{15}N_{caf}$) no composto específico do grão verde de café, a cafeína, e a razão isotópica do estrôncio ($^{87}Sr/^{86}Sr$) foram as que demonstraram maior potencial diferenciador das origens geográficas do café. É importante salientar que as diversas

análises realizadas neste trabalho levam a concluir que a composição isotópica da cafeína presente no grão verde demonstra maior carácter diferenciador da origem geográfica que a composição isotópica do grão. Adicionalmente, a composição isotópica do enxofre ($\delta^{34}S$) mostrou-se interessante do ponto de vista de diferenciação de microrregiões, como foi o caso analisado da ilha Havai *versus* as restantes ilhas do arquipélago do Havai.

O fator que pode ser indicado como principal agente de variabilidade das composições isotópicas entre diferentes origens geográficas é o clima, que é uma característica intrínseca de cada região. Contudo, fenómenos climatéricos podem influenciar fortemente o valor da composição isotópica do oxigénio, levando a que amostras da mesma região mas de colheitas diferentes possam apresentar valores díspares nas composições isotópicas (esta evidência foi encontrada na Secção 3.4). Fatores como a altitude ou distância ao mar foram os que mais se distinguiram na diferenciação de microrregiões, nomeadamente das regiões do Havai. As características do solo onde a planta de café se desenvolveu podem também ser apontadas como diferenciadoras da origem geográfica de grãos verdes de café, por se refletirem em $^{87}Sr/^{86}Sr$ e $\delta^{15}N_{caf}$.

Em especial, a composição isotópica do azoto na cafeína do grão mostrou-se importante na diferenciação das amostras provenientes do Quénia.

A aplicação de métodos de classificação, descrita no Capítulo 5, permitiu obter classificadores para a origem geográfica de grãos verdes de café com elevado desempenho, considerando as regiões produtoras do Havai, Brasil e Quénia. O classificador selecionado é um caso particular da regressão logística multinomial à qual se adiciona um termo que promove a contração dos coeficientes de regressão logística de forma a realizar seleção de variáveis automática (RMLR). Este classificador, que foi descrito na Subsecção 5.3.7, recorre à composição isotópica da molécula de cafeína do grão (nomeadamente, do oxigénio, do carbono e do azoto) e à razão isotópica do estrôncio para classificar, estimando-se que classifica corretamente 90.9% das observações.

Uma limitação do conjunto de dados em estudo é o elevado número de valores omissos, que levou à exclusão de algumas observações do conjunto de dados. Ainda assim, este trabalho destaca-se pelo elevado número de observações, consideradas nas análise estatísticas efetuadas, e pelo elevado nível de informação acerca da localização geográfica das plantações onde se desenvolveram os grãos recolhidos para as amostras. Pois, para a recolha de amostras da composição isotópica do grão verde de café são necessários diversos recursos de difícil acesso.

Outra das limitações deste estudo é o facto da análise da composição isotópica do grão verde de café ser muito recente, dificultando a interpretação e validação de relações existentes entre as diferentes composições isotópicas do grão verde, bem como entre fatores como o clima, a altitude, a composição do solo, a estação e ano de colheita do grão, etc.

Estes dados apesar das limitações representam uma amostra de grande valor no estudo do grão verde de café, dada a sua originalidade nesta área. Não obstante as variadas análises exploradas no âmbito deste trabalho muito ficou por fazer. Em seguida listam-se alguns dos aspetos que se consideram mais promissores para desenvolver em trabalho futuro.

1. Valores omissos: muitos dos métodos de análise de valores omissos baseiam-se na imputação

de estimativas nos valores omissos (e.g., médias ou medianas), seguindo-se a abordagem tradicional. Esta estratégia pode levar ao enviesamento de estimativas obtidas e é bastante polémica. Ficaram por explorar alternativas, eventualmente robustas, para lidar de forma satisfatória com estas omissões.

2. Seleção de variáveis: a estratégia usada baseou-se na construção de subgrupos de variáveis, de acordo com a análise exploratória dos dados, os padrões de valores omissos e a informação *a priori* dos peritos. No entanto, classificadores como árvores de decisão e SMLR conduzem implicitamente a regras de classificação onde está subjacente a identificação das variáveis mais relevantes no processo de classificação. No entanto, outras estratégias podem ser adotadas, conduzindo eventualmente a melhores classificadores.
3. Métodos robustos: os métodos robustos utilizados neste trabalho concentraram-se na análise em componentes principais, deteção de *outliers* e análise discriminante robusta, não tendo esta última conduzido ao melhor classificador. Contudo, outras alternativas poderiam ter sido utilizadas, em particular métodos robustos que contemplem a existência de valores omissos.
4. Validação das hipóteses estatísticas: os compêndios de análise multivariada sugerem que num problema de classificação os dados sejam divididos em três subconjuntos disjuntos: um utilizado para estimar as regras de classificação (amostra de treino), outra para testar o classificador e estimar medidas de desempenho do mesmo (amostra de teste) e uma última amostra, dita de validação, para confirmar hipóteses estatísticas subjacentes aos modelos utilizados. No presente estudo, seleccionou-se como melhor regra de classificação uma baseada em regressão logística. Todavia, as hipóteses subjacentes (relação linear entre o $\ln odds$ e as variáveis, Johnson e Wichern, 2007, p. 635) não foram analisadas. Esta análise não foi realizada por falta de tempo e pela reduzida dimensão da amostra. Assim, o classificador proposto deve ser analisado e validado antes de ser utilizado pela Delta Cafés.
5. Novas amostras de grão verde de café: este trabalho seria enriquecido se existissem novas amostras de café que pudessem ser utilizadas para reforçar o bom desempenho do classificador e despistar potenciais problemas de sobre-ajuste do modelo aos dados (*overfitting*).

Espera-se com este trabalho ter contribuído para chamar a atenção dos peritos da área para a relevância de utilizar métodos automáticos de classificação dos grãos de café segundo a origem geográfica, e com estes métodos colocar ao dispor da empresa ferramentas auxiliares aos procedimentos usualmente utilizados pela mesma.

Referências

- Aggarwal, P. K., L. J. Araguás-Araguás, M. Groening, K. M. Kulkarni, T. Kurttas, B. D. Newman e T. Vitvar (2010). Global Hydrological Isotope Data and Data Networks. *Isoscapes: Understanding movement, pattern, and process on Earth through isotope mapping*. Ed. por J. B. West, G. J. Bowen, T. E. Dawson e K. P. Tu. New York: Springer.
- Amundson, R., A. T. Austin, E. A. G. Schuur, K. Yoo, V. Matzek, C. Kendall, A. Uebersax, D. Brenner e W. T. Baisden (2003). Global patterns of the isotopic composition of soil and plant nitrogen. *Global Biogeochemical Cycles* **17**, 1031, 31–1 –31–10.
- Armstrong, W. (2013). *Versão do pacote rrcov (Pca.R)*. URL: <https://github.com/armstrtw/rrcov/blob/master/R/Pca.R> (acedido em 06/2015).
- Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. New York: Springer.
- Bolón-Canedo, V., N. Sánchez-Maróño e A. Alonso-Betanzos (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems* **34**, 483–519.
- Bosko, J., M. Edwards e A. Hartemink (2007). *SMLR (Sparse Multinomial Logistic Regression) Version 1.1.0*. URL: <https://www.cs.duke.edu/~amink/software/smlr/>.
- Bowen, G. J. e J. Revenaugh (2003). Interpolating the isotopic composition of modern meteoric precipitation. *Water Resources Research* **39**, SWC91–SWC913.
- Cadima, J. e I. T. Jolliffe (1995). Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics* **22**, 203–214.
- Chandrashekar, G. e F. Sahin (2014). A survey on feature selection methods. *Computers & Electrical Engineering* **40**, 16–28.
- ClubeCafé (2015). *História do café*. URL: <http://www.clubecafe.net.br/historia-cafe#historia> (acedido em 11/06/2015).
- Crem B.V. e J. Pierrot (2015). *Product Factsheet Coffee in Belgium*. CBI Market Intelligence. URL: www.cbi.eu/sites/default/files/study/product-factsheet-coffee-belgium-2014.pdf.
- Croux, C., P. Filzmoser e M. Oliveira (2007). Algorithms for Projection–Pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **87**, 218–225.
- Croux, C. e A. Ruiz-Gazen (2005). High breakdown estimators for principal components: the projection–pursuit approach revisited. *Journal of Multivariate Analysis* **95**, 206–226.
- Dawson, T. E. e R. Siegwolf (2007). *Stable Isotopes as Indicators of Ecological Change*. Terrestrial Ecology. Oxford: Elsevier Science.

- Delta Cafés (2011). *Relatório de sustentabilidade*.
- (2015). *O café – Da planta à chavena*. URL: <http://www.delta-cafes.pt/pt/o-cafe/da-planta-a-chavena/> (acedido em 05/2015).
- Faria, R., M. Gomes, D. Epstein e I. White (2014). A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *PharmacoEconomics* **32**, 1157–1170.
- Filzmoser, P., R. Maronna e M. Werner (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis* **52**, 1694–1711.
- Folmer, B. (2014). “How can science help to create new value in coffee?” **63, Part C**. 2nd Conference on Coffee Cocoa and Tea Science (CoCoTea2013), 477–482.
- Gharibnezhad, F., L. E. Mujica e J. Rodellar (2015). Applying robust variant of Principal Component Analysis as a damage detector in the presence of outliers. *Mechanical Systems and Signal Processing* **50–51**, 467–479.
- Ghasemi, A. e S. Zahediasl (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism* **10**, 486–489.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–871.
- Halkidi, M., Y. Batistakis e M. Vazirgiannis (2001). On clustering validation techniques. *Journal of Intelligent Information Systems* **17**, 107–145.
- Hand, D. J., P. Smyth e H. Mannila (2001). *Principles of Data Mining*. Cambridge: MIT Press.
- ICO (2012). “Reexportações de café”. Conselho Internacional do Café, 109ª sessão. Organização Internacional do Café. Londres. URL: www.ico.org/documents/icc-109-2-r1p-re-exports.pdf.
- (2015). *A história do café*. Organização Internacional do Café. URL: http://www.ico.org/pt/coffee_storyp.asp (acedido em 05/2015).
- Inácio, C. T., P. M. Chalk e A. M. Magalhães (2015). Principles and limitations of stable isotopes in differentiating organic and conventional foodstuffs: 1. plant products. *Critical Reviews in Food Science and Nutrition* **55**, 1206–1218.
- Johnson, R. A. e D. W. Wichern (2007). *Applied Multivariate Statistical Analysis*. 6ª ed. Upper Saddle River: Pearson Prentice Hall.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. 2ª ed. New York: Springer.
- Kaufman, L. e P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.
- Keidel, A., D. von Stetten, C. I. I. Rodrigues, C. Máguas e P. Hildebrandt (2010). Discrimination of green Arabica and Robusta coffee beans by raman spectroscopy. *Journal of Agricultural and Food Chemistry* **58**, 11187–11192.
- Kelly, S., K. Heaton e J. Hoogewerff (2005). Tracing the geographical origin of food: the application of multi-element and multi-isotope analysis. *Trends in Food Science & Technology* **16**, 555–567.
- Krishnapuram, B., L. Carin, M. A. T. Figueiredo e A. J. Hartemink (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 957–968.

- Lal, T., O. Chapelle, J. Weston e A. Elisseeff (2006). *Embedded Methods. Feature Extraction*. Ed. por I. Guyon, M. Nikravesh, S. Gunn e L. Zadeh. **207**. Studies in Fuzziness and Soft Computing. Berlin: Springer.
- Liu, H.-C., C.-F. You, C.-Y. Chen, Y.-C. Liu e M.-T. Chung (2013). Geographic determination of coffee beans using multi-element analysis and isotope ratios of boron and strontium. *Food Chemistry* **142**, 439–445.
- Luykx, D. M. e S. M. van Ruth (2008). An overview of analytical methods for determining the geographical origin of food products. *Food Chemistry* **107**, 897–911.
- Marengo, J., R. Schaeffer, D. Zee e H. S. Pinto (2010). Mudanças Climáticas e Eventos Extremos no Brasil. Ed. por FBDS. URL: http://www.fbds.org.br/cop15/FBDS_MudancasClimaticas.pdf.
- Maronna, R. A., D. R. Martin e V. J. Yohai (2006). *Robust Statistics: Theory and Methods*. London: John Wiley e Sons.
- McLeod, R., M. Garland, R. Hale, S. Steiman e R. Frew (2013). Determining the most effective combination of chemical parameters for differentiating the geographic origin of food products: an example using coffee beans. *Journal of Food Chemistry and Nutrition* **1**, 49–61.
- Moldvaer, A. (2014). *Coffee Obsession*. London: DK Publishing.
- Mosteller, F. e J. W. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. London: Addison-Wesley.
- Nanni, L., A. Lumini e S. Brahnam (2012). A classifier ensemble approach for the missing feature problem. *Artificial Intelligence in Medicine* **55**, 1, 37–50.
- Narlikar, L. e A. J. Hartemink (2006). Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics* **22**, 157–163.
- nca (2015). *Coffee Around the World*. National Coffee Association of U.S.A., Inc. URL: <http://www.ncausa.org/About-Coffee/Coffee-Around-the-World> (acedido em 11/2015).
- Park, H. A. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing* **43**, 2, 154–164.
- Pascoal, C., M. R. de Oliveira, R. Valadas, P. Filzmoser, P. Salvador e A. Pacheco (2012). “Robust feature selection and robust PCA for internet traffic anomaly detection”. *Proceedings of the IEEE INFOCOM 2012, Orlando, FL, USA, March 25-30, 2012*, 1755–1763.
- Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. London: Morgan Kaufmann.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <http://www.R-project.org/>.
- Rodrigues, C., R. Maia, M. Ribeirinho, P. Hildebrandt, L. Gautz, T. Prohaska e C. Máguas (2013). Chapter 22 - Coffee. *Food Protected Designation of Origin Methodologies and Applications*. Ed. por M. de la Guardia e A. González. **60**. Comprehensive Analytical Chemistry. Oxford: Elsevier.
- Rodrigues, C. I. I. (2011). *Geographical Origin Discrimination of The Green Coffee Bean and Analytical Qualification of The Roasting Profiles*. Tese de doutoramento. Universidade de Lisboa, Portugal.

- Rodrigues, C. I. I., R. Maia, M. Miranda, M. Ribeirinho, J. Nogueira e C. Máguas (2009). Stable isotope analysis for green coffee bean: a possible method for geographic origin discrimination. *Journal of Food Composition and Analysis* **22**, 463–471.
- Rodrigues, C. I. I., M. Brunner, S. Steiman, G. J. Bowen, J. M. F. Nogueira, L. Gautz, T. Prohaska e C. Máguas (2011a). Isotopes as tracers of the hawaiian coffee-producing regions. *Journal of Agricultural and Food Chemistry* **59**, 10239–10246.
- Rodrigues, C. I. I., C. Máguas e T. Prohaska (2011b). Strontium and oxygen isotope fingerprinting of green coffee beans and its potential to proof authenticity of coffee. *European Food Research and Technology* **232**, 361–373.
- RuleQuest Research (2013). *See5: An informal tutorial*. URL: www.rulequest.com/see5-win.html (acedido em 11/2015).
- Santato, A., D. Bertoldi, M. Perini, F. Camin e R. Larcher (2012). Using elemental profiles and stable isotopes to trace the origin of green coffee beans on the global market. *Journal of Mass Spectrometry* **47**, 1132–1140.
- Serra, F., C. G. Guillou, F. Reniero, L. Ballarin, M. I. Cantagallo, M. Wieser, S. S. Iyer, K. Héberger e F. Vanhaecke (2005). Determination of the geographical origin of green coffee by principal component analysis of carbon, nitrogen and boron stable isotope ratios. *Rapid Communications in Mass Spectrometry* **19**, 2111–2115.
- SICAL (2015). *Origens do Café*. URL: <http://www.sical.pt/#/origens/> (acedido em 06/2015).
- Tan, P.-N., M. Steinbach e V. Kumar (2005). *Introduction to Data Mining*. 1ª ed. Boston: Pearson Addison-Wesley.
- Todorov, V. e P. Filzmoser (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software* **32**, 1–47.
- Weckerle, B., E. Richling, S. Heinrich e P. Schreier (2002). Origin assessment of green coffee (Coffea Arabica) by multi-element stable isotope analysis of caffeine. *Analytical and Bioanalytical Chemistry* **374**, 886–890.

Apêndice A

Como complemento à análise das correlações amostrais clássicas realizada na Secção 3.2, apresentam-se na Secção A.1 as correlações amostrais robustas entre as variáveis do conjunto de dados em estudo.

Nas Secções A.2 e A.3 apresentam-se mapas onde se assinalam as localizações geográficas das observações em estudo. Na Secção A.2 apresentam-se figuras que diferenciam as observações classificadas como *outliers* na Secção 3.4. Na Secção A.3 apresentam-se figuras que localizam geograficamente as observações dos quatro grupos obtidos na análise de agrupamentos efetuada no Capítulo 4. Estes mapas foram obtidos através da aplicação *My Maps* do Google.

A.1 Análise descritiva

Na Tabela A.1 as correlações amostrais robustas dizem respeito a todo o conjunto de dados em estudo, ao passo que na Tabela A.2 referem-se ao Havai.

Estas correlações foram calculadas recorrendo ao MCD *Fast*, utilizando a rotina *CovMcd* do pacote *rrcov* do R. A **negrito** apresentam-se as correlações acima de 0.70 em valor absoluto.

Tabela A.1: Correlações amostrais robustas (Fast MCD).

	Latitude	Longitude	$\delta^{18}O_{prec}$	$C\%$	$N\%$	$\delta^{13}C$	$\delta^{18}O$	$\delta^{15}N$	$\delta^{34}S$	$^{87}Sr/^{86}Sr$	$\delta^{13}C_{ca,f}$	$\delta^{18}O_{ca,f}$	$\delta^{15}N_{ca,f}$
Altitude	-0.77	0.67	-0.98	-0.33	0.05	0.37	-0.64	-0.14	-0.74	-0.34	0.70	-0.63	-0.01
Latitude		-0.94	0.66	0.42	0.25	-0.45	0.81	0.10	0.79	0.32	-0.55	0.74	-0.08
Longitude			-0.54	-0.35	-0.34	0.36	-0.76	-0.11	-0.71	-0.10	0.44	-0.69	0.03
$\delta^{18}O_{prec}$				0.27	-0.13	-0.35	0.54	0.14	0.67	0.32	-0.69	0.56	0.04
$C\%$					0.29	-0.33	0.36	0.24	0.39	0.22	-0.41	0.23	-0.12
$N\%$						0.00	0.23	0.09	0.06	-0.03	0.10	0.15	0.01
$\delta^{13}C$							-0.28	0.00	-0.41	-0.19	0.66	-0.40	0.04
$\delta^{18}O$								-0.03	0.66	0.07	-0.38	0.71	-0.29
$\delta^{15}N$									0.03	0.19	-0.16	-0.05	0.66
$\delta^{34}S$										0.15	-0.69	0.65	-0.18
$^{87}Sr/^{86}Sr$											-0.26	0.07	0.27
$\delta^{13}C_{ca,f}$												-0.52	0.02
$\delta^{18}O_{ca,f}$													-0.27

Tabela A.2: Correlações amostrais robustas (Fast MCD) referentes ao Havai.

	Latitude	Longitude	$\delta^{18}O_{prec}$	$C\%$	$N\%$	$\delta^{13}C$	$\delta^{18}O$	$\delta^{15}N$	$\delta^{34}S$	$^{87}Sr/^{86}Sr$	$\delta^{13}C_{ca,f}$	$\delta^{18}O_{ca,f}$	$\delta^{15}N_{ca,f}$
Altitude	-0.84	0.74	-0.99	-0.59	0.04	0.47	-0.83	-0.15	-0.88	-0.14	0.79	-0.83	0.29
Latitude		-0.94	0.77	0.60	0.09	-0.61	0.86	0.34	0.87	0.18	-0.72	0.88	-0.22
Longitude			-0.65	-0.42	-0.23	0.47	-0.79	-0.39	-0.77	0.04	0.56	-0.80	0.05
$\delta^{18}O_{prec}$				0.56	-0.07	-0.44	0.78	0.11	0.84	0.12	-0.79	0.78	-0.28
$C\%$					-0.17	-0.39	0.41	0.07	0.62	0.38	-0.52	0.56	-0.46
$N\%$						-0.01	-0.02	-0.07	-0.07	-0.19	0.12	0.17	0.03
$\delta^{13}C$							-0.48	-0.10	-0.58	-0.31	0.71	-0.45	0.14
$\delta^{18}O$								0.16	0.75	0.01	-0.57	0.83	-0.32
$\delta^{15}N_{ca,f}$									0.27	0.23	-0.17	0.28	0.60
$^{87}Sr/^{86}Sr$										0.12	-0.83	0.79	-0.23
$\delta^{13}C_{ca,f}$											-0.28	0.16	-0.05
$\delta^{18}O_{ca,f}$												-0.67	0.12
													-0.24

A.2 Detecção de *outliers*

O intuito dos mapas apresentados nesta Secção é identificar a origem geográfica das observações consideradas *outliers* na Secção 3.4. Assim, apresentam-se nas figuras A.1, A.2, A.3 e A.4, os mapas das quatro regiões produtoras de café em estudo, com identificação dos *outliers* detetados no set_3 , set_7 , set_9 e set_{10} , respetivamente. Identifica-se a **vermelho** e com o símbolo em forma de gota as observações consideradas *outliers* no subconjunto correspondente. Atente-se que existem observações que não constam destes mapas por não se dispor das coordenadas geográficas, como é o caso das observações classificadas como *outliers* no set_7 .

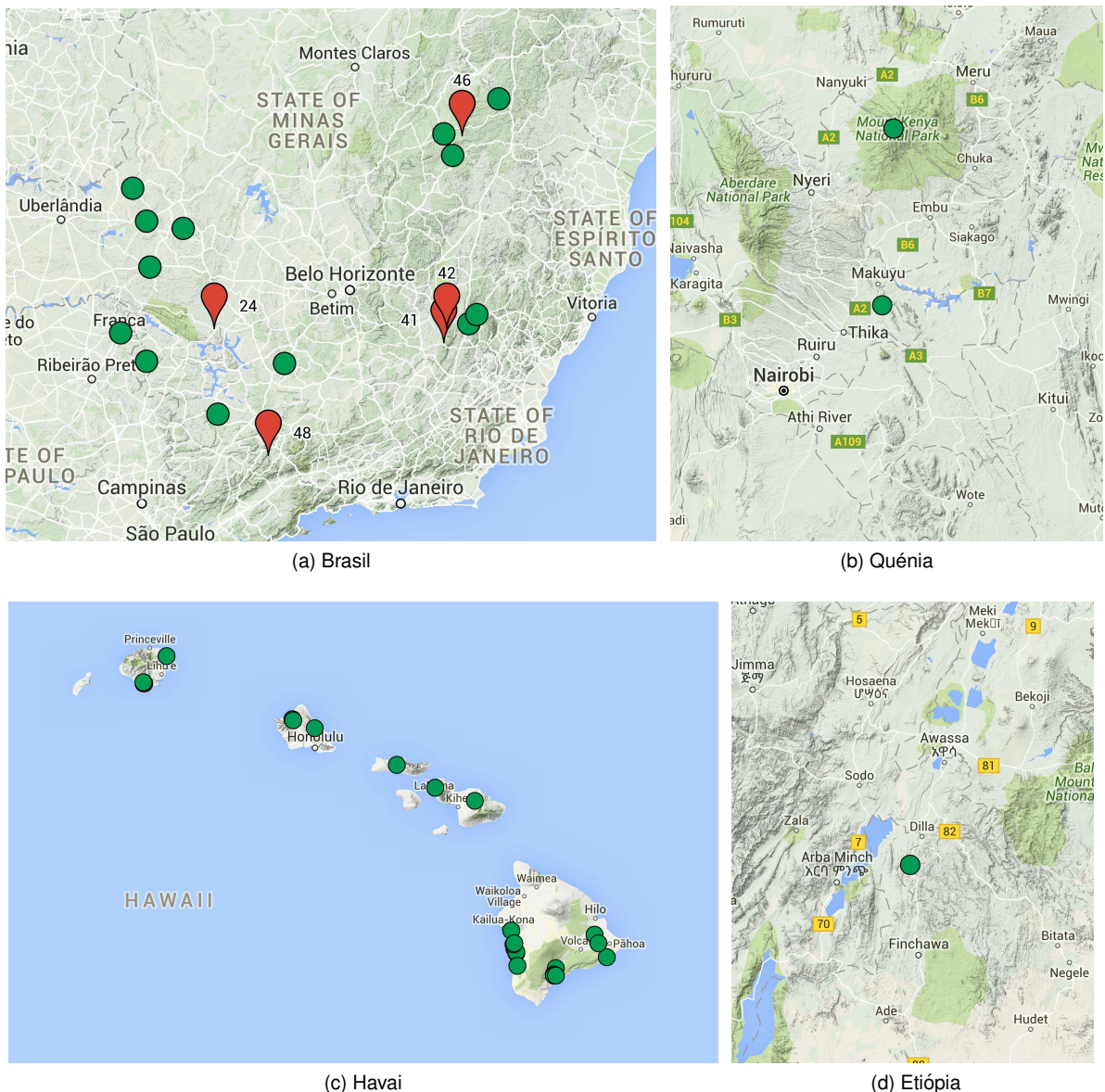
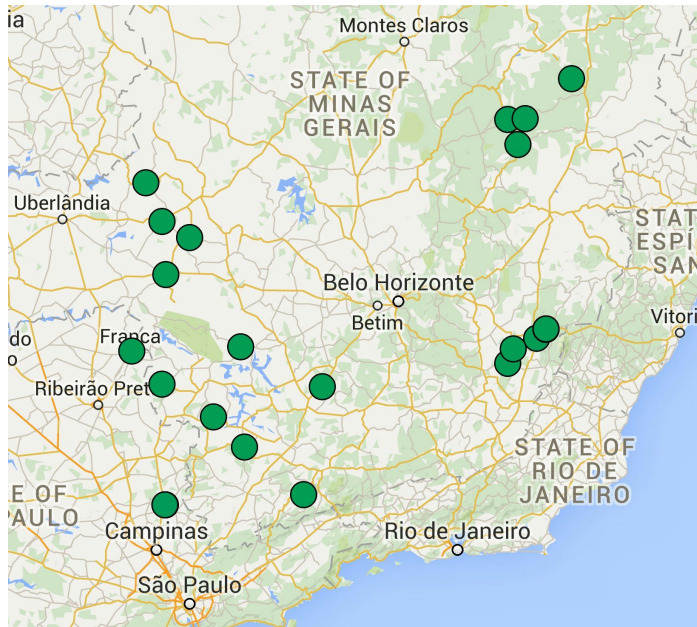
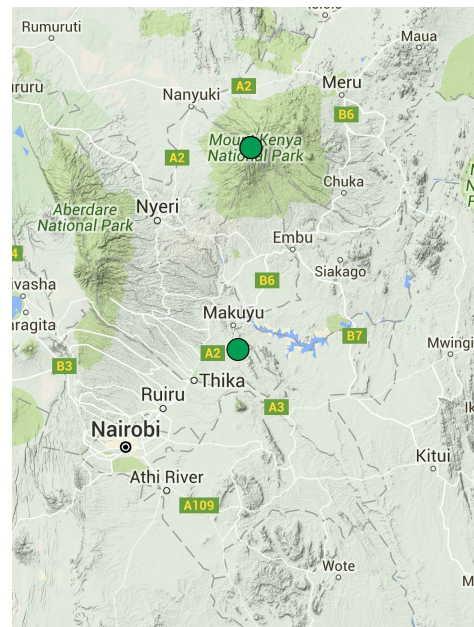


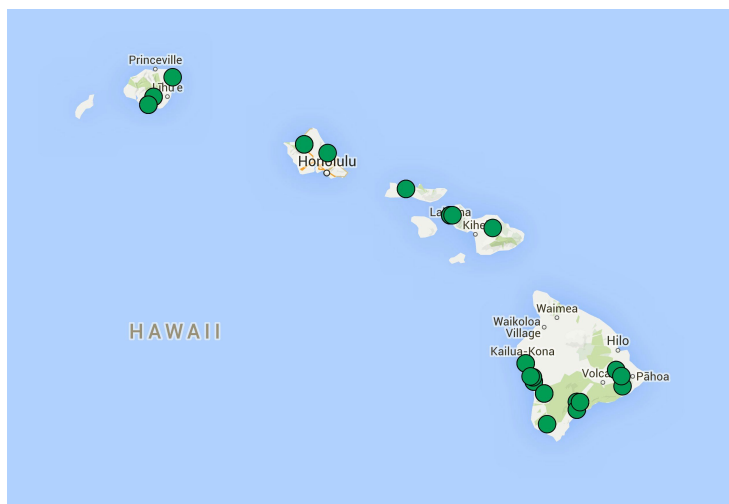
Figura A.1: Mapas do Brasil, Quênia, Havai e Etiópia, respetivamente (com coordenadas geográficas disponíveis), com identificação das observações consideradas *outliers* no set_3 a **vermelho**. (Map data ©2015 Google, INEGI)



(a) Brasil



(b) Quênia

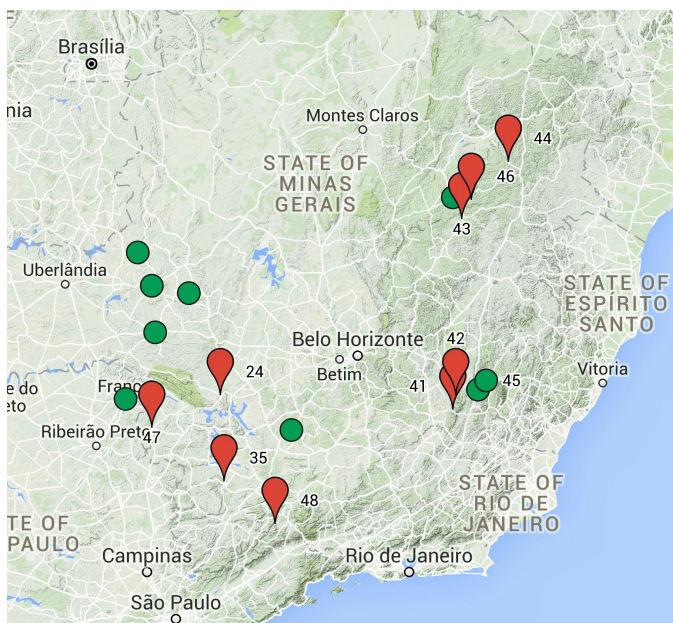


(c) Havai



(d) Etiópia

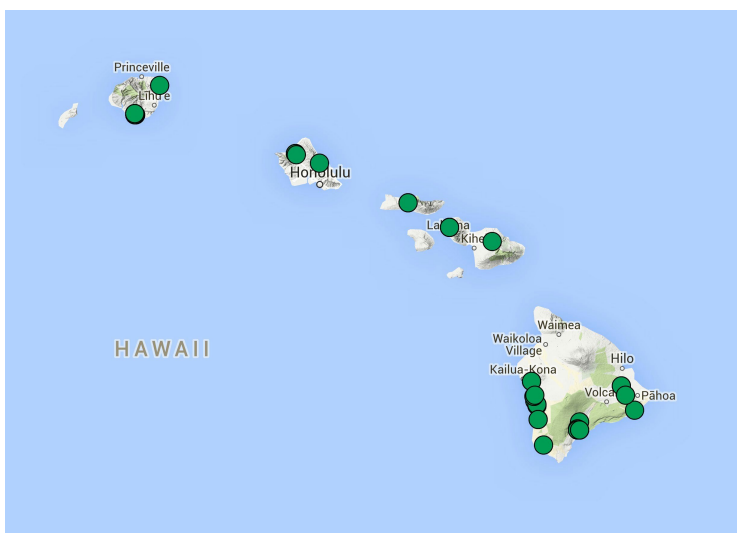
Figura A.2: Mapas do Brasil, Quênia, Havai e Etiópia, respetivamente (com coordenadas geográficas disponíveis), com identificação das observações consideradas *outliers* no *set*₇ a **vermelho**. (Map data ©2015 Google, INEGI)



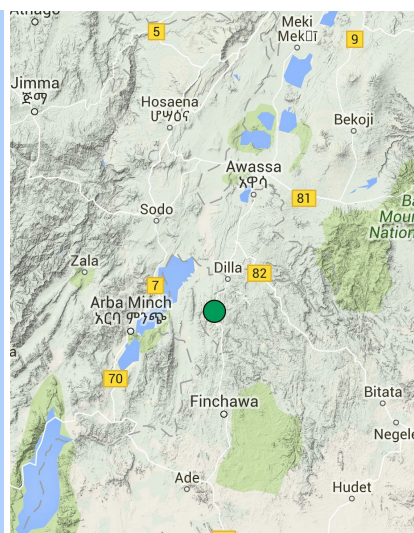
(a) Brasil



(b) Quênia

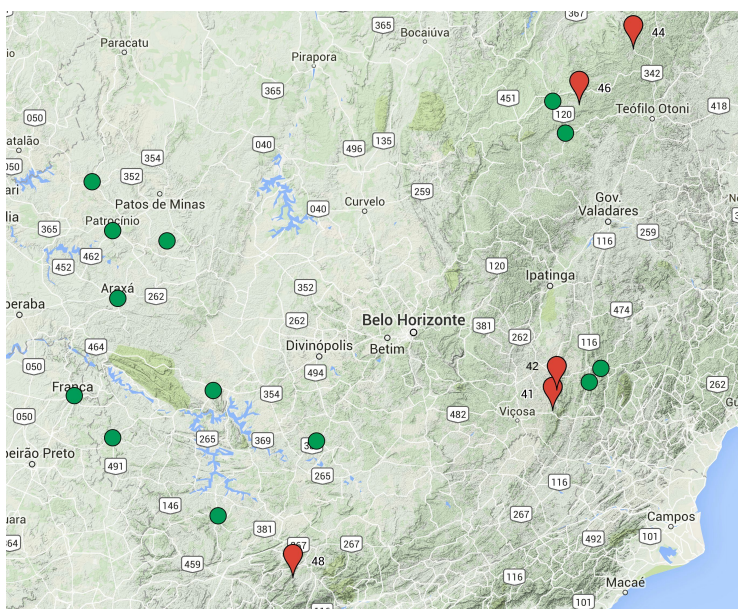


(c) Havai

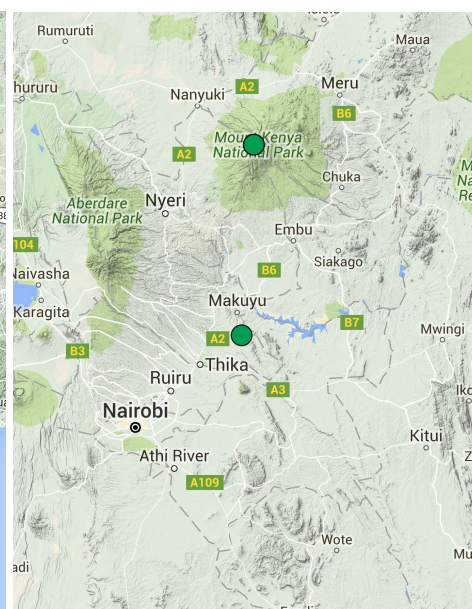


(d) Etiópia

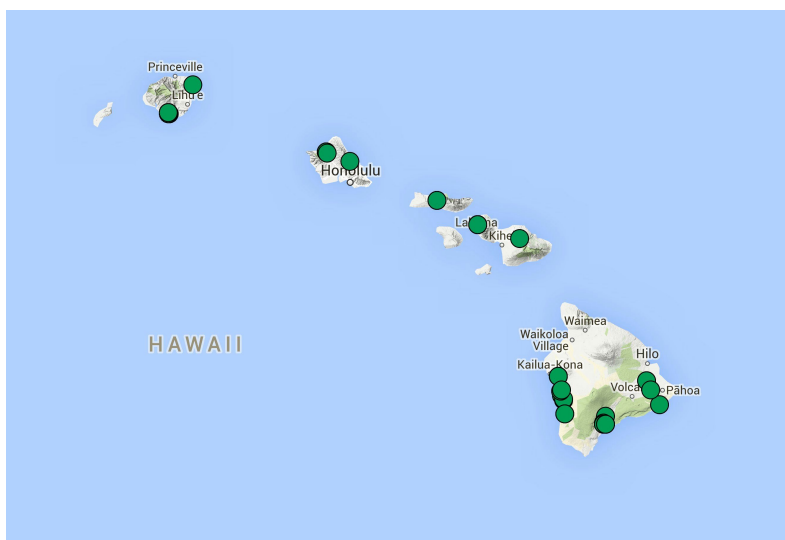
Figura A.3: Mapas do Brasil, Quênia, Havai e Etiópia, respetivamente (com coordenadas geográficas disponíveis), com identificação das observações consideradas *outliers* no *set₉* a **vermelho**. (Map data ©2015 Google, INEGI)



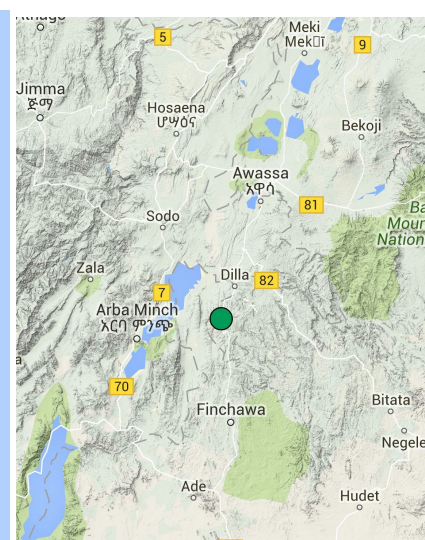
(a) Brasil



(b) Quênia



(c) Havai



(d) Etiópia

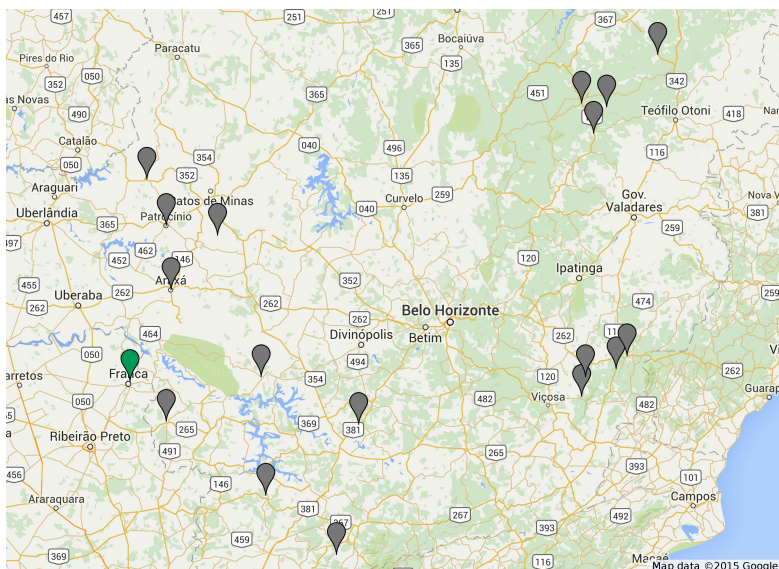
Figura A.4: Mapas do Brasil, Quênia, Havai e Etiópia, respetivamente (com coordenadas geográficas disponíveis), com identificação das observações consideradas *outliers* no set_{10} a **vermelho**. (Map data ©2015 Google, INEGI)

A.3 Agrupamentos

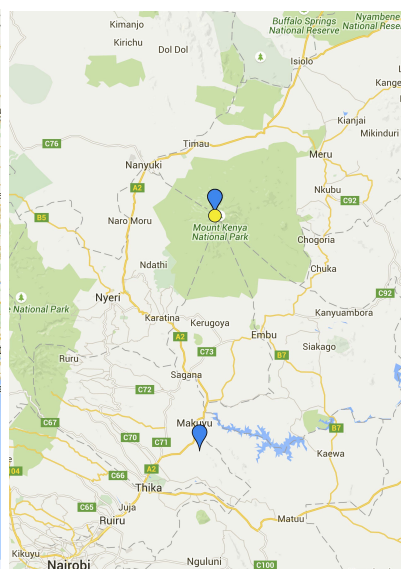
Nesta secção apresentam-se os mapas com a localização das observações do conjunto de dados em estudo na Figura A.5, com identificação dos grupos a que foram atribuídas na análise de agrupamentos aos *scores* da ACP do *set*₃ realizada no Capítulo 4.

A identificação do grupo atribuído a cada observação é feita pelas cores dos símbolos. Ao grupo 1 corresponde o azul, cuja maioria das observações pertence ao Quénia (como se verifica na Figura A.5b). O grupo 2 é identificado pela cor amarela e deste grupo faz parte a maioria das observações do arquipélago do Havai (à exceção da ilha Havai) (*vide* Figura A.5c). O grupo 3 é identificado pela cor verde, sendo que a maioria das observações pertence à ilha Havai. Por fim, ao grupo 4 corresponde o cinza, sendo a maioria das observações pertencente ao Brasil (como se verifica na Figura A.5a).

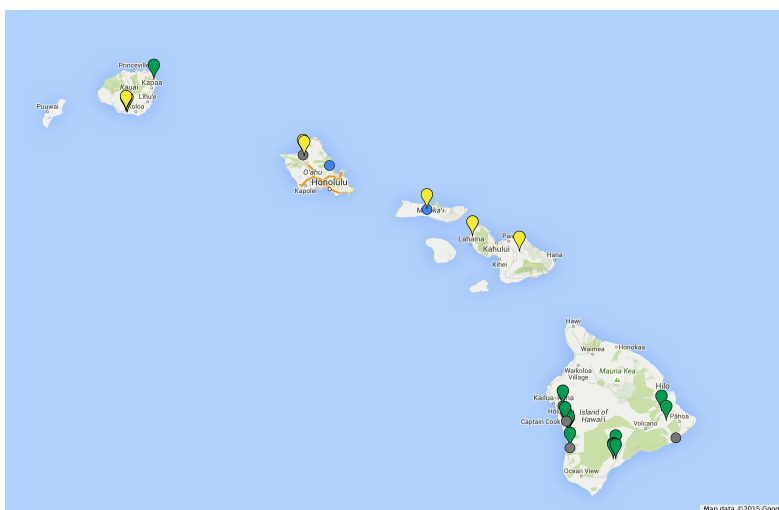
Além da identificação do grupo atribuído e com o intuito de facilitar a identificação das observações *mal alocadas*, estas são representadas pelo símbolo em forma de círculo, enquanto as observações que foram atribuídas ao grupo cuja sua origem geográfica está em maioria têm associado um símbolo em forma de gota.



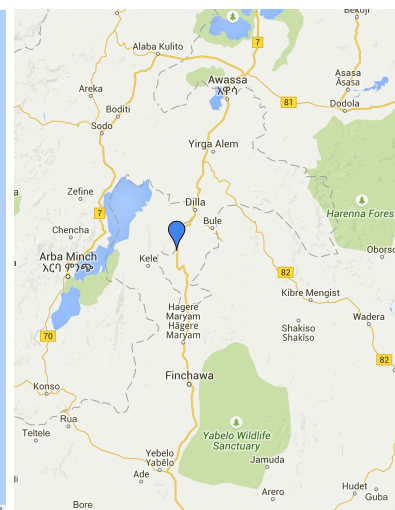
(a) Brasil



(b) Quênia



(c) Havai



(d) Etiópia

Figura A.5: Mapas do Brasil, Quênia, Havai e Etiópia, respetivamente (com coordenadas geográficas disponíveis), com identificação do grupo obtido da análise de agrupamentos aos *scores* da ACP do *set*₃, a azul as observações do *G*₁, a amarelo as observações do *G*₂, a verde as observações do *G*₃ e a cinza as observações do *G*₄. (Map data ©2015 Google, INEGI)